Predictive Modeling

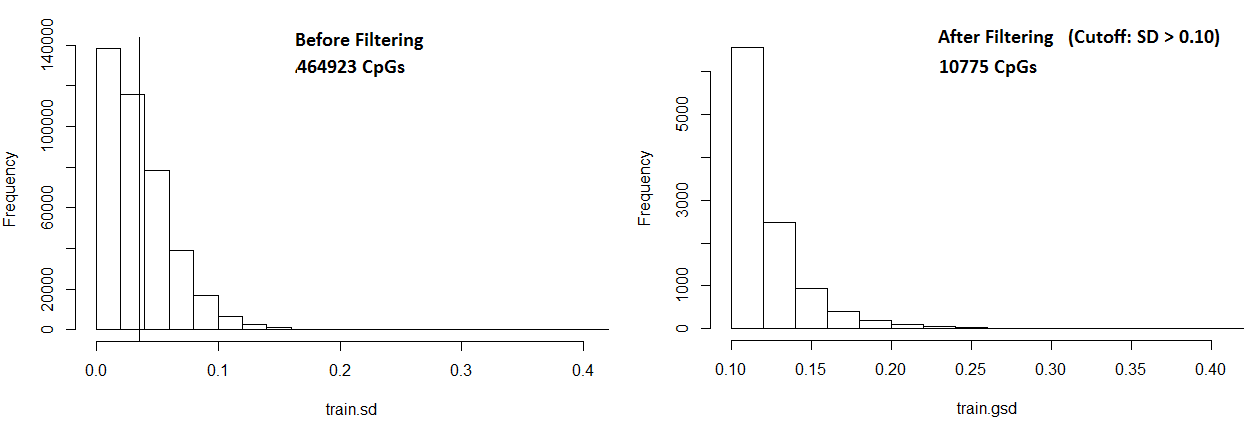
# Background

Support Vector Machines (SVM) is a powerful supervised learning method used to recognize patterns in data. SVM has been used in many bioinformatics applications for classification, regression, and outlier detection because it avoids overfitting, it can account for nonlinear relationships, and is robust to noise [1].

‘Elastic net’ is a regularized regression method that combines the L1 and L2 penalties from LASSO and ridge regularization methods. In this way, Elastic net overcomes some of the limitations of LASSO and ridge. Elastic net regularized regression (GLMnet) has been used to create a multi-tissue age predictor based on DNA methylation before [2], indicating that this might be well-suited for our dataset.

We propose to compare SVM and Elastic net regression methods in generating a classifier that will predict ethnicity based on DNA methylation features (CpGs).

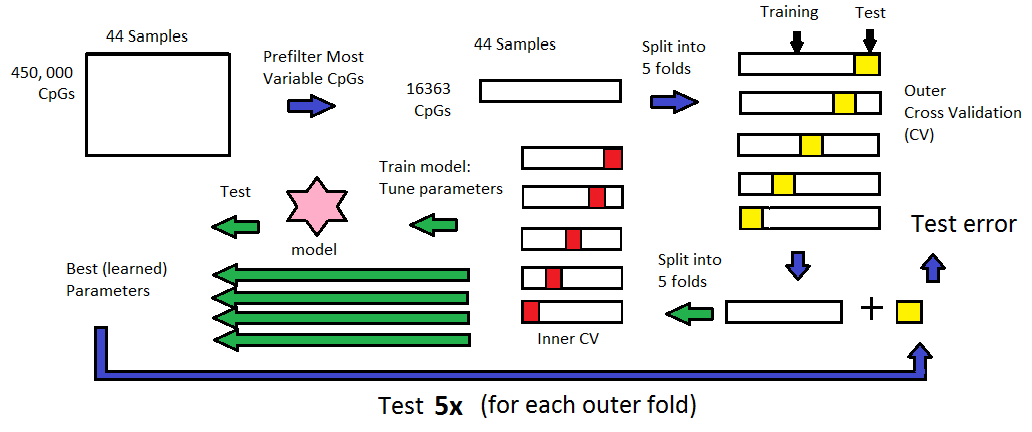
# Step 1: Filtering

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 ethnicity, however we are uncertain with the viability of this strategy in setting a threshold.

