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):

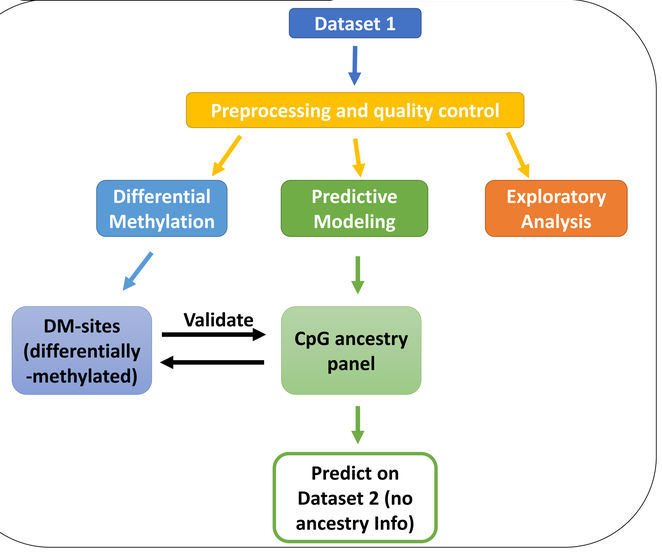
	Caucasian	Asian
Male	16	5
Female	17	7
Sum	33	12

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:

1. C. K. Williams, A. Engelhardt, T. Cooper, Z. Mayer, A. Ziem, L. Scrucca, Y. Tang, C. Candan, M. M. Kuhn, "Package 'caret'", 2015.
2. 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.
3. 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.
4. Vapnik VN. Statistical Learning Theory. Wiley, New York (1998)
5. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.

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.

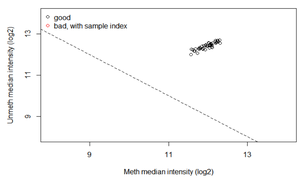
Future Directions

- 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)

Normalization and Quality Checks

Quality checks

Samples with bad detection P-values and abnormal distributions of methylation intensities will removed. All samples based these quality checks.

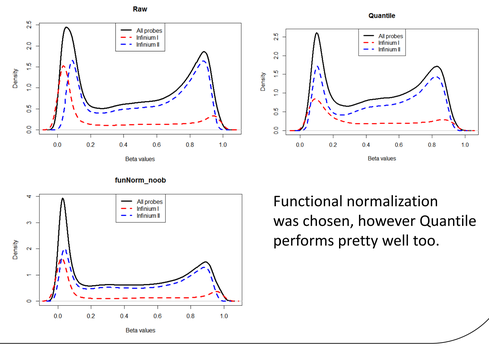


Filtering probes

- Filtered 485,512 down to 464,923 probes
- We removed 20,589 probes based on bad detection p value > 0.01, bead count > 3, and sex probes.

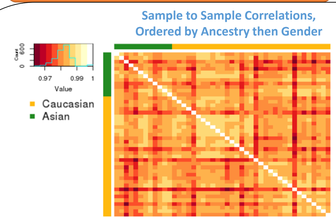
Normalization

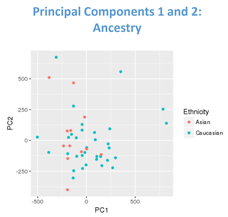
Type I and Type II probes have different beta value distributions. It is important to adjust for these different types of distributions to make the probes comparable. There are a couple methods out there to do this. We compare Quantile and Functional normalization.

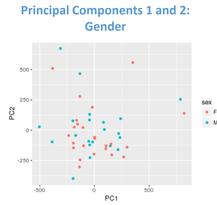


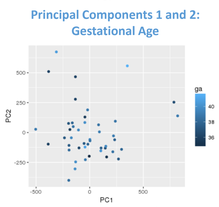
Functional normalization was chosen, however Quantile performs pretty well too.

Exploratory Analysis









Sample-sample correlations confirm that outliers were mostly removed in preprocessing, but clustering by genetic ancestry is not qualitatively apparent. Similar results are seen when plotting the first three principal components of genetic ancestry. Principal component plotting of other variables (sex, gestational age and sample group) does not clearly show what is driving the variance in the data. A specific interaction of variables is the likely driver, and is explored in differential methylation analysis.

Differential Methylation

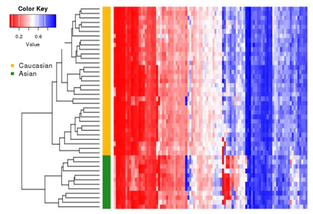
Here we used a linear model to identify differentially methylated probes with limma.

