Differential methylation analysis of preeclamptic placental tissue using the Illumina HumanMethylation450K array

Abdullah Farouk , Amy Inkster, Nikolas Krstic & Yue Yang Shen

STAT540 UBC, April 2018

INTRODUCTION

Background: Preeclampsia (PE) is a hypertensive disorder affecting 2-8% of all pregnancies [1]. The symptoms range from hypertension and proteinuria to more severe complications such as renal/liver failure and seizures [2]. Multiple studies have demonstrated that PE is associated with distinctive placental DNA methylation [3, 4].

Purpose: The objectives of our study are 1) to use methylation data to predict patients' PE status; and 2) to identify targets and pathways associated with preeclampsia for further study.

Method: We combined two DNA methylation datasets publicly available on GEO (GSE57767 [5] and GSE44667 [6]). All the data were generated with the genome-wide Illumina Infinium Methylation 450 BeadChip array. Analytical methods that were used include principal component analysis, agglomerative hierarchical clustering, limma ("linear models for microarray data") and several supervised learning approaches.

DATA

Data	Cases (PE)	Controls
GSE44667 [6]	20	20
GSE57767 [5]	19 term /12 preterm	14

Table 1. Publicly available Illumina HumanMethylation450K array PE placental datasets used in this analysis:

EXPLORATORY ANALYSIS

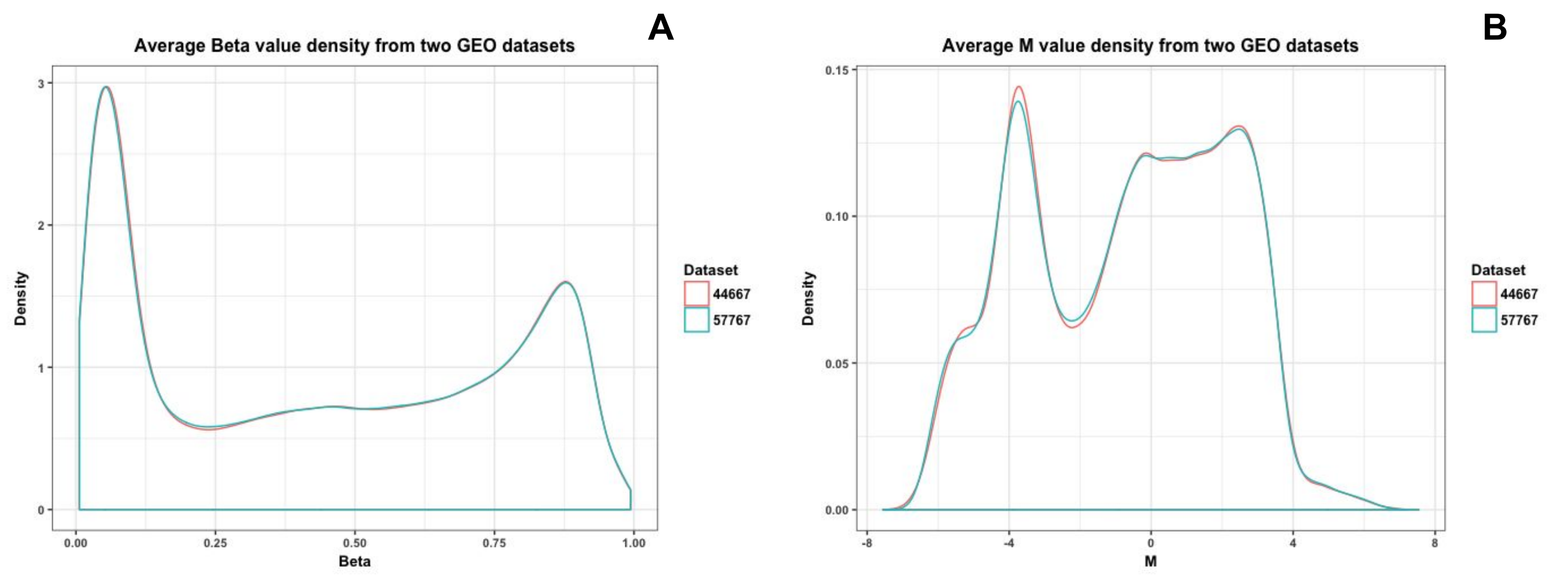

Figure 1. A: Density plots of beta-values after normalization. Beta values from the two datasets show similar distribution. **B:** Density plot of M values after normalization. When examining the M values of individual CpG-islands, they appeared relatively normally distributed.