# Step 2: Nested Cross Validation for Model Performance

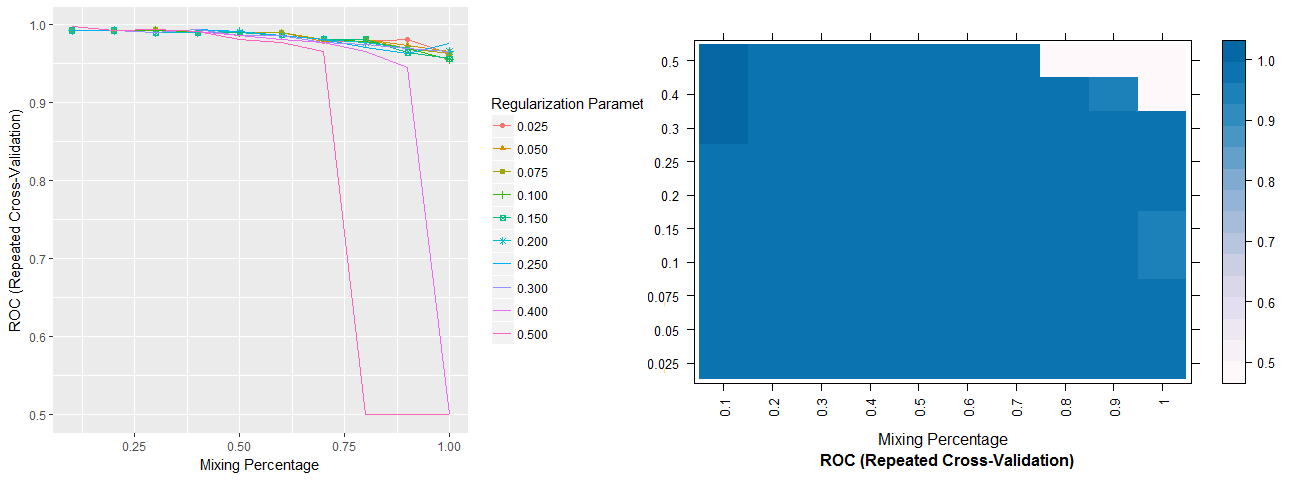
## Cross Validation

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.



## 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).



## 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 utilizes 11 CpGs as predictors.

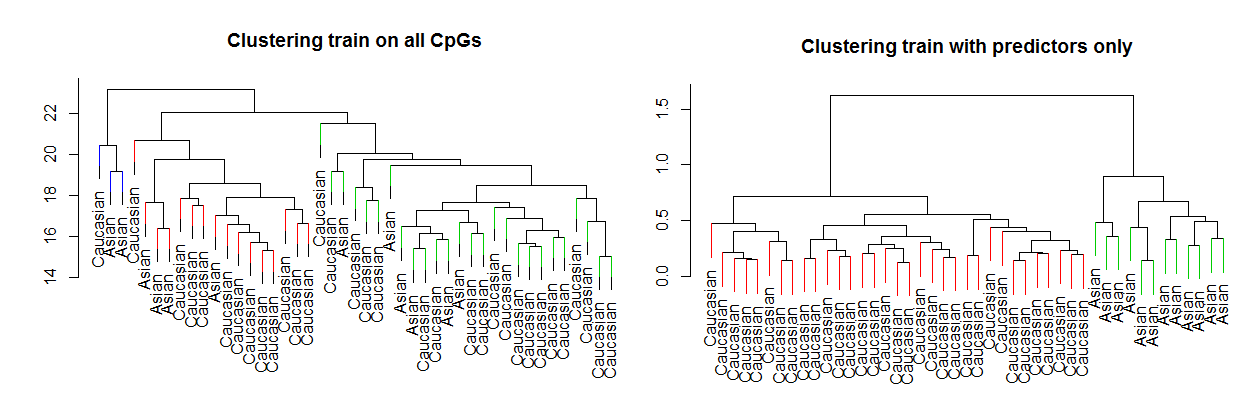
# Step 3: Testing model on unlabeled test data

## Predicting ethnicity on test data

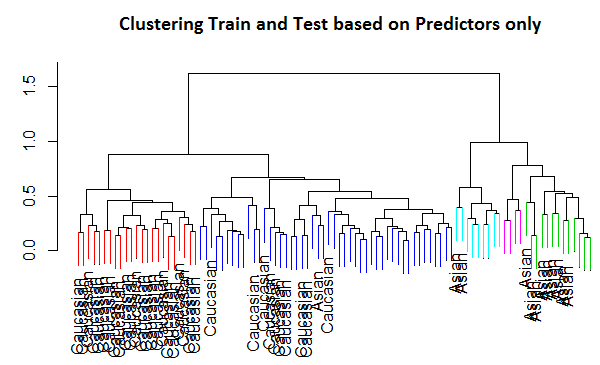
Using the glmnet model, which utilizes 11 predictor CpGs to predict Ethnicity, we assessed the heterogeneity of our unlabeled test data. We found that all Samples were classified as Caucasian, with a probability ranging from 0.57 to 0.82.

