TYPE Original Research
PUBLISHED 05 September 2023
DOI 10.3389/fgene.2023.1248088

Check for updates

OPEN ACCESS

EDITED BY
Vanessa Meyer,
University of the Witwatersrand,
South Africa

REVIEWED BY
Abhimanyu Abhimanyu,
The University of Chicago, United States
James Breen,
University of Adelaide, Australia
Carlos Galaviz-Hernandez,
National Polytechnic Institute (IPN),
Mexico

*CORRESPONDENCE
W. P. Robinson,
✉ wrobinson@bcchr.ca

†These authors share senior authorship

RECEIVED 26 June 2023
ACCEPTED 08 August 2023
PUBLISHED 05 September 2023

CITATION
Fernández-Boyano I, Inkster AM, Yuan V and Robinson WP (2023), eoPred: predicting the placental phenotype of early-onset preeclampsia using public DNA methylation data.
Front. Genet. 14:1248088.
doi: 10.3389/fgene.2023.1248088

eoPred: predicting the placental phenotype of early-onset preeclampsia using public DNA methylation data

I. Fernández-Boyano[1,2], A. M. Inkster[1,2], V. Yuan[1,2] and W. P. Robinson[1,2]*†

[1]BC Children's Hospital Research Institute (BCCHR), Vancouver, BC, Canada, [2]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Background: A growing body of literature has reported molecular and histological changes in the human placenta in association with preeclampsia (PE). Placental DNA methylation (DNAme) and transcriptomic patterns have revealed molecular subgroups of PE that are associated with placental histopathology and clinical phenotypes of the disease. However, the clinical and molecular heterogeneity of PE both across and within subtypes complicates the study of this disease. PE is most strongly associated with placental pathology and adverse fetal and maternal outcomes when it develops early in pregnancy. We focused on placentae from pregnancies affected by preeclampsia that were delivered before 34 weeks of gestation to develop eoPred, a predictor of the DNAme signature associated with the placental phenotype of early-onset preeclampsia (EOPE).

Results: Public data from 83 placental samples (HM450K), consisting of 42 EOPE and 41 normotensive preterm birth (nPTB) cases, was used to develop eoPred—a supervised model that relies on a highly discriminative 45 CpG DNAme signature of EOPE in the placenta. The performance of eoPred was assessed using cross-validation (AUC = 0.95) and tested in an independent validation cohort (*n* = 49, AUC = 0.725). A subset of fetal growth restriction (FGR) and late-PE cases showed a similar DNAme profile at the 45 predictive CpGs, consistent with the overlap in placental pathology between these conditions. The relationship between the EOPE probability generated by eoPred and various phenotypic variables was also assessed, revealing that it is associated with gestational age, and it is not driven by cell composition differences.

Conclusion: eoPred relies on a 45-CpG DNAme signature to predict a homogeneous placental phenotype of EOPE in a discrete or continuous manner. Using this classifier should 1) aid in the study of placental insufficiency and improve the consistency of future placental DNAme studies of PE, 2) facilitate identifying the placental phenotype of EOPE in public data sets and 3) importantly, standardize the placental diagnosis of EOPE to allow better cross-cohort comparisons. Lastly, classification of cases with eoPred will be useful for investigating the relationship between placental pathology and genetic or environmental variables.

KEYWORDS

placenta, prediction model, preeclampsia, epigenetics, DNA methylation, placental epigenetics

Background

Preeclampsia (PE) is a hypertensive disorder of pregnancy, and is one of the leading causes of maternal and perinatal morbidity and mortality worldwide. The Society of Obstetricians and Gynecologists of Canada defines PE as gestational hypertension with new-onset proteinuria or one or more adverse conditions, which include maternal end-organ complications and evidence of uteroplacental dysfunction (Magee et al., 2014). While the global incidence of PE is estimated as 4.6% (Abalos et al., 2013), this is an approximation limited by the lack of data from many regions, and varies by population, race, and socioeconomic status (Silva et al., 2008; Lisonkova and Joseph, 2013; Wang et al., 2021). Encompassing a diverse range of clinical presentations and outcomes, PE is a complex and multifactorial disease, which exists on a spectrum of severity. Among others, chronic hypertension and prior PE are well-known risk factors for PE (Bartsch et al., 2016). Placental, maternal, and paternal genetics also significantly contribute to the pathogenesis of PE (Williams and Broughton Pipkin, 2011; Wang et al., 2022). The combination of such genetic and environmental factors plays an important role in the development of PE, which is a two-stage process whereby placental dysfunction leads to maternal onset of disease (Redman et al., 2014); several studies have observed PE-associated molecular and histopathological changes in the human placenta, likely reflecting the observed placental dysfunction (Leavey et al., 2016; Benton et al., 2018; Wilson et al., 2018). Characterization of placental PE-associated changes can complement clinical findings to gain a better understanding of the disease and assist with its classification into more homogeneous subtypes.

The umbrella syndrome of PE is likely the junction where at least two distinct, and likely interacting pathways to disease converge, initially coined as “placental” and “maternal” (Roberts and Escudero, 2012; Redman et al., 2014; Redman, 2017; Staff and Redman, 2018; Staff, 2019, 201). Clinically, diagnosis before or at/after 34 weeks gestational age (GA) may be used to define two subtypes of PE: early-onset PE (EOPE) and late-onset PE (LOPE) (Brown et al., 2018). EOPE, which often has a greater placental involvement (Redman, 2017), is more severe, frequently overlaps with other pathologies such as fetal growth restriction (FGR), and presents with placental pathology such as villous infarctions and hypermaturation. In contrast, LOPE may be more heavily influenced by predisposing maternal factors despite normal placentation, and is often milder (Staff, 2019). EOPE is thought to originate with reduced placental perfusion because of incomplete vascular remodelling during placentation (Staff, 2019). Oxidative stress secondary to malperfusion then leads to trophoblast damage, which eventually results in the systemic inflammatory response and ensuing hypertensive syndrome (Roberts and Hubel, 2009; Staff, 2019), although the mechanisms linking the two stages are still being elucidated (Roberts and Hubel, 2009). In contrast, the development of LOPE does not necessitate inadequate placentation but instead may result from compression of the terminal villi at term, which impedes appropriate perfusion, leading to syncytiotrophoblast hypoxia as in EOPE (Staff, 2019). All PE is thus associated with placental syncytiotrophoblast stress (Staff, 2019), albeit with distinct underlying causes and timing. Maternal factors such as obesity also contribute to all stages of the disease and play an important role in disease severity.

In the human placenta, DNA methylation (DNAme) has a unique profile that shifts throughout gestation in association with changes in cell composition, gene expression, and in response to pregnancy complications, among other factors (Robinson and Price, 2015; Yuan et al., 2021). Many studies have reported wide-spread PE-associated alterations in placental DNAme (Blair et al., 2013; Chu et al., 2014; Martin et al., 2015; Yeung et al., 2016; Kim, 2017; Wilson et al., 2018; Wang et al., 2019; Lim et al., 2020; Tsegaselassie et al., 2020), but findings are not consistently reproduced across studies (Cirkovic et al., 2020). Discrepancies between studies in definitions of PE used, study design (e.g., analysing all PE compared to analysing specific subtypes such as EOPE, LOPE), the use of different metrics to assess reproducibility (e.g., a CpG found to be significantly differentially methylated in one cohort may not replicate in an independent cohort, but a proximal and correlated CpG might), population differences in contributing genetic and/or environmental factors, and platforms used to measure DNAme, among others, can contribute to poor reproducibility across studies. Identification of homogeneous subtypes of PE is essential to successfully identify and manage each subtype (Myatt et al., 2014).

PE-associated molecular variation in the placenta may be used to refine current classification of PE. Wilson et al. found that the placental DNAme profile of intermediate onset PE (<36 weeks) co-occurring with FGR is similar to that of EOPE, suggesting that the clinical threshold of 34 weeks is imperfect and will misclassify cases (Wilson et al., 2018). Further, integration of DNAme with transcriptional information revealed up to four molecular PE subtypes and several histopathological findings were found to associate with each molecular PE subcluster (Leavey et al., 2016). Severity also existed on a gradient within each subcluster, supporting the hypothesized connection between molecular changes and pathology in the placenta, as well as illustrating the value of studying DNAme and transcriptomic profiles to characterize PE.

To address these challenges, we developed eoPred, a DNAme-based model that uses previously collected 450K Illumina DNAme microarray data from placental chorionic villus samples taken at delivery to find a placental signature of EOPE that is robust to cohort differences. This model was validated by predicting the disease status of samples in an independent cohort set of samples. Furthermore, the model outputs a continuous probability that can be valuable in correlating placental changes phenotype with environmental and genetic variables, as well as with postnatal outcomes.

Methods

Study data

Placental DNAme data (n = 401) were collected from eight Illumina Infinium HumanMethylation450 Beadchip array (HM450K) datasets (GSE100197 (Leavey et al., 2018; Wilson et al., 2018), GSE103253 (Herzog et al., 2017), GSE120981 (Monteagudo-Sánchez et al., 2019), GSE73375 (Martin et al., 2015), GSE75196 (Yeung et al., 2016), GSE98224 [Leavey et al., 2018; Wilson et al., 2018), GSE49343 (Blair et al., 2014)] and one

TABLE 1 Arrangement of datasets from the Gene Expression Omnibus into the three study groups. Samples in the training and validation groups only included those samples labelled as EOPE and nPTB by dataset authors. Samples from GSE100197, GSE103253, GSE57767, GSE73375, GSE125605, GSE98224, and GSE203396 that were not included in the training and validation groups (i.e., from etiologies other than EOPE or nPTB) are in the exploratory group.