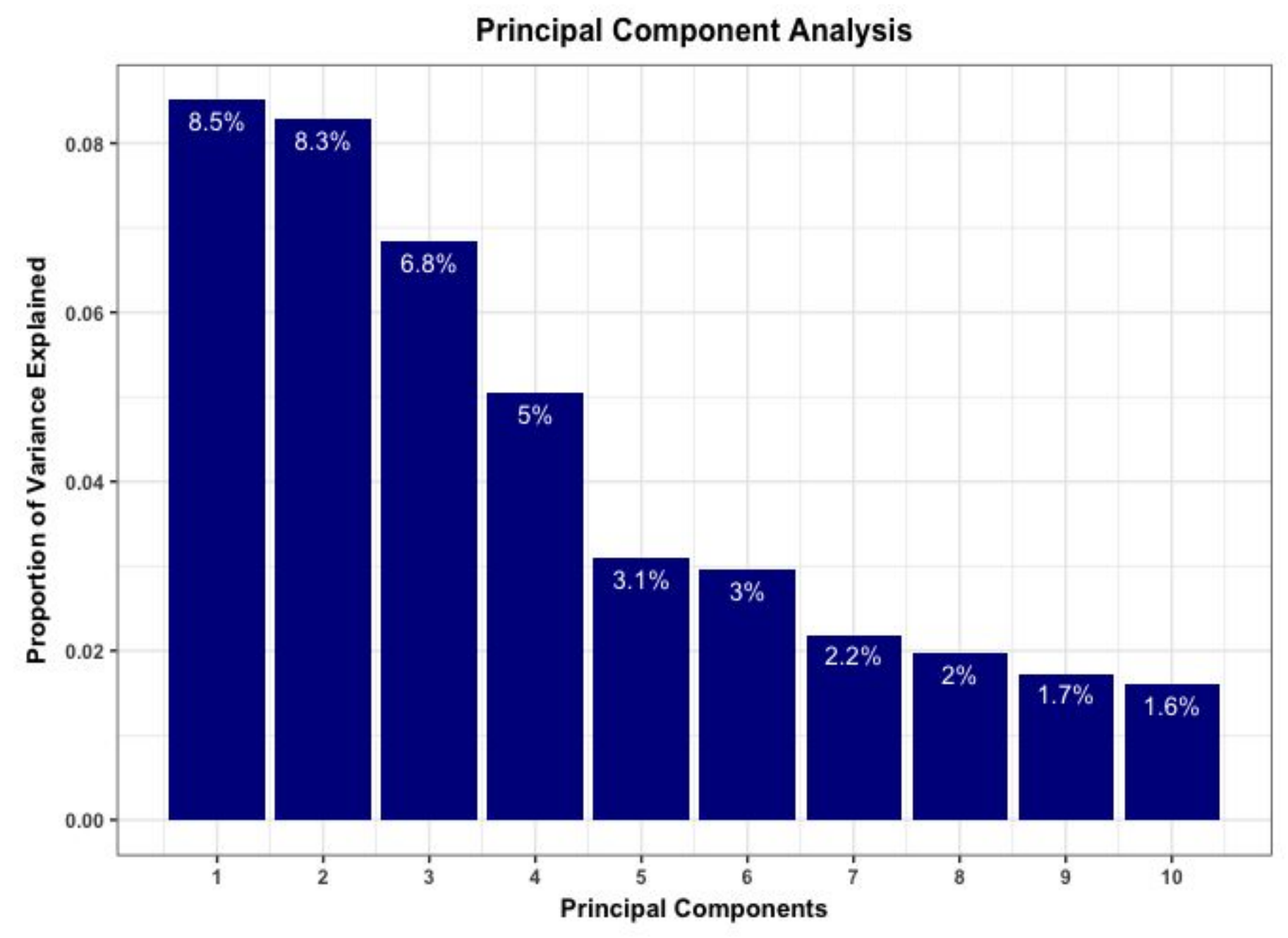

Figure 2. Principal component analysis was performed per CpG island. The top 3 PC's capture 23.6% of the total variance in our data, and the 10 PC's capture 41.8%. Our datasets were not corrected for batch effects and gestational age, so we included PCs 1-3 as covariates in our subsequent linear model to account for batch effects and cell-type heterogeneity (known to be a factor in the placenta) [7].