## Clustering Analysis

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

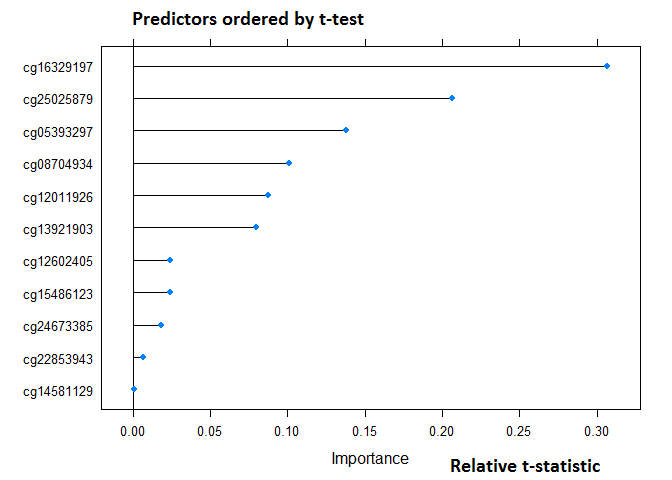


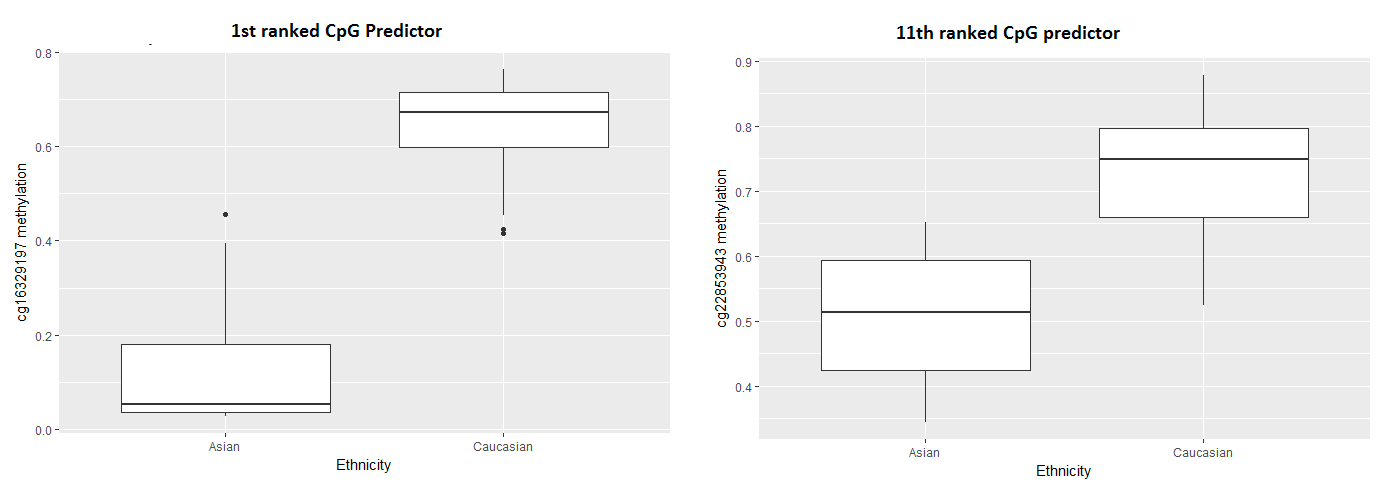
Clustering combined Train and Test data using the predictor CpGs only results in separation of two main clusters, consisting of primarily Asians or Caucasians. The test training samples (unlabeled) primarily fall into the Caucasian cluster.