	Training	Validation	Exploratory
GSE100197	●		●
GSE103253	●		●
GSE57767	●		●
GSE73375	●		●
GSE125605		●	●
GSE98224		●	●
GSE203396		●	●
GSE75196			●
GSE120981			●
GSE49343			●

Illumina Infinium HumanMethylation850 (EPIC) dataset (GSE203396 (Inkster et al., 2022) available on Gene Expression Omnibus (GEO). Using samples from these datasets, we created three separate groups for i) model training, ii) model validation and iii) model exploration (Table 1). A flowchart of the steps followed to create and process these cohorts is described in Supplementary Figure S1.

To develop a predictive model for canonical EOPE, the training group was restricted to EOPE and normotensive preterm birth (nPTB) samples to avoid confounding by inclusion of LOPE or normotensive fetal growth restriction (FGR), which are heterogeneous groups that may partly overlap with the placental phenotype of EOPE, and normotensive term samples, which may differ from EOPE due to GA-associated changes rather than due to pathology. We chose to generate a binary classifier that differentiated between EOPE and nPTB rather than a multi-class predictor (i.e., also including FGR and LOPE) because of the placental dysfunction observed in EOPE compared to other pathologies. The validation group was also comprised of EOPE and nPTB samples. The exploratory group consisted of a broader mix of samples with different pathologies, exclusive of those used in the training and validation groups, to better understand the application of our model to a variety of placental phenotypes. Cord blood DNAme data was downloaded from GSE110829 to test the tissue-specificity of eoPred.

The training group ($n = 89$) was assembled with the criteria that each of the datasets included (GSE100197, GSE103253, GSE57767, GSE73375) must contain a mixture of EOPE ($n = 47$) and nPTB ($n = 42$) cases, balanced by dataset (i.e., if a dataset included only 1 EOPE sample and several nPTB samples, it was excluded). The validation group ($n = 49$) was selected from three datasets (GSE125605, GSE98224, GSE203396) with a total of 38 EOPE and 11 nPTB samples. An exploratory group ($n = 329$) was then constructed from all remaining samples (GSE100197, GSE103253, GSE57767, GSE73375, GSE125605, GSE98224, GSE120981, GSE75196, GSE49343, GSE203396) that had a diagnosis other than EOPE or nPTB (i.e., late-PE, nFGR, CPM16 and nTB). This third group was assembled separately from the validation group to avoid any bias that might result from processing the data in the validation group with a dataset that was also included in the training cohort, even if the specific samples were not actually used for training the model. Demographic characteristics of each group are summarized in Table 2.

As sub-classification of PE into EOPE and LOPE was not always reported by dataset authors for all samples, PE was classified as follows: samples reported as "PE" were labelled as EOPE if delivered prior to 34 weeks and were otherwise classified as "late-PE." No EOPE-reported cases were labelled as late-PE using this approach, since cases delivered prior to 34 weeks would by necessity have been diagnosed before 34 weeks. While PE cases with clinical onset <34 weeks (EOPE) but delivery after 34 weeks would have been labelled as late-PE using this approach, none of these were included in model development and testing but were reserved only for exploratory analysis.

There were also inconsistencies in the reporting of GA across datasets. GA was provided in weeks and days by five datasets (GSE100197, GSE98224, GSE125605, GSE110829, GSE203396), two datasets reported weeks without days (GSE73375, GSE75196), one dataset (GSE57767) only distinguished between term and preterm samples, and two datasets (GSE103253, GSE120981) did not provide any metric of GA. To harmonize the GA metric across studies, placental epigenetic GA was calculated using the robust placental clock (RPC) (Lee et al., 2019) for all samples, using the R package planet (Lee et al., 2019; Yuan, 2023). The difference between reported and predicted GA in the datasets in which reported GA was available is summarized in Supplementary Figure S2 and Supplementary Table S1; reported and predicted GA were highly correlated in the training group ($R = 0.94$) and there was a median absolute error of 4 days across all samples in the three groups. We note that the robust placental clock was trained on GSE100197, amongst other datasets. Using RPC-predicted rather than reported GA did not change the classification of samples into EOPE and late-PE; importantly, no cases clinically classified as LOPE by the authors were re-labelled as EOPE using predicted GA (Supplementary Table

TABLE 2 Demographic characteristics of each constructed group included in this study. The data reflects the number of samples after processing.

	Training cohort N (%)	Validation cohort N (%)	Exploratory cohort N (%)	GSE110829 N (%)
Early-onset preeclampsia (EOPE)	42 (51%)	38 (78%)	2 (1%)	10 (9%)
Gestational age range (weeks)	31.6 (25–37)	33.25 (26–39)	32.78 (32–33)	29.91 (26–33)
Placental sex (XY/XX)	23/19	16/22	1/1	2/8
Late preeclampsia (Late-PE)	0	0	86 (26%)	10 (9%)
Gestational age range (weeks)	NA	NA	36.97 (34–41)	35.57 (34–38)
Placental sex (XY/XX)	NA	NA	45/41	3/7
Fetal growth restriction (FGR)	0	0	42 (14%)	0
Gestational age range (weeks)	NA	NA	36.43 (28–40)	NA
Placental sex (XY/XX)	NA	NA	24/22	NA
Normotensive preterm birth (nPTB)	41 (49%)	11 (22%)	3 (1%)	47 (43%)
Gestational age range (weeks)	33.76 (25–36)	34.42 (31–36)	33.11 (28–37)	30.44 (23–37)
Placental sex (XY/XX)	23/18	5/6	2/1	29/18
Normotensive term birth (nTB)	0	0	181 (55%)	43 (39%)
Gestational age range (weeks)	NA	NA	38.62 (37–42)	38.58 (37–41)
Placental sex (XY/XX)	NA	NA	92/89	19/24
Trisomy 16 confined to the placenta (CPM16)	0	0	10 (3%)	0
Gestational age range (weeks)	NA	NA	34.10 (25–39)	NA
Placental sex (XY/XX)	NA	NA	5/5	NA

S1). Genetic ancestry and cell composition were also calculated for all samples using planet (Yuan et al., 2019; 2021; Yuan, 2023).

Data processing

Data were downloaded as IDAT files wherever possible, otherwise methylated and unmethylated intensities were used. The training, validation, and exploratory group, and the cord blood dataset, were each processed independently using the same sample exclusion criteria, probe filtering criteria, and normalization method, beta-mixture quantile (BMIQ) from the R package wateRmelon (Pidsley et al., 2013).

After normalization, samples were removed from downstream analyses if they failed any of the following checks: i) mean inter-array correlation <95% (Edgar et al., 2017) ii) discordance between reported sex and chromosomal sex as inferred by X and Y chromosome fluorescence intensities (Heiss and Just, 2018), or iii) a measure of sample contamination based on allelic ratios of single-nucleotide polymorphism (SNP) probes implemented in the ewastools package (Heiss and Just, 2018). CpG probes were removed if they had poor quality signal in more than 5% of samples (detection p value > 0.01 or bead count < 3) (Aryee et al., 2014) or targeted loci cross-hybridizing to multiple sites (Price et al., 2013; Zhou et al., 2017), overlapped polymorphisms in the genome (Zhou et al., 2017), mapped to the X and Y chromosomes (Bibikova et al., 2011), or were non-variable both in this dataset and were reported as non-variable probes in the placenta (Edgar et al., 2017). To allow future application of eoPred on data collected with the Infinium MethylationEPIC v1.0 Beadchip array (EPIC), probes were filtered to those in common between the HM450K and EPIC platforms prior to developing the model. Probe filtering was only applied to the training group; to ensure that all 45 CpGs were present when applying eoPred, the other groups were normalized and checked for sample quality, but not filtered for poor-quality probes.

Lastly, batch correction was applied to the training group only using ComBat with the R package sva (Edgar et al., 2017) to correct for dataset differences. A model matrix was created with variables of interest (condition, sex, GA, and European and Asian ancestry coordinates assigned to each sample by planet (Yuan, 2023)) to preserve their variation, as recommended in (Leek et al., 2012). No batch correction was applied to the validation, exploratory, or cord blood groups.

After processing, 341,281 CpGs and 83 samples remained in the training group. The validation group consisted of 49 samples, the exploratory group was composed of 329 samples, and the cord blood data was comprised of 110 samples. Demographic data of the three constructed groups is shown in Table 1.

Model development

R Markdown source files and knitted reports for the analysis can be found on GitHub at https://github.com/iciarfernandez/eoPred/.

TABLE 3 Average performance metrics from cross-validation folds in the seven models tested during development, and in the permuted model. OER is the overall error rate, defined as the number of all incorrect predictions divided by the total number of samples in the data. OER (EOPE) is calculated exclusively over the EOPE samples (*n* = 38), and OER (nPTB) is calculated over the nPTB samples (*n* = 11). AUC is the area under the receiver operating characteristic curve, it is a measure of separability. The Brier score measures the accuracy of probabilistic predictions.

Model	OER	OER SD	OER (EOPE)	OER (nPTB)	AUC	Brier score
90 CpGs	0.10	0.01	0.13	0.07	0.95	0.14
75 CpGs	0.11	0.02	0.13	0.08	0.95	0.14
60 CpGs	0.11	0.02	0.14	0.08	0.95	0.14
45 CpGs (Final Model)	0.11	0.01	0.13	0.08	0.95	0.14
45 CpGs (Permuted Model)	0.53	0.05	0.52	0.55	0.45	0.26
30 CpGs	0.12	0.02	0.14	0.09	0.95	0.14
15 CpGs	0.12	0.02	0.15	0.09	0.95	0.14
5 CpGs	0.13	0.13	0.15	0.10	0.93	0.14

The R package mixOmics was used for all steps in model development and assessment (Rohart et al., 2017). Other R packages used to calculate model metrics were cvms (Renbo, 2016), MLmetrics (Yan, 2017), and caret (Kuhn, 2008). To maximize the use of the data, repeated (*n* = 50) M-fold (*n* = 3) cross-validation was used to develop and assess the performance of eoPred. Three folds were chosen to ensure that enough samples were present in each fold, given the relatively small sample size of the cohort. One fold was reserved to assess model performance, and the remaining two folds were used to train the model. This process was repeated 50 times, and each of the 50 performance estimations were averaged to produce a single estimation.