RESULTS

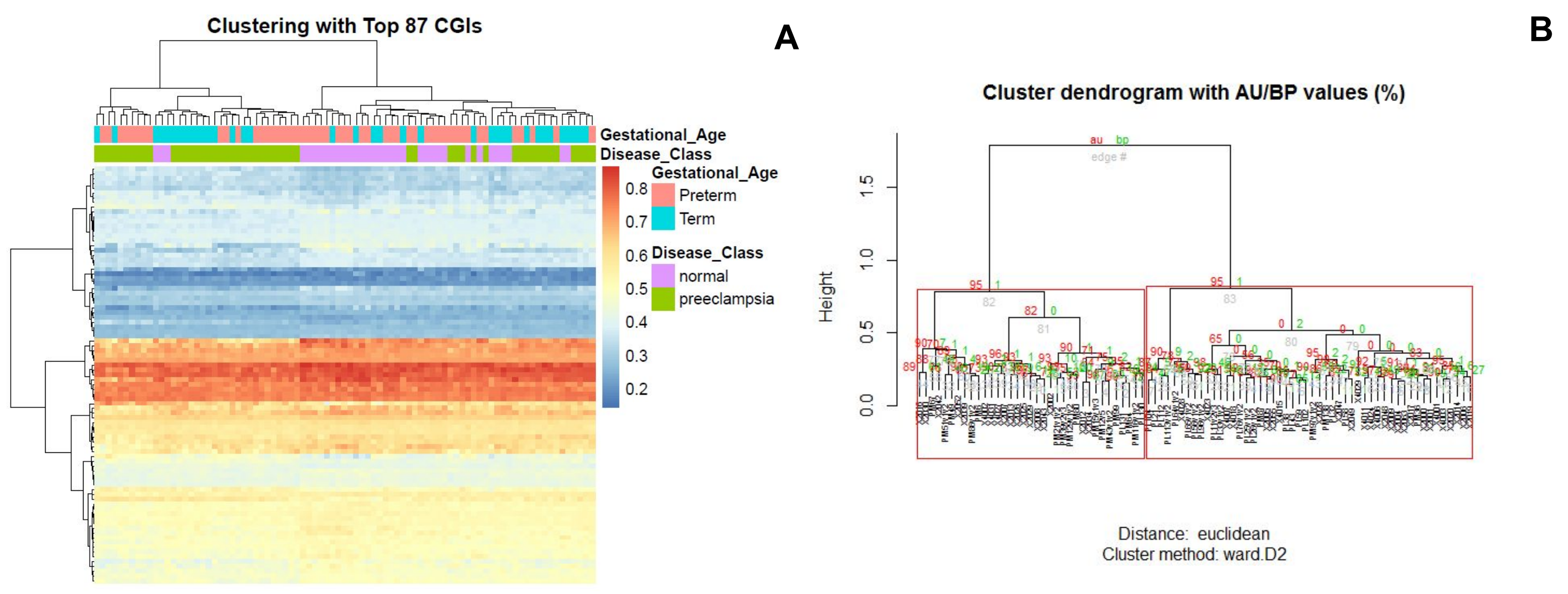

Figure 3. A: Heatmap of the clustering result of the top 87 differentially methylated CpG islands at FDR < 0.01 (20 hyper-, and 67 hypo-methylated). The clustering method applied was Ward's method; our distance metric was "euclidean". **B:** Using multiscale bootstrap resampling we verified the stability of the clusters found using hierarchical clustering.