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.

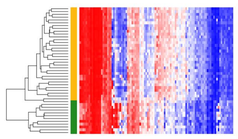
However, DNA methylation is known to associate with gender, so the interaction effect ancestry and gender was accounted for in another linear model. Using a cutoff of p value = 0.01, we identified just 13 CpG sites that are differentially methylated between Caucasian and Asian genetic ancestry, when the interaction effect of ethnicity 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

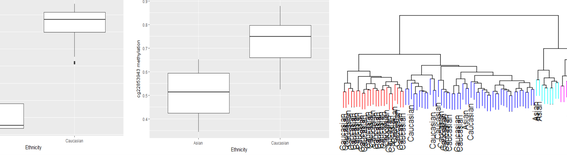
Top 100 Limma hits when accounting for Ancestry and Gender interaction effect



Top 100 Limma hits by Ancestry



Three CpG sites prioritized by both limma and glmnet



Three random CpG sites not prioritized by limma nor glmnet

There is an overlap of 5 CpG sites between those detected by GLMnet and those detected in linear regression analysis. Three of these sites are plotted below. Three random CpG sites that are not detected as differentially methylated between ancestry groups are also shown.

Functional analysis

Using the COHCAP (City of Hope CpG Island Analysis Pipeline) package, CpGs identified as predictors by glmnet and prioritized by limma were annotated with chromosome mapping, location, gene name and CpG island information.

The package mygene was used to find GO terms for genes.

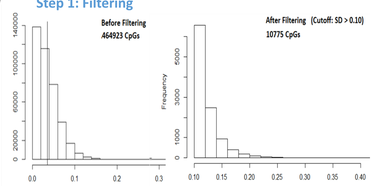
- glmnet:
  Chromosomes 1, 2, 3, 12, 15 or 17
  3 genes: SH2D5, IVL and C3orf21
- limma:
  Chromosomes 1, 2, 8, 12, 15, 16, X
  7 genes: FANCA, VPS37A, WDR90, CCNL2, LOC391322, ARSD, KCNS3

Predictive Modeling

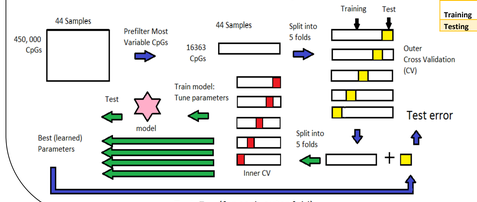
We compare SVM [4] and elastic net logistic regression (*glmnet*) [5] to generate a classifier that will predict ancestry based on DNA methylation features (CpGs).

We filtered CpGs with a standard deviation (SD) greater than 0.10, leaving 10,775 CpGs for model building.

Step 1: Filtering



Step 2: Nested Cross Validation



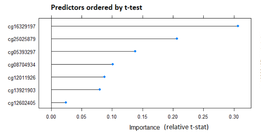
Model Performance

The training performance was AUC = 0.981, 0.988 for *glmnet* and SVM, respectively. However, *glmnet* was more stable across repeats (we used repeated CV, 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

.

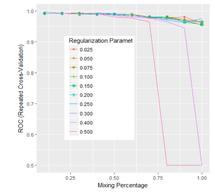
	glmnet			SVM		
	AUC	Sens	Spec	AUC	Sens	Spec
Training	0.9809524	0.933333	1	0.9880952	0.916667	0.960318
Testing	0.977+-0.024			0.978+-0.05		

The final model utilized 11 predictors.



Tuning Alpha and Lambda in Elastic Net Regularization

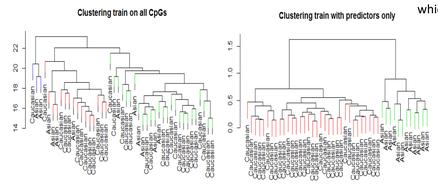
Elastic net regression employs a penalty that is a combination of L1 and L2 norm controlled by α and λ tuning parameters. We wanted to utilize more L1-regularization to obtain a small panel of biomarkers (α = 0.75, λ = 0.25). We show the relationship between α, λ, and training error (AUC).

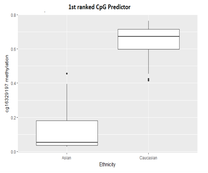


Step 3: Predicting Ethnicity on Unlabeled Data

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.





Follow-up: PCA on Merged Data

We suspect that the classification is performing poorly on the test data, as we doubt that it is truly entirely Caucasian. From this PCA plot on merged test and training 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 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.