The model was trained on mean DNAme beta values, using sparse partial least squares discriminant analysis (sPLS-DA), which performs simultaneous dimensionality reduction and feature selection to find a molecular signature that can discriminate samples based on an outcome category (in this case, EOPE or nPTB). sPLS-DA has been previously shown to perform well on high-dimensional genomics data (Lê Cao et al., 2011). After the optimal parameters (number of components, and number of features per component) were selected during cross-validation, a final model was fit to the entire training data.

Overall misclassification error rate, receiver operating characteristic and area under the curve (ROC-AUC), sensitivity, and Brier score were the metrics used to assess the performance of the final model. A permutation test was run where a model with randomly permuted labels was trained and compared to the final model. The stability of the predictive DNAme-signature was also assessed using the perf function of the mixOmics package, which computes the frequency at which each CpG is selected in each cross-validation run, and uses this to evaluate each CpG's stability (i.e., a CpG selected with low frequency across all cross-validation runs may suggest that it is relevant to a certain split of the training data, but not to others and is thus "low stability") (Rohart et al., 2017).

Model validation

To test the performance of eoPred, the predict function in mixOmics (Rohart et al., 2017) was used to assign a predicted class to each of the 48 samples in the validation group. The maximum prediction distance was chosen as it resulted in the lowest error rate during the cross-validation process. Each sample is thus assigned two prediction distances (one per outcome category) and the predicted class is the outcome category with the largest predicted dummy value. Class probabilities were calculated from transformed prediction distances using the softmax function, such that each sample was assigned two estimated probabilities of being classified as either EOPE or nPTB, which sum to 1. Each sample was then predicted to be the class with the largest class probability.

Results

Development of a supervised model using placental DNAme that predicts EOPE

To develop a placental DNAme-based classifier that predicts the placental phenotype of EOPE, we used sparse partial least discriminant analysis (sPLS-DA) to select the CpG sites with the greatest ability to discriminate between EOPE and controls. We developed this model in a training group composed of 4 distinct cohorts, and then applied it to predict the disease status of samples in an independent validation group composed of three separate cohorts.

Repeated cross-validation was used to select the optimal number of components and features (CpGs) per component, with a component being constructed from a linear combination of the features. One component constructed from 90 CpGs resulted in the lowest average overall misclassification error rate (OER) across all cross-validation folds. The optimal number of components is typically K-1, with K being the number of classes to be predicted by the model. Introducing more components in the model did not significantly improve cross-validation performance. While this model performed well (OER = 0.10, AUC = 0.97), 35% of the 90 CpGs selected had low stability (frequency of selection across cross-validation folds < 0.5). These CpGs also contributed less than those with higher stability to the model's discriminative ability, suggesting that a simpler model with less CpGs may be equally successful in predicting the outcome of new samples. To test this, we

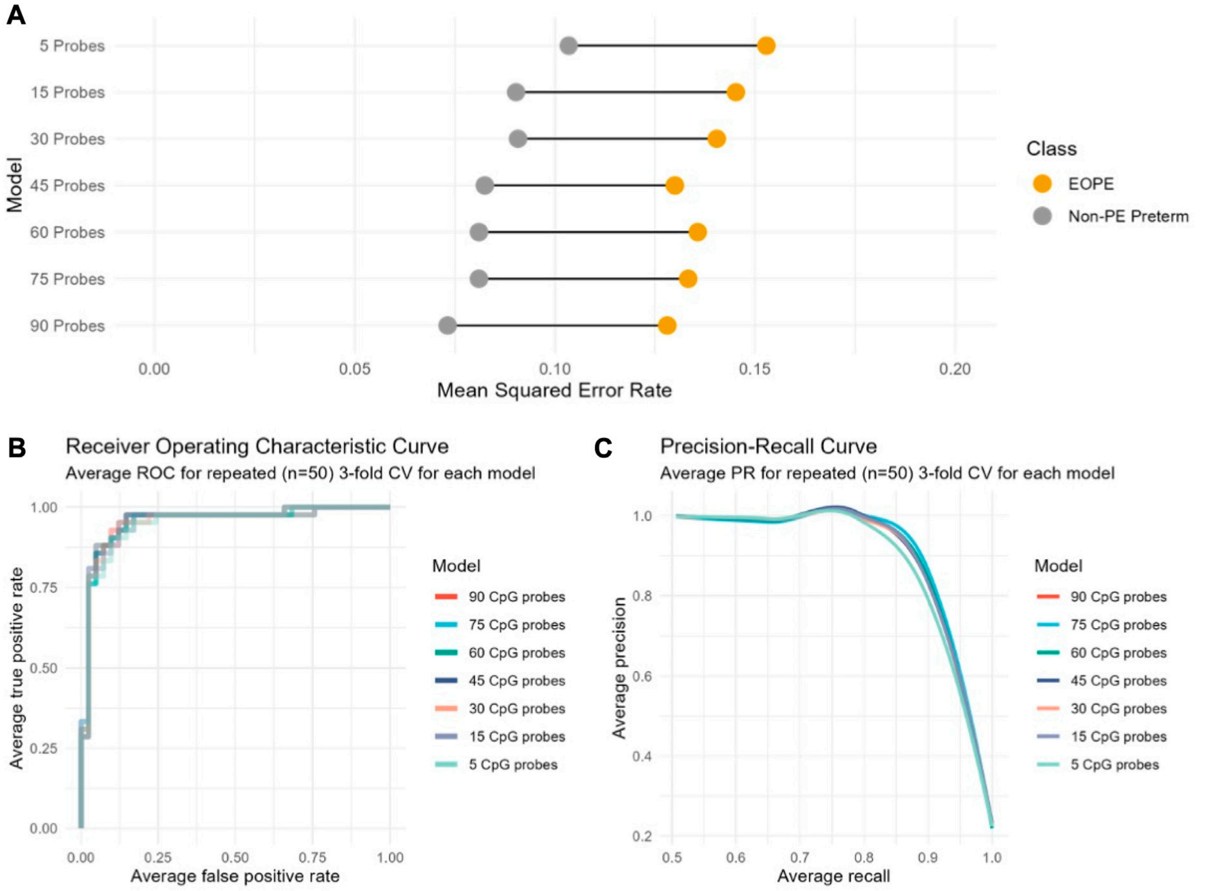

FIGURE 1
Cross-validation performance of seven models based on varying number of CpGs indicates no substantial performance difference between models regardless of the number of CpGs included. The performance of each model computed across cross-validation folds is indicated according to **(A)** to the overall classification error rate (OER); **(B)** The receiver operating characteristic (ROC) curve; **(C)** The precision-recall curve.

developed six additional models by selecting the number of CpGs *a priori* (*n* = 75, 60, 45, 30, 15, or 5) and assessed their performance on cross-validation.

Decreasing the number of CpGs did not significantly affect model performance compared to the 90-CpG model, as evaluated using two discrete classification metrics (OER, ROC-AUC) and one probability-based metric (Brier score) (Table 3). The mean OER value of the seven models with different numbers of CpGs was 0.11 (±0.01 standard deviation); we saw a slight but non-significant increase in OER as the number of CpGs decreased (Figure 1A; Table 3). For all seven models, we also assessed the area under the curve (AUC, Table 3) of the receiver operating characteristic (ROC, Figure 1B) curve, which informs about the trade-off between the true positive rate and the false positive rate. Once again, the difference between the seven models on this metric was not significant, and all seven showed outstanding discrimination with an AUC of 0.95 (Applied Logistic Regression, 2023). Lastly, we computed the average Brier score across cross-validation folds for each model, which measures the mean squared difference between the predicted probability assigned to the possible outcomes for a given sample, and its actual outcome. All seven models had an average Brier Score of 0.14 (Table 3), which confirmed that the predicted probability of most samples matched the true likelihood of the outcome of interest, i.e., most true EOPE samples had a probability of EOPE closer to 1, where 1 is EOPE and 0 is nPTB.

While a model relying on a very small number of CpGs (e.g., 5) may demonstrate a good average performance on cross-validation, it might not be as generalizable to new data. At the same time, the larger model (90 CpGs) initially selected by cross-validation does not perform significantly better than any of the smaller models (Figure 1, Table 3). Therefore, we chose a model with an intermediate number of features (45 CpGs) to evaluate on new data (Supplementary Table S2).