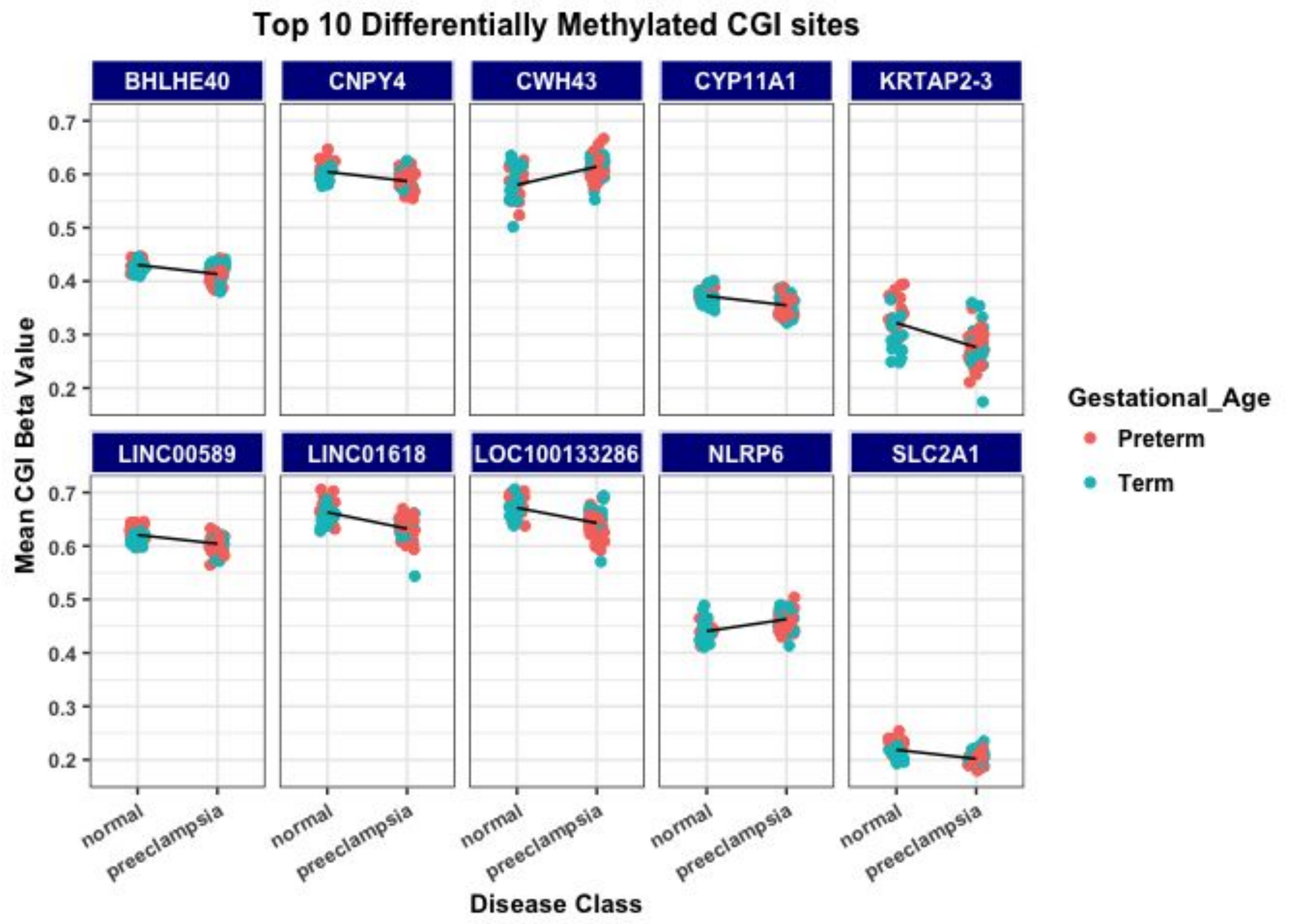

Figure 4. Top 10 differentially methylated sites using limma. To account for batch effects and cell type heterogeneity, we performed reference-free deconvolution using the top 3 principal components as covariates in our linear model (figure not shown).

