DNA Methylation Ancestry Predictor on Placenta 450k data

Victor Yuan[1], Ming Wan[2],Michael Yuen[3], Nivretta Thatra[4], Anni Zhang[1]
[1]Department of Genome Science and Technology, University of British Columbia, [2]Department of Statistics, University of British Columbia,
[3]Department of Medical Genetics, University of British Columbia, [4]Department of Bioinformatics, University of British Columbia

Introduction

Cells use DNA methylation (DNAm) to control gene expressions and DNAm is linked to many diseases like cancer. Many factors like cell types can affect DNAm marks and they are all taken into account when studying DNAm-related diseases. In recent years, several DNAm studies have suggested that a large portion of DNAm variability is associated with genetic ancestry and is heritable, making DNAm a potential confounding factor which is not given enough consideration in the context of DNA methylation analysis. Differentially methylated CpG sites associated with pathology can be confounded by CpGs associated with genetic ancestry causing spurious results. Therefore, genetic ancestry, as a covariate, needs to be accounted for in any epigenome-wide association study (EWAS).

DNAm profiles across tissue types is extremely variable and the amount of variability that can be accounted for by ancestry in placenta samples have not yet been examined. Therefore, in order to investigate how DNA methylation affects prenatal health, it is important for us to identify genetic ancestry -associated CpGs to figure out true positives. This DNAm variability in the placenta due to ancestry needs to be accounted for in large scale DNAm studies, or else no meaningful interpretation of results can be done to assess prenatal health.

Hypothesis: DNA in placental tissue is differentially methylated across populations of different ethnicities.

Dataset 1 (Training):

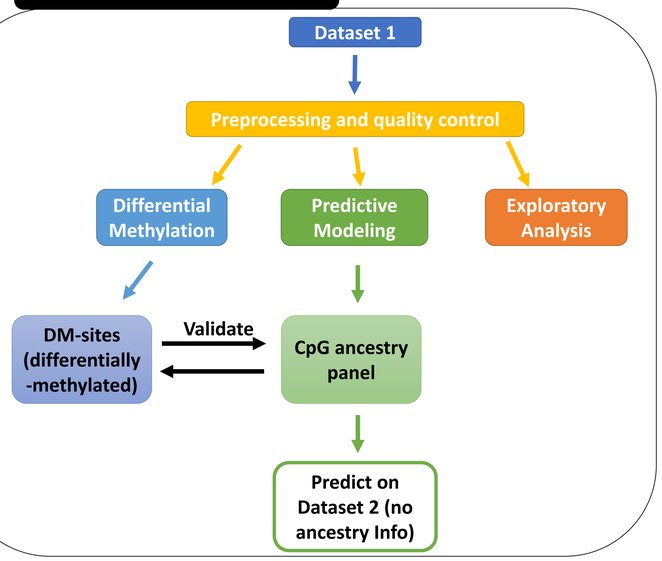
	Caucasian	Asian
Male	16	5
Female	17	7
Sum	33	12

Note: DNAm was measured by 450K microarray from Illumina

Dataset 2 (Test):

- Raw DNAm datasets of 52 placental tissue from Price et al. paper (Pice et al., 2016).
- The genetic ancestry of dataset 2 samples is unknown.
- DNAm was measured by 450K technology.

Workflow



References:

- C. K. Williams, A. Engelhardt, T. Cooper, Z. Mayer, A. Ziem, L. Scrucca, Y. Tang, C. Candan, M. M. Kuhn, "Package 'caret'", 2015.
- Martin J Aryee, Andrew E Jaffe, and et al.. 2014. "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays." Bioinformatics 30 (10): 1363–69.
- Price, E. M., M. S. Penaherrera, and et al. 2016. 'Profiling placental and fetal DNA methylation in human neural tube defects', *Epigenetics Chromatin*, 9: 6.

Summary

- SVM performed slightly better than glmnet (for both training and testing error)
- Final model used 11 CpG predictors and was built with glmnet with a AUC of 0.981 and 0.977+-0.024 for training and testing error respectively (α = 0.75, λ = 0.25).
- The classifier predicted all of the unlabeled test set to Caucasian, which we doubt is the true case.
- We suspect the test set is too 'different' from the training data set for the classifier to perform accurately on the test set.