Lastly, we compared the average performance across cross-validation folds of the final 45 CpG model to a 45 CpG model that was trained on randomly permuted labels. The goal of this permutation test was to assess whether our final model performs better than a random model, in which case we can infer that the CpGs selected in our model are likely biologically relevant to preeclampsia. Across all metrics, it was clear that the model with randomly permuted labels performed significantly worse (0.52 OER, AUC = 0.45) as compared to our final model trained on true EOPE/nPTB labels (Table 3), confirming the biological relevance of the 45-CpG signature. The characteristics of the final 45 CpGs that

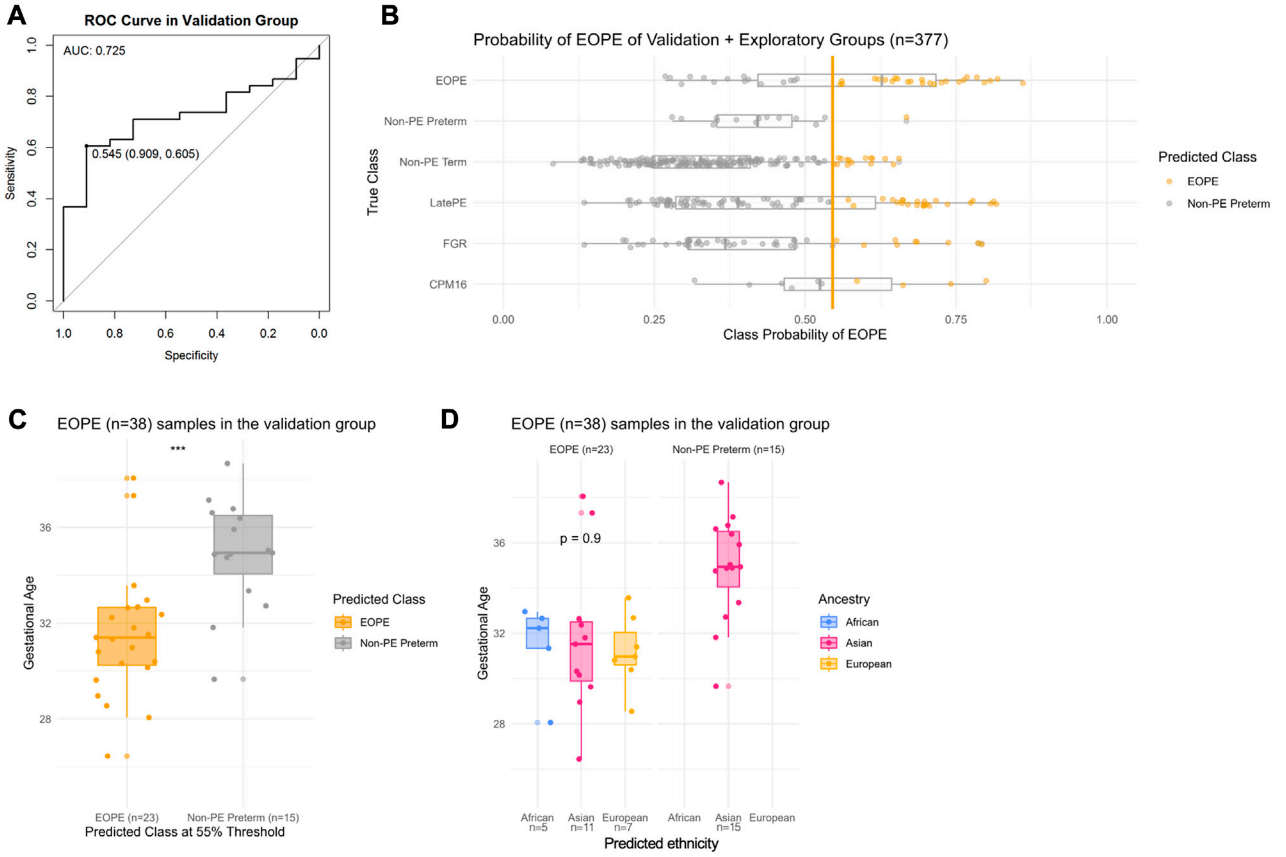

FIGURE 2
Probability of EOPE in the validation group (n = 48) and in the exploratory group (n = 328). eoPred was used to predict the outcome of BMIQ normalized samples in the validation and exploratory groups (the two groups were normalized separately). There are 11 nPTB samples, and 38 EOPE samples in the validation group. **(A)** Receiver operating characteristic (ROC) curve in the validation group. A classification threshold of 55% was selected to maximize Youden's J statistic (sensitivity + specificity—1). The diagonal line indicates the curve for a classifier that predicts the majority class in all cases. **(B)** Probability of EOPE of samples in the validation and exploratory groups. EOPE and nPTB samples in the plot belong to the validation group. **(C)** Gestational age of EOPE samples in the validation group. eoPred predicted 23 samples as EOPE and 15 as nPTB, using a 55% probability of EOPE as the classification threshold. The difference in gestational age between EOPE samples classified as EOPE and those classified as nPTB is significant. **(D)** Gestational age of EOPE samples in the validation group, faceted by ancestry groups. Ancestry was assigned to samples using planet probabilities. The probability of EOPE is not significantly different across ancestries in those samples correctly classified as EOPE.

comprise the predictive signature (genomic location, overlapping genes, etc.) are described in Supplementary Table S2.

As the performance of machine learning models can be affected by data transformations, we evaluated the performance of eoPred with 7 common normalization methods using a subset of the validation group data with IDATs available (GSE98224 and GSE125605, n = 44). These methods included beta-mixture quantile normalization (BMIQ) (Teschendorff et al., 2013), quantile normalization (QN) (Aryee et al., 2014), subset-quantile within array normalization (SWAN) (Maksimovic et al., 2012), Dasen (Pidsley et al., 2013), noob (Triche et al., 2013), QN + BMIQ, and noob + BMIQ. While our model was trained on BMIQ normalized data, it performed better on other iterations with different normalization methods (Supplementary Table S3). Specifically, eoPred applied to quantile-normalized data demonstrated the best performance across all metrics evaluated (Supplementary Table S3). However, this evaluation is strictly based on machine learning evaluation metrics and does not consider model interpretability. As our sample size is small, we cannot conclude that a given normalization method will consistently outperform others and testing on more data in future will be required to systematically assess the effect of normalization on eoPred's performance.

eoPred demonstrated good discriminatory ability (AUC = 0.73) in an independent validation group

To test the ability of eoPred to identify the placental phenotype of EOPE, we applied it to predict the disease status of 49 samples (38 EOPE, 11 nPTB) in the validation group. Each sample was assigned a probability (from 0 to 1) of being classified as either EOPE or nPTB; some were labelled with a high confidence and others were intermediate (Figure 2B). We then wanted to determine what probability cut-off would best discriminate between EOPE and nPTB. Youden's J statistic, defined as the sum of sensitivity and specificity minus one, was used to choose an optimal threshold of

TABLE 4 Characterization of late-PE and FGR cases in the exploratory cohort, by their predicted outcome. *p*-values for categorical values were calculated with Fisher's exact test. Welch's t-test was used to compute the *p*-values of the continuous variables.

	Late-preeclampsia (*n* = 86)			Fetal growth restriction (*n* = 46)		*p*
Predicted outcome	EOPE	nPTB		EOPE	nPTB	
Total N (%)	24 (28%)	62 (72%)		9 (20%)	37 (80%)	
Probability of EOPE (range)	0.70 (0.57–0.82)	0.34 (0.13–0.54)	**<2.2e−16**	0.68 (0.55–0.79)	0.35 (0.13–0.55)	**3.05E-08**
Inferred weeks of gestational age (range)	35.65 (34.02–38.80)	37.48 (34.65–40.69)	**5.41E-07**	33.37 (28.17–37.65)	37.18 (32.16–39.75)	**8.22E-05**
Placental sex (XY/XX)	14/10	31/31	0.63	4/5	18/19	1
Predicted ancestry			0.32			0.17
African	5	23		2	5	
Asian	3	9		3	5	
European	16	30		4	27	
FGR status			**0.03**			NA
Yes	4	2		9	37	
No	4	5		0	0	
Unknown	16	55		0	0	

55% to classify samples in our data (Figure 2A). This threshold can be modified by the user in the eoPred function such that users can explore the spread of probabilities in their data and determine a threshold that fits their question best. For example, higher probabilities of EOPE may be used to select samples with a more homogeneous placental phenotype, while lower probabilities might capture those placentas with sub-clinical levels of pathology.

Of the 38 EOPE samples in the validation group, 23 were predicted as EOPE (60% sensitivity) with a mean EOPE probability of 70% (Figure 2). The EOPE samples that were correctly classified as EOPE (31.9 weeks) had significantly lower average gestational age ($p = 0.0004$) than those that were classified as nPTB (35.5 weeks) (Figure 2C). In addition, of 15 EOPE samples that were misclassified as nPTB, all originated from the same dataset (GSE125605), which consisted of samples of Han Chinese ancestry (Figure 2D). In contrast, all samples from a separate dataset in the validation group, GSE98224 (16 EOPE, 5 nPTB), were correctly classified, suggesting either a dataset or ancestry effect in GSE125605. There was a wide range of EOPE probability values in the EOPE group, ranging from 56% to 82% (Figure 2B), which may reflect the spectrum of placental findings that is observed within this subtype of PE.

In the small sample size ($n = 11$) of nPTB cases in the validation group, eoPred had a high specificity of 91%. The probability of EOPE assigned to nPTB cases, 10 of which were correctly classified as nPTB using a 55% probability threshold, ranged from 24% to 53% (Figure 2B), with an average of 41%. The misclassified nPTB sample had a EOPE probability of 67%. While we do not have information on the cause for preterm birth for these samples labelled as nPTB but predicted to be EOPE, we can assume a variety of etiologies within this group—thus, the placental phenotype is likely also heterogeneous and, in some cases may share certain features with EOPE.

Applying eoPred to the validation cohort produced an area under the curve (AUC) value of 0.725, meaning that 73% of the time, for a randomly selected pair of nPTB and EOPE samples, the model will correctly assign a higher absolute EOPE risk to the sample with EOPE than to the nPTB sample. Importantly, this indicates that regardless of the varied causes of preterm birth within the nPTB group, eoPred is successful at discriminating between the placental phenotype of EOPE and that of other etiologies that may lead to preterm birth.

Lastly, we also tested the 5 CpG model in the validation group. The predictions made by eoPred (45 CpGs) and the 5 CpG model were very similar (MAE in EOPE probability between the two models = 0.4). The 5 CpG model predicted EOPE with 61% sensitivity, and correctly classified 25 EOPE samples. The 5 CpG model demonstrated 91% specificity, only misclassifying 2 more nPTB samples than eoPred (Supplementary Figure S3). The threshold chosen as optimal for the 5 CpG model using Youden's J statistic (56%) was also remarkably similar to the 55% threshold identified as optimal for the 45 CpG model.

