Epigenomic Aging

Eric Kramer

1 June 2015

Introduction

DNA methylation is the binding of a methyl group to a guanine on DNA. Multiple studies have documented correlations between methylation in a gene's promoter and the expression of that gene; however, the exact causal relationship between these variables is unknown and likely depends on a variety of factors, such as the presence of eQTLs and the density of CpG sites in the region.

DNA methylation plays a crucial role in tissue differentiation. Studies have tracked increasing methylation as a cell proceeds down the differentiation pathway for innate immune cells [1] [^Aryee2014], while the first principle component of methylation data will separate tissues, if there are multiple tissues present.

[1]

More recent work has shown that environmental factors also affect DNA methylation. Age, socio-economic status and ancestral background can all affect DNA methylation.[2] In general, studies have shown increased methylation with age and low socio-economic status. Nseuronal pathways appear to be the most highly methylated with agining; while stress and immune pathways demonstrate the most increase in methylation due to low socio-economic status.

[2] Disentangling the effects of ancestry and socio-economic presents problems in many datasets where only one of these variables is measured although they are strongly correlated

Hovarth's seminal 2013 paper provided a predictive model of age based on DNA methylation. This model is able to predict a sample's age with astonishing accuracy. More importantly, the model is applicable to a wide-range of tissues. However, we noticed that this model performs worse on samples of African ancestry, likely due to the above-mentioned environmental effects. Because of this, we decided to recalibrate the epigenetic so that is provides improved performance in non-Europeans, while still maintaining performance in diverse tissues.

The following work is an attempt to rebuild the epigenetic clock of Horvath to ensure that the model performs equally well on European, African and Asian samples; while still demonstrating the cross-tissue validity of the original epigenetic clock.

This work proceeeds in four primary steps:

1. Preprocessing of publically available methylation data
2. Creation of predictive model for ethnicity
3. Promoter-level feature selection
4. Recalibrarting the epigenetic clock

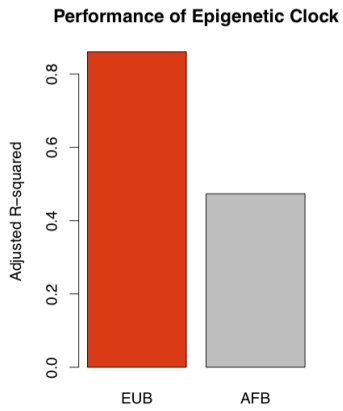


Figure 1: Horvath's epigenetic clock performs significantly worse on Belgians of african ancestry (AFB) than Belgians of European ancestry (EUB)

Preprocessing of publically available methylation data

In order to recalibrate the epigenetic clock so that is stable across ancestries and tissues, we require a training set with samples from multiple ancestries and multiple tissues.

Gene Expression Omnibus data

Gene Expression Omnibus (GEO) contains over 20,000 Illumina 450k DNA methylation arrays. Among these, approximately 7247 samples are annotated with an age. I downloaded and processed raw data for these GEO series and placed these data into a SQLite database.[3] These samples range in age from newborns to 105 years old.

[3] See ./data/BMIQ.db. Recent work has allowed SQLite databases to be used within R as *data.frames*. See the *dplyr* package for more detail