Future Directions

- Normalizing and QCing the test and training datasets together may be necessary for DNA methylation classifiers to perform well
- Using MDS ancestry coordinates from population stratification meta-analyses may provide 'labels' to assess classifier performance or iprove model building. (self-reported ancestry can be unreliable)

Preprocessing and QC

s

Exploratory Analysis

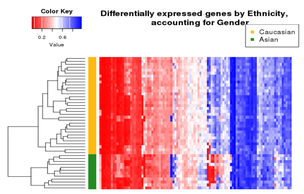
s

Differential Methylation

Here we used a linear model to identify differentially methylated probes with limma.
We first fit a linear model with ancestry as the only covariate to obtain top differentially methylated CpG sites and use a cutoff of p value= 0.01, we identified 106 CpG sites that are differentially methylated between Caucasian and Asian genetic ancestry.
However, DNA methylation is known to associate with gender, so the interaction effect ancestry and gender was accounted for in another linear model. Using a cutoff of p value = 0.01, we identified just 13 CpG sites that are differentially methylated between Caucasian and Asian genetic ancestry, when the interaction effect of ancestry and gender was accounted for. Here are 6 of those sites:

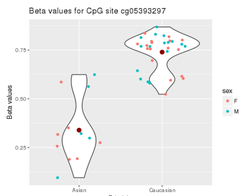
	logFC	AveExpr	t	P.Value	adj.P.Val	B
cg16329197	0.5368513	0.4996451	9.546477	0	0.0000020	17.319139
cg25025879	0.4343004	0.4535298	9.205678	0	0.0000028	16.272412
cg05393297	0.4229273	0.6325893	8.211144	0	0.0000427	13.142162
cg14581129	0.2265901	0.5049442	6.992750	0	0.0016940	9.185289
cg26513180	-0.0294624	0.0339633	-6.732052	0	0.0025085	8.327817
cg19041462	0.1018915	0.8764931	6.689685	0	0.0025085	8.188292

Plot top 100 Limma hits when accounting for ancestry and gender interaction effect

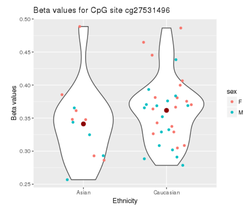


Check overlap with GLMnet predictions

There is an overlap of 5 CpG sites between the those detected by GLMnet and the ones detected in linear regression analysis and the plots of 2 of these CpG sites:

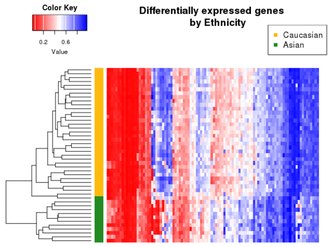


Random CpG sites for comparison



Other: table and plot for methylated sites by ancestry only

We fit a linear model with ancestry as the only covariate to obtain top differentially methylated CpG sites. We are testing the null hypothesis that samples in the two ethnic groups are drawn from populations with the same mean CpG site methylation.
Using a cutoff of p value = 0.01, we identified 106 CpG sites that are differentially methylated between Caucasian and Asian genetic ancestry.



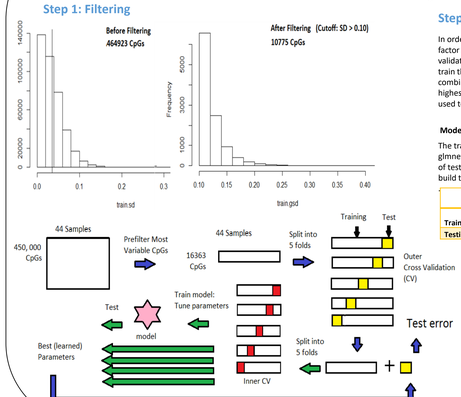
Predictive Modeling

Background

We propose to compare SVM and Elastic net regression (glmnet) to generate a classifier that will predict ancestry based on DNA methylation features (CpGs).