FGR and late-PE cases with earlier gestational ages tend to be classified as EOPE

An overlap in placental findings from pregnancies affected by EOPE and those affected by FGR without preeclampsia or by late onset PE has been reported based on DNAme (Wilson et al., 2018), gene expression (Leavey et al., 2018), as well as placental pathology (Roberts and Escudero, 2012). To further explore the overlap between these conditions, we investigated the EOPE probability value for cases from pathologies that may overlap with the placental phenotype of EOPE, such as FGR and late-PE. We thus applied eoPred to 328 diverse samples in the exploratory group, which included 2 EOPE, 86 late-PE, 3 nPTB, 181 nTB, 46 FGR, and 10 CPM16 cases.

Of cases known with high confidence to be normotensive, 91% were correctly classified using the 55% probability cutoff as nPTB ($n = 184$, 181 nTB and 3 nPTB), with a mean EOPE probability of 31%. The

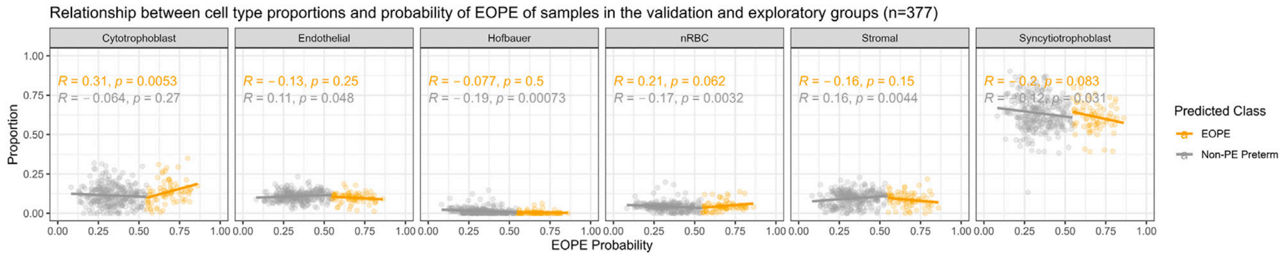

FIGURE 3
Relationship of eoPred's EOPE probability and predictive CpG signature with placental cell types. Relationship between cell type proportions and probability of EOPE of 38 EOPE samples in the validation group. Syncytiotrophoblast and trophoblast are moderately associated with the probability of EOPE of samples misclassified as nPTB.

2 EOPE cases in the exploratory group were correctly assigned a high EOPE probability of 73% and 86%, respectively (Figure 2B). In addition, 22 of 86 (28%) late-PE cases and 9 of 46 FGR (20%) cases were predicted as EOPE, which could reflect the subset of cases in these groups with similar molecular pathology to EOPE. Both late-PE and FGR cases classified as EOPE had significantly lower gestational ages than their counterparts predicted to be normotensive (Table 4), with late-PE cases classified as EOPE having an average GA of 35.65 weeks and nPTB classified as EOPE having an average GA of 33.37 weeks. Moreover, of the 6 late-PE cases that presented with co-occurring FGR (which is commonly associated with EOPE), 4 were classified as EOPE with a mean EOPE probability of 69%, whereas 2 were predicted to be normotensive with 54% and 28% EOPE probability, respectively. Late-PE + FGR cases were more likely to be predicted EOPE than nPTB (mean p(EOPE) = 69%); but were not significantly more likely to be predicted EOPE than late-PE samples without FGR (mean P(EOPE) = 41%, p-value = 0.15).

Confined placental mosaicism for chromosome 16 (CPM16) is highly associated with EOPE and almost always results in FGR. DNAme patterns of CPM16 placentae were previously reported to overlap with those observed in EOPE (Yong et al., 2003). Using the same data as in (Yong et al., 2003), we evaluated how eoPred would classify CPM16 samples (n = 10). Five of the 10 CPM16 samples were also diagnosed as EOPE; eoPred correctly predicted four of these samples as EOPE, while the fifth fell just below our 55% probability cut-off. All 5 CPM16 samples in which PE had not been diagnosed were correctly classified as nPTB.

Influence of biological variables on eoPred

Several variables have been reported to influence the placental DNA methylome throughout gestation, including cell composition and genetic ancestry (Khan et al., 2023). We therefore sought to investigate the effect of placental cell composition on eoPred CpGs, tissue-specificity of eoPred, and the effect of genetic ancestry on eoPred CpGs as well as the EOPE estimates.

There were significant differences (Bonferroni adjusted p-value < 0.05) in the PlaNET-predicted proportion of stromal and Hofbauer cells between the original author labels of EOPE and nPTB in the training data, and in the proportion of stromal but not Hofbauer cells by EOPE status in the author-labelled validation data (Supplementary Figure S4). However, none of the 45 predictive CpGs overlapped with any of the first or third trimester cell-type specific CpGs used by PlaNET to estimate cell composition, suggesting that these CpGs are not strongly driven by cell composition differences and should be robust to minor cell composition variation arising from sampling differences. The eoPred-derived probability of EOPE was not strongly and significantly associated with any cell type proportion (R > 0.3 and p-value<0.05), with the exception of cytotrophoblast, which was moderately (R = 0.31, p = 0.005) associated with the probability of EOPE in samples in the validation and exploratory groups predicted as EOPE (Figure 3). Overall, these results suggest a limited effect of cell composition on eoPred predictions.

We also used the placental methylome browser (Yuan, 2021) to assess the cell-specific DNAme signature of the top five most predictive CpGs (cg10994126, cg26787199, cg14605117, cg10246581) in the model (i.e., those that were selected as most stable in the cross-validation process). While the cell types did display differences in average DNAme at these eoPred CpGs, there were no consistent changes: some CpGs had higher DNAme in trophoblast cells than other cell types, whereas Hofbauer cells had the highest DNAme at other eoPred CpGs (Supplementary Figure S5). These results further confirm the absence of a consistent cell composition effect at the most predictive eoPred CpGs.

Moreover, PE-associated DNAme alterations have been reported both in cord blood and placenta, with little overlap between CpGs reported to be differentially methylated in the two, likely due to tissue-specificity of DNAme. We thus applied eoPred to 110 cord blood samples (including EOPE, LOPE, nPTB and nTB) to test whether the 45 CpG sites in our placental model could predict disease status in cord blood. All 110 samples were predicted as nPTB (average probability of 93%), demonstrating that the DNAme signature of EOPE developed in this study is specific to placental tissue.

Ancestry probabilities computed with PlaNET indicated that 65% of samples in the training group and 61% in the validation and exploratory groups were primarily of European ancestry (>75% probability) (Supplementary Figure S6). We measured the strength of association between the probability of EOPE and each of the three ancestry probabilities using Spearman's rank order correlation for all samples in the validation and exploratory cohorts. None of the three PlaNET ancestry probabilities (African/Asian/European) were strongly associated with the probability of EOPE (ρ < 0.2).

We then compared how the β values at the 45 predictive CpGs varied by ancestry within the EOPE and nPTB groups. Ancestry was not

TABLE 5 Average β values at 45 predictive CpGs, by dataset and by ancestry. The mean β at the 45 predictive CpGs was computed for EOPE and nPTB samples separately, using author-assigned labels. Δβ is defined as the absolute difference between EOPE mean β and nPTB mean β. *p*-values were computed with an unpaired t-test.

	EOPE mean β (N)	nPTB mean β (N)	Δβ	*p*-value
Training group (N = 83)				
All samples	0.47 (42)	0.61 (41)	0.14	< 2.2e-16
By inferred ancestry				
Asian ancestry	0.49 (7)	0.62 (6)	0.13	< 2.2e-16
African ancestry	0.48 (7)	0.62 (5)	0.14	< 2.2e-16
European ancestry	0.48 (28)	0.62 (30)	0.14	< 2.2e-16
By dataset				
GSE100197	0.48 (22)	0.62 (28)	0.14	< 2.2e-16
GSE103253	0.47 (11)	0.61 (8)	0.15	< 2.2e-16
GSE57767	0.52 (5)	0.59 (3)	0.07	0.000408
GSE73375	0.38 (4)	0.48 (2)	0.1	3.57E-06
Validation (N = 49)				
All samples	0.49 (38)	0.55 (11)	0.06	5.73E-10
By inferred ancestry				
Asian ancestry	0.51 (26)	0.55 (2)	0.04	0.06439
African ancestry	0.43 (5)	0	NA	NA
European ancestry	0.43 (7)	0.55 (9)	0.12	< 2.2e-16
By dataset				
GSE98224	0.42 (16)	0.56 (5)	0.14	< 2.2e-16
GSE125605	0.53 (22)	0.54 (1)	0.006	0.8235
GSE203396	0	0.53 (5)	NA	NA

associated with DNAme at 43 of the 45 (96%) eoPred CpGs in nPTB, suggesting that these CpGs are not systematically affected by genetic ancestry (Supplementary Figure S7). However, ancestry was associated with CpG methylation within the author–assigned EOPE group at 35 of the 45 CpGs. While cohort was confounded with ancestry, differences in diagnostic criteria between cohorts might explain this effect. Interestingly, when using the eoPred-predicted groups, rather than author-assigned diagnosis, the differences at these 45 CpGs across ancestries were much smaller (Supplementary Figure S7, Supplementary Figure S8). In addition, none of the 45 predictive CpG probes in eoPred have been reported as influenced by nearby SNPs in the placenta (Delahaye et al., 2018).