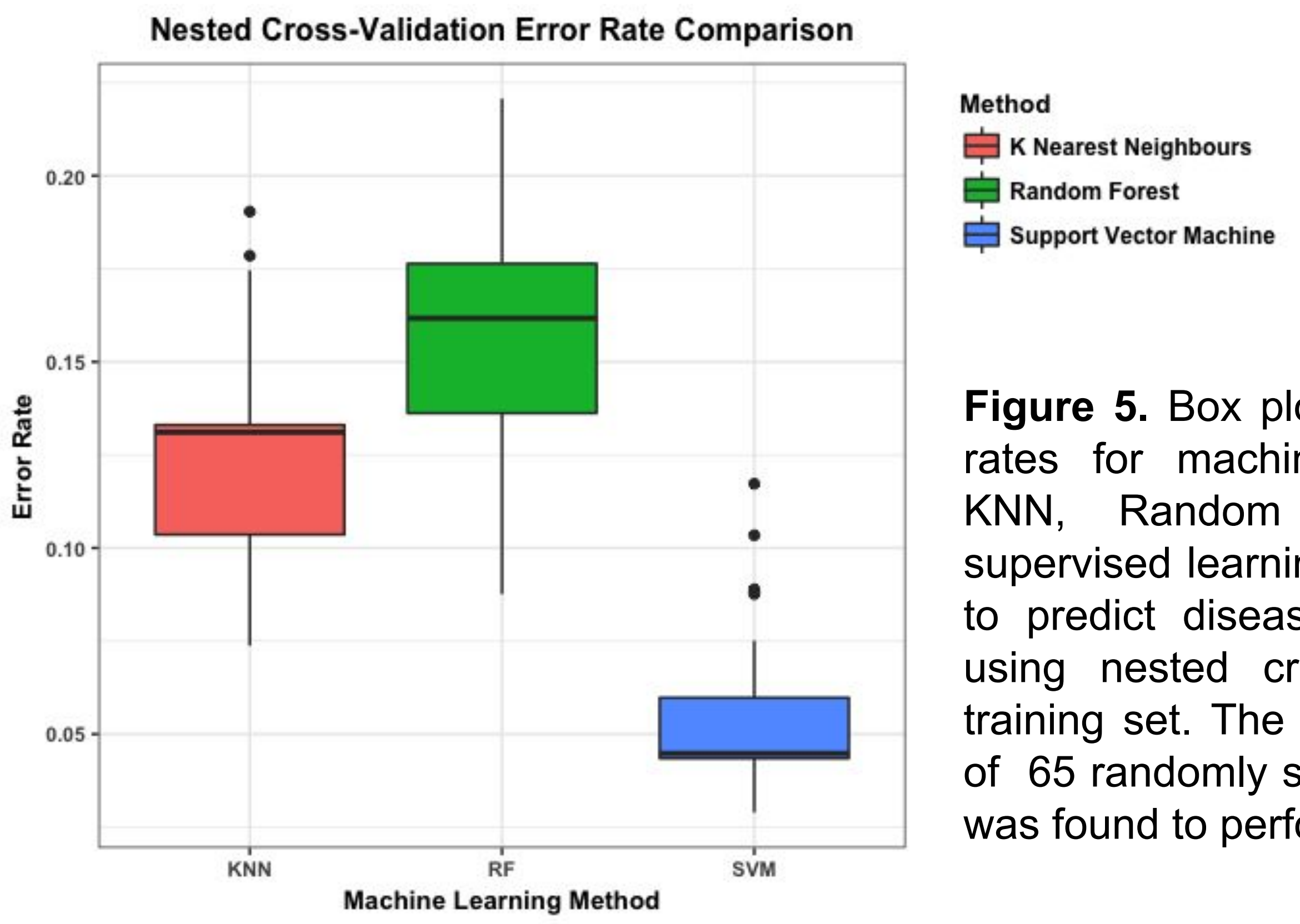

Figure 5. Box plot comparison of error rates for machine learning methods. KNN, Random Forest, and SVM supervised learning methods were used to predict disease status (PE/healthy) using nested cross-validation on the training set. The training set comprised of 65 randomly selected samples. SVM was found to perform best.