We chose to use an arbitrary threshold to prefilter CpGs to retain. CpGs with a standard deviation (SD) greater than 0.10, leaving 10775 CpGs for model building. We reason that only variable CpGs are likely to be able to be used to distinguish ancestry, however we are uncertain with the viability of this strategy in setting a threshold.

Step 1: Filtering



Step 2: Nested Cross Validation for Model Performance

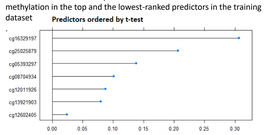
In order to estimate the optimal parameters of these models (Elastic net: α, λ; SVM: penalty factor C) and to estimate prediction accuracy, data was randomly divided into 5 sets for cross validation (CV), using a testing and training set. The samples in the training set was used to train the model with a grid of tuning parameters. Five values of α, and λ (25 different combinations) were tested for elastic net. The parameters resulting in the model with the highest ROC on the testing sets were chosen. An outer layer of cross validation (nested) was used to estimate test error.

Model Performance

The training performance was AUC = 0.981, 0.988 for glmnet and SVM, respectively. However, glmnet was more stable across repeats (we used repeatedcv, repeats = 3) during the estimation of test error (0.977 +- 0.024 vs 0.978 +- 0.05). Despite higher performance of SVM, we chose to build the final model with glmnet. The final tuning parameters used were α = 0.75, λ = 0.25

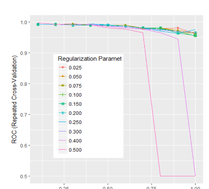
	glmnet			SVM		
	AUC	Sens	Spec	AUC	Sens	Spec
Training	0.9809524	0.633333	1	0.9880952	0.916667	0.960318
Testing	0.977+-0.024			0.978+-0.05		

The final model utilized 11 predictors. Here we examine the methylation in the top and the lowest-ranked predictors in the training dataset



Tuning Alpha and Lambda in Elastic Net regularization

Elastic net regression employs a penalty that is a combination of L1 and L2 norm controlled by α and λ tuning parameters. We wanted to use more L1-regularization to obtain a small panel of biomarkers (α = 0.75, λ = 0.25). We show the relationship between α, λ, and training error (AUC).

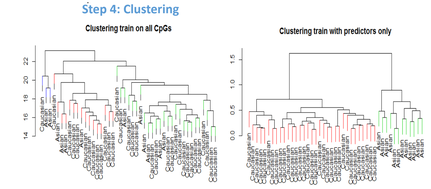


Step 3: Testing model on unlabeled test data

Predicting ancestry on test data

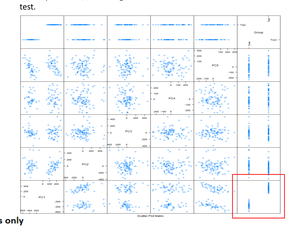
Using the glmnet model, which utilizes 11 predictor CpGs to predict ancestry, we assessed the heterogeneity of our unlabeled test data. We found that all Samples were classified as Caucasian, with a probability ranging from 0.57 to 0.82.

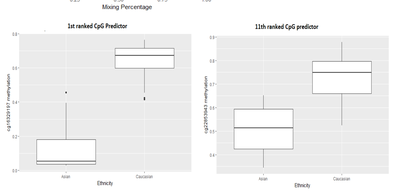
Step 4: Clustering



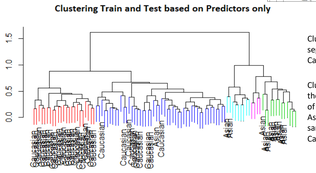
Step 4: PCA on test

We suspect that the classification is performing poorly on the test data because we doubt that it is truly entirely Caucasian. From this PCA plot on test and train data, we can see that the first PC is much different in train vs test (bottom right panel). This indicates that our two datasets are very different, which might make the classifier unsuitable for the test.





Clustering Train and Test based on Predictors only



Clustering based on predictors only improves the separation of the training data into Asian and Caucasian clusters.

Clustering combined Train and Test data using the predictor CpGs only results in separation of two main clusters, consisting of primarily Asians or Caucasians. The test training samples (unlabeled) primarily fall into the Caucasian cluster.