Lastly, we explored how eoPred's predictions may be affected by ancestry (European, Asian, and African) by assessing the average DNAme β values of EOPE versus nPTB samples in the training and validation cohorts at the 45 predictive CpGs (Table 5). DNAme was, as expected, significantly lower in EOPE compared to nPTB samples (p-value < 0.05) within each ancestry group, and within each of the datasets included in the training and validation groups, with the exception of GSE125605 (Table 5). The mean β of EOPE samples in this dataset, which is comprised exclusively of samples of Asian ancestry, was also higher than all the other datasets in both the training and validation groups (Table 4). In contrast, for the training group, the mean β of EOPE samples of Asian ancestry was very similar to those of European ancestry (Table 4), and the mean β value of the 4 EOPE samples of Asian ancestry from GSE98224 was also similar (0.40). This suggests that the misclassification of a greater proportion of EOPE samples in GSE125605 compared to GSE98224 may be due to a dataset effect rather than to genetic ancestry. We confirmed that there were no major cell composition differences due to sampling between GSE125605 and GSE98224, which could have been a potential source for this dataset effect (Supplementary Figure S9).

Discussion

Previous studies of placental DNAme have shown that placentae from pregnancies affected by preeclampsia exhibit DNA methylome differences as compared to placentae from uncomplicated pregnancies (Blair et al., 2013; Yeung et al., 2016; Leavey et al., 2018; Wilson et al., 2018). We used placental DNAme to develop a supervised model, "eoPred," to classify placentae according to their DNAme patterns into those with and without EOPE. This tool can be applied by users to define i) homogeneous subgroups of placentae with an EOPE-associated DNAme phenotypes, rather than relying on clinical measures for group definition or ii) assigning a probability score allowing the phenotype to be assessed on a continuous scale.

We anticipate that one of the major values of eoPred will be that it can be used by researchers to identify placentae that share a similar placental phenotype with EOPE, even if the pregnant parent was not diagnosed with PE. For instance, pregnancy complications such as FGR or preterm birth, as well as environmental conditions during gestation (i.e., smoking, stress) are associated with an increased risk of developing PE (Lisonkova and Joseph, 2013; Bartsch et al., 2016). DNAme patterns have been reported to be similar (Wilson et al., 2018) between samples with EOPE and those with co-occurring late-PE and FGR diagnosed between 34 and 36 weeks; an overlap between the nPTB and EOPE phenotypes is also reflected by the rate at which nPTB samples were predicted to be EOPE by eoPred. We thus hypothesize that eoPred will be helpful in investigating the extent to which such aforementioned environmental exposures affect the placental DNA methylome at different degrees and time points in gestation. This will broaden the scope of available studies beyond current analysis limitations of clinically-confirmed and reported cases of PE and provide more insight into the pathogenesis of PE. Lastly, eoPred has the potential to generate translational impact in the development of early-diagnostic tools based on cell-free placental DNA methylation: as non-invasive methods are developed to assess fetal health during gestation (e.g., placental-derived cell free DNA and DNA methylation in maternal blood) the utility of the signature of EOPE described in our study can be assessed (Kwak et al., 2020; Palei, 2021; Moufarrej et al., 2022).

The relatively low sensitivity of our model (60%) primarily arose from one of the two datasets in the validation group. All samples in one of our validation datasets (GSE98224) were correctly classified, while in the other (GSE125605), which is comprised exclusively of samples of Asian ancestry, 15/22 EOPE samples were misclassified as nPTB. Notably, the 7 samples in this dataset which were correctly predicted as EOPE had a significantly lower GA than those predicted as nPTB. Without further independent clinical data, we cannot be sure that the diagnostic criteria were identical in each dataset. The interpretation of this finding is currently limited by the fact that all 15 misclassified samples are affected by the confounding effects of dataset, ancestry, and gestational age. Our finding that 1) the 45 predictive CpGs do not vary by ancestry within normotensive preterm samples and 2) there is a significant difference in β values at the 45 predictive CpGs between EOPE and nPTB samples in samples of Asian ancestry from datasets other than GSE125605, both in the training and validation groups, suggests that the high misclassification rate in this dataset may not be a result of ancestry but rather a dataset effect (e.g., systemic differences in how cases were ascertained and classified as EOPE or in sample and array processing). However, European ancestry is overrepresented in a majority of placental DNAme datasets, and more genetically diverse data is needed to appropriately assess the utility of eoPred in different populations. Ideally, a new iteration of the tool can be built with a larger and more diverse sample size to further investigate this.

While more validation is needed, many of the top CpGs selected as predictive by our model are relevant to preeclampsia. For example, there were 3 eoPred CpGs in the *PAPPA*-2 gene in chromosome 1, which were some of most stable features in our model. *PAPPA*-2 is highly expressed in the placenta (Wang et al., 2009; Sifakis et al., 2018) and has been reported to be upregulated in cases of severe PE (Macintire et al., 2014; Kramer et al., 2016). In addition, two predictive CpGs mapped to each of *SYDE1* and *FLNB*, which have also been reported as potentially involved in FGR and PE (Tejera et al., 2013; Lo et al., 2017; Leavey et al., 2018; Huang et al., 2022). In addition, the model demonstrated a high specificity in both validation and exploratory groups, where 171 of 195 normotensive samples were predicted correctly, adding confidence to the utility of eoPred.

The performance of machine learning models on new data may be affected by data transformations (e.g., normalization), particularly if these differ from the transformations that were applied to the data that the model was trained on. We found that that the prediction ability of eoPred is improved with quantile normalization, while our training data was normalized using BMIQ. The better performance of eoPred on quantile normalized data may be explained by the fact that quantile normalization is a between-array method, which means it considers all samples in the data, and transforms the data to minimize global DNAme differences between samples (Hicks et al., 2018). As such, quantile normalization could be correcting for dataset-associated differences in a similar way to batch correction, thus reducing noise arising from technical variation, which could improve the performance of eoPred. As gestational age was lower in EOPE samples classified as EOPE than in those classified as nPTB, it is also possible that EOPE samples with a low EOPE probability have a milder placental phenotype that is more similar to that of late-onset preeclampsia. However, given the overrepresentation of samples from GSE125605 among the misclassifications, this is difficult to assess. We encourage users to make an informed choice on what the best normalization method is in the context of their data, and if choosing a method other than BMIQ, we recommend comparing eoPred's predictions to those achieved in BMIQ-normalized data.

Placental DNAme from whole tissue is a composite of cell-specific DNAme patterns, and while placental cell composition often varies between cohorts due to sampling differences, it has also been proposed to vary in preeclamptic placentae for biological reasons (Yuan et al., 2021; Campbell et al., 2023). Furthermore, preeclampsia-associated differentially expressed genes in cell-free RNA were reported to be tissue- or cell-type specific (Moufarrej et al., 2022). Although EOPE samples in our training and validation groups had less predicted stromal and Hofbauer cells compared to nPTB, the magnitude of change was small, and no difference was observed for the two trophoblast subtypes, which make up the bulk of the placental (chorionic villus) sample. Furthermore, DNAme variation at the 45 EOPE predictive CpGs did not appear to be driven by cell composition. We observed that the probability of EOPE was moderately associated with increased cytotrophoblast estimates particularly within samples predicted as EOPE in the validation and exploratory groups, but not at any of the other cell types. Overall, the EOPE-associated DNAme signature used by eoPred does not seem to be driven by cell composition.

There are several strengths and limitations inherent to how we have constructed eoPred given the available data and how it was clinically characterized. The Illumina HM450K array is the most common platforms used for placental studies of PE, and eight Illumina HM450K publicly available on GEO were used to train, validate, and explore the utility of our model. Our use of the HM450K platform in this study will increase eoPred's relevance to future EWAS studies, as will our choice to train the model on CpGs that are also present on the EPIC array, which will enable the use of eoPred with new datasets as they are generated. While the HM450K array only measures DNA methylation at ~1% of all CpGs in the genome, there are widespread changes in PE at several regions covered by this array. Secondly, re-utilization of public data is essential to maximize research time and resources, but it comes with limitations when the purpose is to ask new questions. In compliance with privacy and confidentiality agreements, public data often lacks extensive clinical and/or phenotypical characterization of samples. This limited our ability to interpret our findings; in particular, we would like to investigate whether the probability of EOPE is associated with severity of clinical measures and/or placental histopathology, and whether the placental samples from other reported etiologies with high EOPE probability exhibited characteristic pathological features also seen in EOPE. More generally, public datasets of chorionic villi samples of various etiologies are sparse, which limited the design of our predictor. As suggested by Myatt et al. (Leslie et al., 2014), comparing PE to "normal" outcomes will inevitably lead to false predictive capability, begging the question of whether PE will be adequately identified as distinct from other outcomes, both normal and abnormal. While this is certainly a limitation of our study, we chose a binary classification model as opposed to a multiclass model due to the size constraints of our data, and we developed eoPred as a tool with research utility in mind, with the aim that its application to future datasets may provide insight into the pathophysiology of PE in the placenta, and other associated outcomes.

Conclusion

In this study, we develop eoPred, which is a tool that can be applied to DNAme data to classify placentae as likely to be EOPE or nPTB, as well as to output a continuous measure of the placental phenotype of EOPE. This tool has been included in the PlaNET R package and is available at https://www.bioconductor.org/packages/release/bioc/html/planet.html. We anticipate that eoPred will be useful in future EWAS studies to measure the influence of EOPE on the placental DNA methylome, and to study the effect of other variables of interest such as maternal exposures, chromosomal sex, and stress on a molecularly homogeneous subtype of preeclampsia. Importantly, we confirmed that this measure of EOPE probability is robust to cell composition differences that can arise from sampling and/or pathology, and that it works adequately with a variety of DNAme normalization methods. Lastly, the 45 CpG DNAme signature of EOPE appears to be shared with other pregnancy complications that have some overlapping clinical and placental phenotypes with EOPE, illustrating the overlap between pregnancy complications mediated by the placenta.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: eoPred is available as a function in the PlaNET R package (https://www.bioconductor.org/packages/release/bioc/html/planet.html). All datasets used in this study are publicly available via the Gene Expression Omnibus, with dataset accession numbers GSE100197, GSE98224, GSE125605, GSE110829, GSE73375, GSE75196, GSE57767, GSE103253 and GSE120981. The 45 EOPE-specific CpGs selected in our model can be found in Supplementary Table S2. Source code and R Markdown scripts with the processing and analysis pipeline can be found at https://github.com/iciarfernandez/eoPred/.