RESULTS

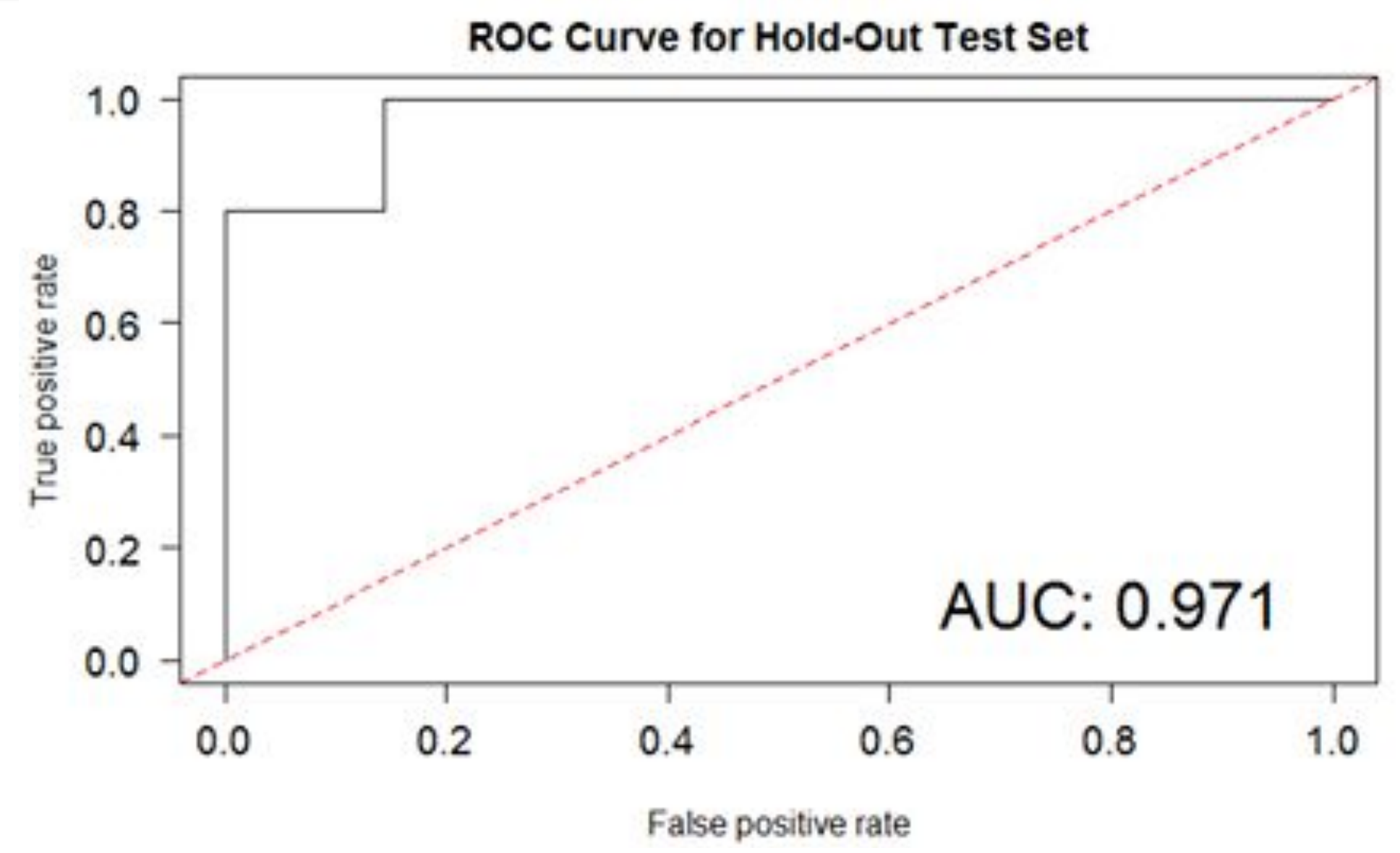

Figure 6. Receiver operating characteristic (ROC) curve for SVM supervised learning method. 17 samples were randomly selected to comprise the "test" set, and were withheld from training. The SVM model was fit with the full training set, and then was used to obtain the class probabilities in the 17-sample withheld test set.