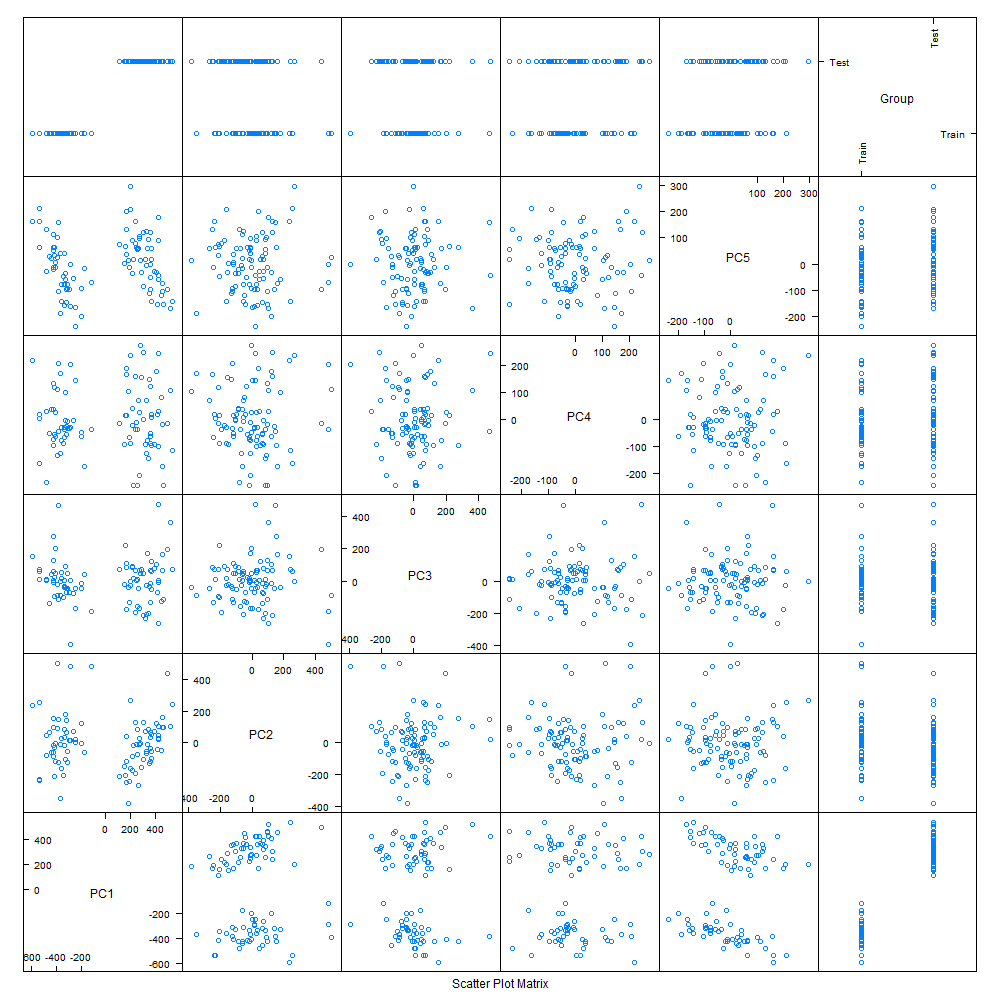
# Step 4: Analyze predictors

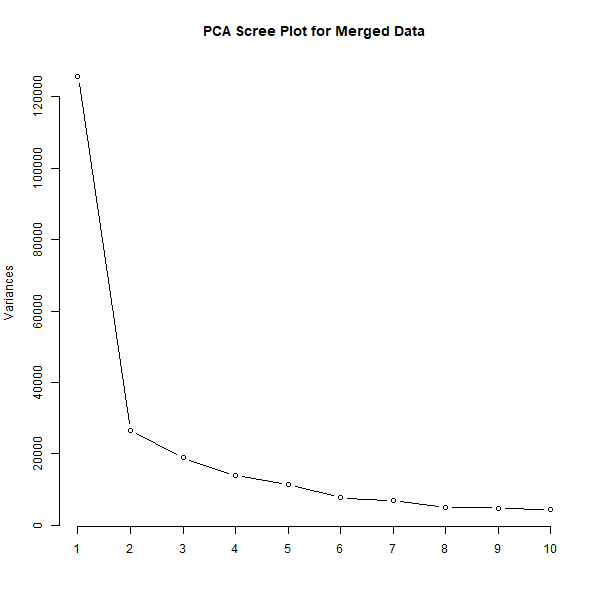
Plot of all 11 predictors



Next we plot the top and last predictor to see their difference in methylation.

# Step 5: 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, 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.

[1] - Vapnik VN. Statistical Learning Theory. Wiley, New York (1998)

[2] - Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.

[3] - Kuhn, M. (2008). Caret package. Journal of Statistical Software, 28(5)