Ethics statement

The studies involving humans were approved by BC Women's and Children's Hospital ethics committees in Vancouver, BC, Canada (H04–704488). The studies were conducted in accordance with the local legislation and institutional requirements.

Author contributions

WR, AI, and VY contributed to project design. IF-B performed analysis of data and wrote manuscript. AI processed gene expression data. VY helped develop eoPred functions for the PlaNET package. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by a Canadian Institutes of Health Research (CIHR) grant to WPR (F19-04091). WPR holds a CIHR Research Chair in Sex and Gender Science (GSK-171375) and an investigatorship award from the BC Children's Hospital Research Institute. IFB receives support from a UBC Four-Year Fellowship.

Acknowledgments

We thank pregnant people who generously choose to donate their placentae for the advancement of research, and the scientific community for their commitment to making data publicly available. This work would not be possible without current and past members of the Robinson lab, we especially acknowledge Dr. Maria Peñaherrera A. for thoughtful discussion and feedback on the design and data analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1248088/full#supplementary-material

References

Abalos, E., Cuesta, C., Grosso, A. L., Chou, D., and Say, L. (2013). Global and regional estimates of preeclampsia and eclampsia: a systematic review. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 170, 1–7. doi:10.1016/j.ejogrb.2013.05.005

Applied Logistic Regression (2023). Available at: https://learning.oreilly.com/library/view/applied-logistic-regression/9781118548356/ (Accessed April 21, 2023). 3rd Edition.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369. doi:10.1093/bioinformatics/btu049

Bartsch, E., Medcalf, K. E., Park, A. L., and Ray, J. G. (2016). Clinical risk factors for pre-eclampsia determined in early pregnancy: systematic review and meta-analysis of large cohort studies. *BMJ* 353, i1753. doi:10.1136/bmj.i1753

Benton, S. J., Leavey, K., Grynspan, D., Cox, B. J., and Bainbridge, S. A. (2018). The clinical heterogeneity of preeclampsia is related to both placental gene expression and placental histopathology. *Am. J. Obstet. Gynecol.* 219, 604. e1–e604. doi:10.1016/j.ajog.2018.09.036

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* 98, 288–295. doi:10.1016/j.ygeno.2011.07.007

Blair, J. D., Langlois, S., McFadden, D. E., and Robinson, W. P. (2014). Overlapping DNA methylation profile between placentas with trisomy 16 and early-onset preeclampsia. *Placenta* 35, 216–222. doi:10.1016/j.placenta.2014.01.001

Blair, J. D., Yuen, R. K. C., Lim, B. K., McFadden, D. E., von Dadelszen, P., and Robinson, W. P. (2013). Widespread DNA hypomethylation at gene enhancer regions in placentas associated with early-onset pre-eclampsia. *Mol. Hum. Reprod.* 19, 697–708. doi:10.1093/molehr/gat044

Brown, M. A., Magee, L. A., Kenny, L. C., Karumanchi, S. A., McCarthy, F. P., Saito, S., et al. (2018). The hypertensive disorders of pregnancy: ISSHP classification, diagnosis & management recommendations for international practice. *Pregnancy Hypertens.* 13, 291–310. doi:10.1016/j.preghy.2018.05.004

Campbell, K. A., Colacino, J. A., Puttabyatappa, M., Dou, J. F., Elkin, E. R., Hammoud, S. S., et al. (2023). Placental cell type deconvolution reveals that cell proportions drive preeclampsia gene expression differences. *Commun. Biol.* 6, 264–315. doi:10.1038/s42003-023-04623-6

Chu, T., Bunce, K., Shaw, P., Shridhar, V., Althouse, A., Hubel, C., et al. (2014). Comprehensive analysis of preeclampsia-associated DNA methylation in the placenta. *PLOS ONE* 9, e107318. doi:10.1371/journal.pone.0107318

Cirkovic, A., Garovic, V., Milin Lazovic, J., Milicevic, O., Savic, M., Rajovic, N., et al. (2020). Systematic review supports the role of DNA methylation in the pathophysiology of preeclampsia: a call for analytical and methodological standardization. *Biol. Sex. Differ.* 11, 36. doi:10.1186/s13293-020-00313-8

Delahaye, F., Do, C., Kong, Y., Ashkar, R., Salas, M., Tycko, B., et al. (2018). Genetic variants influence on the placenta regulatory landscape. *PLOS Genet.* 14, e1007785. doi:10.1371/journal.pgen.1007785

Edgar, R. D., Jones, M. J., Robinson, W. P., and Kobor, M. S. (2017). An empirically driven data reduction method on the human 450K methylation array to remove tissue specific non-variable CpGs. *Clin. Epigenetics* 9, 11. doi:10.1186/s13148-017-0320-z

Heiss, J. A., and Just, A. C. (2018). Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. *Clin. Epigenetics* 10, 73. doi:10.1186/s13148-018-0504-1

Herzog, E. M., Eggink, A. J., Willemsen, S. P., Slieker, R. C., Wijnands, K. P. J., Felix, J. F., et al. (2017). Early- and late-onset preeclampsia and the tissue-specific epigenome of the placenta and newborn. *Placenta* 58, 122–132. doi:10.1016/j.placenta.2017.08.070

Hicks, S. C., Okrah, K., Paulson, J. N., Quackenbush, J., Irizarry, R. A., and Bravo, H. C. (2018). Smooth quantile normalization. *Biostatistics* 19, 185–198. doi:10.1093/biostatistics/kxx028

Huang, S., Cai, S., Li, H., Zhang, W., Xiao, H., Yu, D., et al. (2022). Prediction of differentially expressed genes and a diagnostic signature of preeclampsia via integrated bioinformatics analysis. *Dis. Markers* 2022, 5782637. doi:10.1155/2022/5782637

Inkster, A. M., Konwar, C., Peñaherrera, M. S., Brain, U., Khan, A., Price, E. M., et al. (2022). Profiling placental DNA methylation associated with maternal SSRI treatment during pregnancy. *Sci. Rep.* 12, 22576. doi:10.1038/s41598-022-26071-8

Khan, A., Inkster, A. M., Peñaherrera, M. S., King, S., Kildea, S., Oberlander, T. F., et al. (2023). The application of epiphenotyping approaches to DNA methylation array studies of the human placenta. *Res. Sq.* rs.3, rs-3069705. doi:10.21203/rs.3.rs-3069705/v1

Kim, J.-H., Cheong, H. S., Lee, D. S., Shin, H. D., and Kim, Y. N. (2017). Genome-wide DNA methylation profiles of maternal peripheral blood and placentas: potential risk factors for preeclampsia and validation of GRK5. *Genes Genom* 10, 197–206. doi:10.1007/s13258-016-0486-3

Kramer, A. W., Lamale-Smith, L. M., and Winn, V. D. (2016). Differential expression of human placental PAPP-A2 over gestation and in preeclampsia. *Placenta* 37, 19–25. doi:10.1016/j.placenta.2015.11.004

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi:10.18637/jss.v028.i05

Kwak, D. W., Kim, S. Y., Kim, H. J., Lim, J. H., Kim, Y.-H., and Ryu, H. M. (2020). Maternal total cell-free DNA in preeclampsia with and without intrauterine growth restriction. *Sci. Rep.* 10, 11848. doi:10.1038/s41598-020-68842-1

Lê Cao, K.-A., Boitard, S., and Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinforma.* 12, 253. doi:10.1186/1471-2105-12-253

Leavey, K., Benton, S. J., Grynspan, D., Kingdom, J. C., Bainbridge, S. A., and Cox, B. J. (2016). Unsupervised placental gene expression profiling identifies clinically relevant subclasses of human preeclampsia. *Hypertens. Dallas Tex* 68, 137–147. doi:10.1161/HYPERTENSIONAHA.116.07293

Leavey, K., Wilson, S. L., Bainbridge, S. A., Robinson, W. P., and Cox, B. J. (2018). Epigenetic regulation of placental gene expression in transcriptional subtypes of preeclampsia. *Clin. Epigenetics* 10, 28. doi:10.1186/s13148-018-0463-6

Lee, Y., Choufani, S., Weksberg, R., Wilson, S. L., Yuan, V., Burt, A., et al. (2019). Placental epigenetic clocks: estimating gestational age using placental DNA methylation levels. *Aging* 11, 4238–4253. doi:10.18632/aging.102049

Leek, J. T., Taub, M. A., and Rasgon, J. L. (2012). A statistical approach to selecting and confirming validation targets in -omics experiments. *BMC Bioinforma.* 13, 150. doi:10.1186/1471-2105-13-150

Leslie, M., Redman Christopher, W., Staff, A. C., Stefan, H., Wilson Melissa, L., Hannele, L., et al. (2014). Strategy for standardization of preeclampsia research study design. *Hypertension* 63, 1293–1301. doi:10.1161/HYPERTENSIONAHA.113.02664

Lim, J. H., Kang, Y.-J., Bak, H. J., Kim, M. S., Lee, H. J., Kwak, D. W., et al. (2020). Epigenome-wide DNA methylation profiling of preeclamptic placenta according to severe features. *Clin. Epigenetics* 12, 128. doi:10.1186/s13148-020-00918-1