GO ID	GO Term	Number of Genes	Raw Score	P Value	Corrected P value
GO:0065007	biological regulation	52	0.8430626	0.0708	1.0000000
GO:0044249	cellular biosynthetic process	21	0.4103888	0.1532	0.9000500

Table 3. Top terms from gene set enrichment analysis. These terms may relate to insufficient placentation observed in preeclampsia or synthetic cycles disrupted by the pathology [8].

DISCUSSION

We identified 87 differentially methylated CpG islands in our standard linear regression analysis. We also performed reference-free deconvolution taking PCs 1-3 into consideration as covariates in our linear model. Five CGIs were found to be differentially methylated in both approaches: CYP19A1, KRT19, GPR37L1, RNF217-AS1, and ZNF814. Proteins of the GPR, RNF, and ZNF families have previously been associated with differential methylation in the placenta of preeclamptic women [4].

Supervised learning was used to train a method to predict the disease status of samples based on methylation data. Feature selection was performed on 80% of samples using limma by 5-fold cross-validation. We then performed nested cross-validation (CV) to compare methods (KNN, SVM, and random forest) while simultaneously tuning parameters to select our approach. After choosing Repeated CV was used to tune the parameters of SVM; the chosen approach with the lowest error rate.

In terms of limitations, our datasets were missing information about experimental batch. We tried to account for potential batch and gestational age effects by including PCs 1-3 as covariates in our linear model. Validation of our SVM model on a separate test dataset (GSE73375) was not possible at this time, due to misalignments in the distributions of covariates common to both test and training datasets, so the generalizability of our model is yet to be tested.

REFERENCES

1. Duley L. 2009. The Global impact of pre-eclampsia and eclampsia. Semin Perinatol. 33:130–137.
2. [ACOG] The American College of Obstetricians and Gynecologists' Task Force on Hypertension in Pregnancy. 2013. Hypertension in Pregnancy. Obstet Gynecol. 122:1122-31.
3. Yeung KR, Chiu CL, Pidsley R, Makris A, Hennessy A, Lind JM. 2016. DNA methylation profiles in preeclampsia and healthy control placentas. Am J Physiol Heart Circ Physiol. 310:H1295-303.
4. Anderson CM, Ralph JL, Wright ML, Linggi B, Ohm JE. 2014. DNA methylation as a biomarker for preeclampsia. Biol Res Nurs. 16:409-20.
5. Anton L, Brown AG, Bartolomei MS, Elovitz MA. 2014. Differential methylation of genes associated with cell adhesion in preeclamptic placentas. PLoS One. 9:e100148.
6. Blair JD, Yuen RK, Lim BK, McFadden DE, von Dadelszon P, Robinson WP. 2013. Widespread DNA hypomethylation at gene enhancer regions in placentas associated with early-onset pre-eclampsia. Mol Hum Reprod. 19:697-708.
7. Martin E, Ray PD, Smeester L, Grace MR, Boggess K, Fry RC. 2015. Epigenetics and preeclampsia: defining functional epimutations in the preeclamptic placenta related to the TGF-β pathway. PLoS One. 10:e0141294.
8. Karteris E, Vatish M, Hillhouse EW, Grammatopoulos DK. 2005. Preeclampsia is associated with impaired regulation of the placental nitric oxide-cyclic guanosine monophosphate pathway by corticotropin-releasing hormone (CRH) and CRH-related peptides. J Clin Endocrinol Metab. 90:3680-3687.

ACKNOWLEDGEMENTS

We are grateful to Sarah Mostafavi and Emma Graham for their support and advice.