Lisonkova, S., and Joseph, K. S. (2013). Incidence of preeclampsia: risk factors and outcomes associated with early-versus late-onset disease. *Am. J. Obstet. Gynecol.* 209, 544. e1–e544. doi:10.1016/j.ajog.2013.08.019

Lo, H.-F., Tsai, C.-Y., Chen, C.-P., Wang, L.-J., Lee, Y.-S., Chen, C.-Y., et al. (2017). Association of dysfunctional synapse defective 1 (SYDE1) with restricted fetal growth - SYDE1 regulates placental cell migration and invasion. *J. Pathol.* 241, 324–336. doi:10.1002/path.4835

Macintire, K., Tuohey, L., Ye, L., Palmer, K., Gantier, M., Tong, S., et al. (2014). PAPPA2 is increased in severe early onset pre-eclampsia and upregulated with hypoxia. *Reprod. Fertil. Dev.* 26, 351–357. doi:10.1071/RD12384

Magee, L. A., Pels, A., Helewa, M., Rey, E., Dadelszen, P. V., Magee, L. A., et al. (2014). Diagnosis, evaluation, and management of the hypertensive disorders of pregnancy: executive summary. *J. Obstet. Gynaecol. Can.* 36, 416–441. doi:10.1016/S1701-2163(15)30588-0

Maksimovic, J., Gordon, L., and Oshlack, A. (2012). Swan: subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.* 13, R44. doi:10.1186/gb-2012-13-6-r44

Martin, E., Ray, P. D., Smeester, L., Grace, M. R., Boggess, K., and Fry, R. C. (2015). Epigenetics and preeclampsia: defining functional epimutations in the preeclamptic placenta related to the TGF-β pathway. *PloS One* 10, e0141294. doi:10.1371/journal.pone.0141294

Monteagudo-Sánchez, A., Sánchez-Delgado, M., Mora, J. R. H., Santamaría, N. T., Gratacós, E., Esteller, M., et al. (2019). Differences in expression rather than methylation at placenta-specific imprinted loci is associated with intrauterine growth restriction. *Clin. Epigenetics* 11, 35. doi:10.1186/s13148-019-0630-4

Moufarrej, M. N., Vorperian, S. K., Wong, R. J., Campos, A. A., Quaintance, C. C., Sit, R. V., et al. (2022). Early prediction of preeclampsia in pregnancy with cell-free RNA. *Nature* 602, 689–694. doi:10.1038/s41586-022-04410-z

Myatt, L., Muralimanoharan, S., and Maloyan, A. (2014). Effect of preeclampsia on placental function: influence of sexual dimorphism, microRNA's and mitochondria. *Adv. Fetal Neonatal Physiol.* 814, 133–146. doi:10.1007/978-1-4939-1031-1_12

Palei, A. C. (2021). Cell-free DNA as a potential biomarker for preeclampsia. *Expert Rev. Mol. diagn.* 21, 1253–1256. doi:10.1080/14737159.2021.1997593

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 14, 293. doi:10.1186/1471-2164-14-293

Price, M. E., Cotton, A. M., Lam, L. L., Farré, P., Emberly, E., Brown, C. J., et al. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* 6, 4. doi:10.1186/1756-8935-6-4

Redman, C. W. (2017). Early and late onset preeclampsia: two sides of the same coin. *Pregnancy Hypertens. Int. J. Womens cardiovasc. Health* 7, 58. doi:10.1016/j.preghy.2016.10.011

Redman, C. W., Sargent, I. L., and Staff, A. C. (2014). IFPA senior award lecture: making sense of pre-eclampsia – two placental causes of preeclampsia? *Placenta* 35, S20–S25. doi:10.1016/j.placenta.2013.12.008

Renbo, L. (2016). Cross-validation for model selection. Available at: https://github.com/ludvigolsen/cvms.

Roberts, J. M., and Escudero, C. (2012). The placenta in preeclampsia. *Pregnancy Hypertens.* 2, 72–83. doi:10.1016/j.preghy.2012.01.001

Roberts, J. M., and Hubel, C. A. (2009). The two stage model of preeclampsia: variations on the theme. *Placenta* 30, S32–S37. doi:10.1016/j.placenta.2008.11.009

Robinson, W. P., and Price, E. M. (2015). The human placental methylome. *Cold Spring Harb. Perspect. Med.* 5, a023044. doi:10.1101/cshperspect.a023044

Rohart, F., Gautier, B., Singh, A., and Cao, K.-A. L. (2017). mixOmics: an R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* 13, e1005752. doi:10.1371/journal.pcbi.1005752

Sifakis, S., Androutsopoulos, V. P., Pontikaki, A., Velegrakis, A., Papaioannou, G. I., Koukoura, O., et al. (2018). Placental expression of PAPPA, PAPPA-2 and PLAC-1 in pregnacies is associated with FGR. *Mol. Med. Rep.* 17, 6435–6440. doi:10.3892/mmr.2018.8721

Silva, L. M., Coolman, M., Steegers, E. A., Jaddoe, V. W., Moll, H. A., Hofman, A., et al. (2008). Low socioeconomic status is a risk factor for preeclampsia: the generation R study. *J. Hypertens.* 26, 1200–1208. doi:10.1097/HJH.0b013e3282fcc36e

Staff, A. C., and Redman, C. W. G. (2018). "The differences between early- and late-onset pre-eclampsia," in Preeclampsia: basic, genomic, and clinical *comprehensive gynecology and obstetrics*. Editor S. Saito (Singapore: Springer), 157–172. doi:10.1007/978-981-10-5891-2_10

Staff, A. C. (2019). The two-stage placental model of preeclampsia: an update. *J. Reprod. Immunol.* 134, 1–10. –135. doi:10.1016/j.jri.2019.07.004

Tejera, E., Bernardes, J., and Rebelo, I. (2013). Co-expression network analysis and genetic algorithms for gene prioritization in preeclampsia. *BMC Med. Genomics* 6, 51. doi:10.1186/1755-8794-6-51

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., et al. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinforma. Oxf. Engl.* 29, 189–196. doi:10.1093/bioinformatics/bts680

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

Tsegaselassie, W., Marion, O., Shrestha, D., Wu, J., Grantz Katherine, L., and Tekola-Ayele, F. (2020). Differential DNA methylation in placenta associated with maternal blood pressure during pregnancy. *Hypertension* 75, 1117–1124. doi:10.1161/HYPERTENSIONAHA.119.14509

Wang, J., Qiu, Q., Haider, M., Bell, M., Gruslin, A., and Christians, J. K. (2009). Expression of pregnancy-associated plasma protein A2 during pregnancy in human and mouse. *J. Endocrinol.* 202, 337–345. doi:10.1677/JOE-09-0136

Wang, L. Q., Fernandez-Boyano, I., and Robinson, W. P. (2022). Genetic variation in placental insufficiency: what have we learned over time? *Front. Cell Dev. Biol.* 10, 1038358. doi:10.3389/fcell.2022.1038358

Wang, T., Xiang, Y., Zhou, X., Zheng, X., Zhang, H., Zhang, X., et al. (2019). Epigenome-wide association data implicate fetal/maternal adaptations contributing to clinical outcomes in preeclampsia. *Epigenomics* 11, 1003–1019. doi:10.2217/epi-2019-0065

Wang, W., Xie, X., Yuan, T., Wang, Y., Zhao, F., Zhou, Z., et al. (2021). Epidemiological trends of maternal hypertensive disorders of pregnancy at the global, regional, and national levels: a population-based study. *BMC Pregnancy Childbirth* 21, 364. doi:10.1186/s12884-021-03809-2

Williams, P. J., and Broughton Pipkin, F. (2011). The genetics of pre-eclampsia and other hypertensive disorders of pregnancy. *Best. Pract. Res. Clin. Obstet. Gynaecol.* 25, 405–417. doi:10.1016/j.bpobgyn.2011.02.007

Wilson, S. L., Leavey, K., Cox, B. J., and Robinson, W. P. (2018). Mining DNA methylation alterations towards a classification of placental pathologies. *Hum. Mol. Genet.* 27, 135–146. doi:10.1093/hmg/ddx391

Yan, Y. (2017). Machine learning evaluation metrics. Available at: http://github.com/yanyachen/MLmetrics.

Yeung, K. R., Chiu, C. L., Pidsley, R., Makris, A., Hennessy, A., and Lind, J. M. (2016). DNA methylation profiles in preeclampsia and healthy control placentas. *Am. J. Physiol.-Heart Circ. Physiol.* 310, H1295–H1303. doi:10.1152/ajpheart.00958.2015

Yong, P. J., Barrett, I. J., Kalousek, D. K., and Robinson, W. P. (2003). Clinical aspects, prenatal diagnosis, and pathogenesis of trisomy 16 mosaicism. *J. Med. Genet.* 40, 175–182. doi:10.1136/jmg.40.3.175

Yuan, V., Hui, D., Yin, Y., Peñaherrera, M. S., Beristain, A. G., and Robinson, W. P. (2021). Cell-specific characterization of the placental methylome. *BMC Genomics* 22, 6. doi:10.1186/s12864-020-07186-6

Yuan, V. (2023). *Placental DNA methylation analysis tools.* doi:10.18129/B9.bioc.planet

Yuan, V. (2021). Placental methylome browser. Available at: https://robinsonlab.shinyapps.io/Placental_Methylome_Browser/.

Yuan, V., Price, E. M., Del Gobbo, G., Mostafavi, S., Cox, B., Binder, A. M., et al. (2019). Accurate ethnicity prediction from placental DNA methylation data. *Epigenetics Chromatin* 12, 51. doi:10.1186/s13072-019-0296-3

Zhou, W., Laird, P. W., and Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* 45, e22. doi:10.1093/nar/gkw967