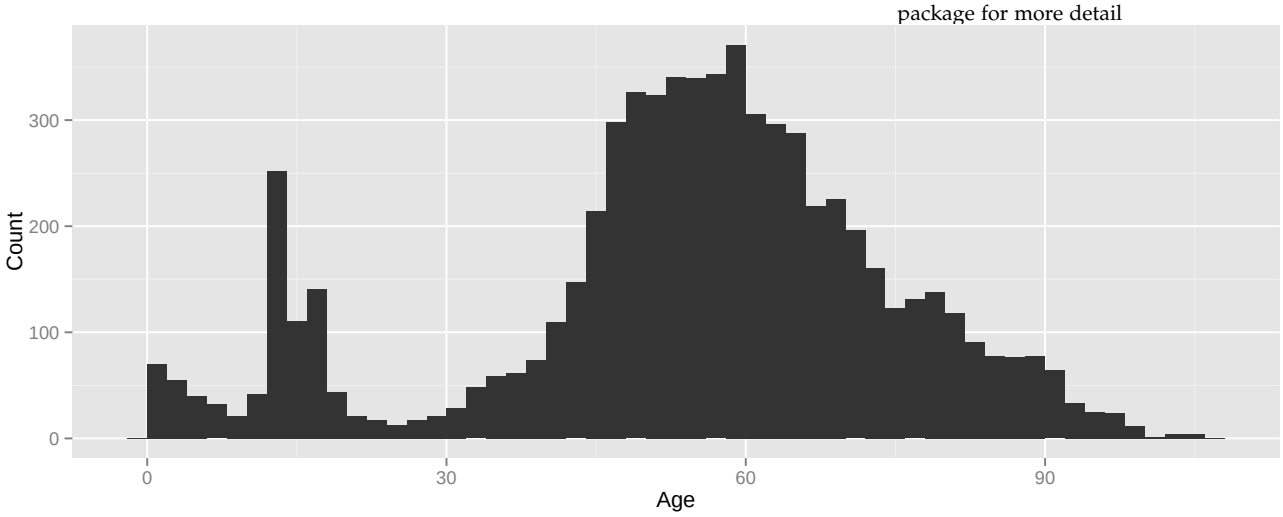


Figure 2: Age distribution of GEO samples assembled for this project

This database contains samples from 18 tissues. A majority of the samples are whole blood or other blood-based cell types (e.g. monocytes). While solid-tissue samples, such as samples from lung or muscle, are comparatively more rare. Because of the highly unbalanced nature of the dataset, careful statistics must be used to ensure that measured effects are not caused by confounding.

Table 1: Number of samples for each tissue

Tissue	Count
Adipose Tissue	174
Bone Marrow	72
Brain	832
Breast	14
Buccal	96

Tissue	Count
Colon	53
Leukocytes	923
Liver	231
Lung	11
Lymphoblasts	322
Lymphocytes	614
Monocytes	1202
Muscle	86
Neuronal	145
Pancreas	221
T-cells	262
Thyroid	82
Whole Blood	1907

Normalization

BMIQ normalization is one of the most common methods to normalize DNA methylation data. It uses a mixture of beta-distributions and quantile normalization to adjust for differences between the probe types on the chip. After normalization, we see that the PCA nicely separates samples based on tissue. The first principle component separates liquid tissues (whole blood, monocytes, T-cells, leukocytes, etc) from the solid tissues (lung, brain, muscle, pancreas, etc). The second principle component is correlated with age and could represent age-related changes in methylation.

If we color the plot according to tissue, rather that tissue group, we see that the two outlying clusters of solid tissues are samples from the cerebellum and the brain, excluding the cerebellum. Additionally, we see sub-structure within each of the clusters. For instance, monocytes in blue and T-cells in purple are clearly separated on the PCA.

Creation of a predictive model for ancestry

Unfortunately, most GEO entires are not annotated with ancestral background. Because of this, we decided to create a predictive model, which is able to impute a sample's ancestry based on it's DNA methylation levels. To do this, we used methylation sites where common SNPs are located. These methylation sites allow us to infer genotype of a sample, and therefore the ancestry.

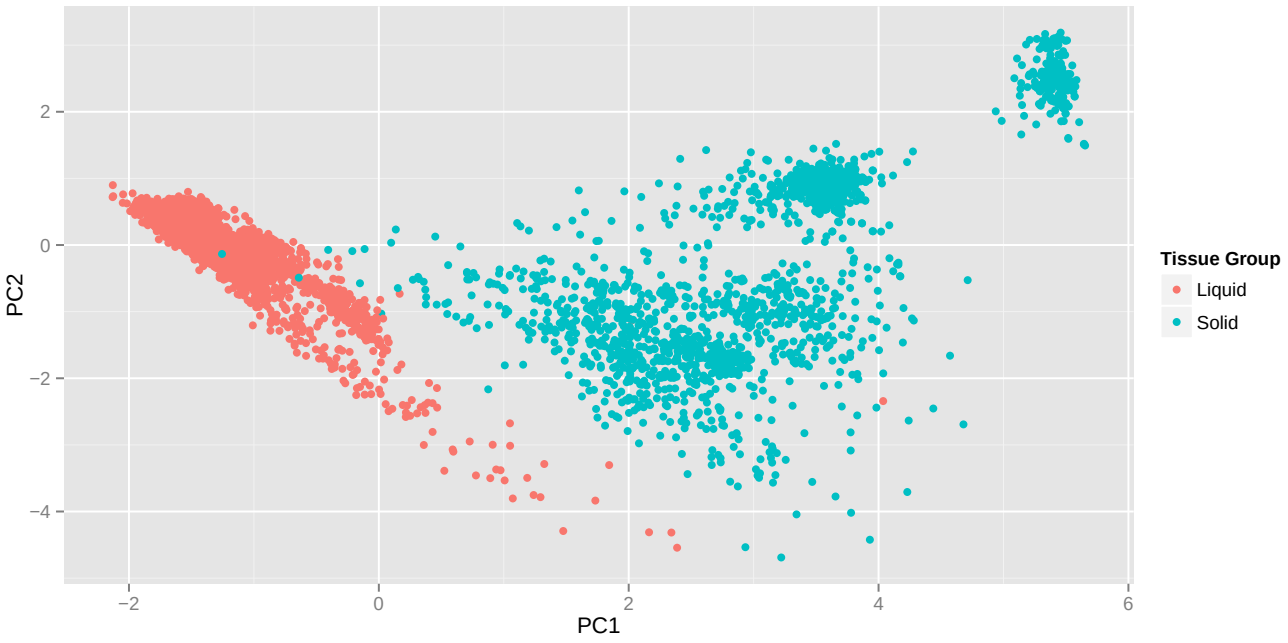


Figure 3: PCA of GEO samples colored according to tissue group

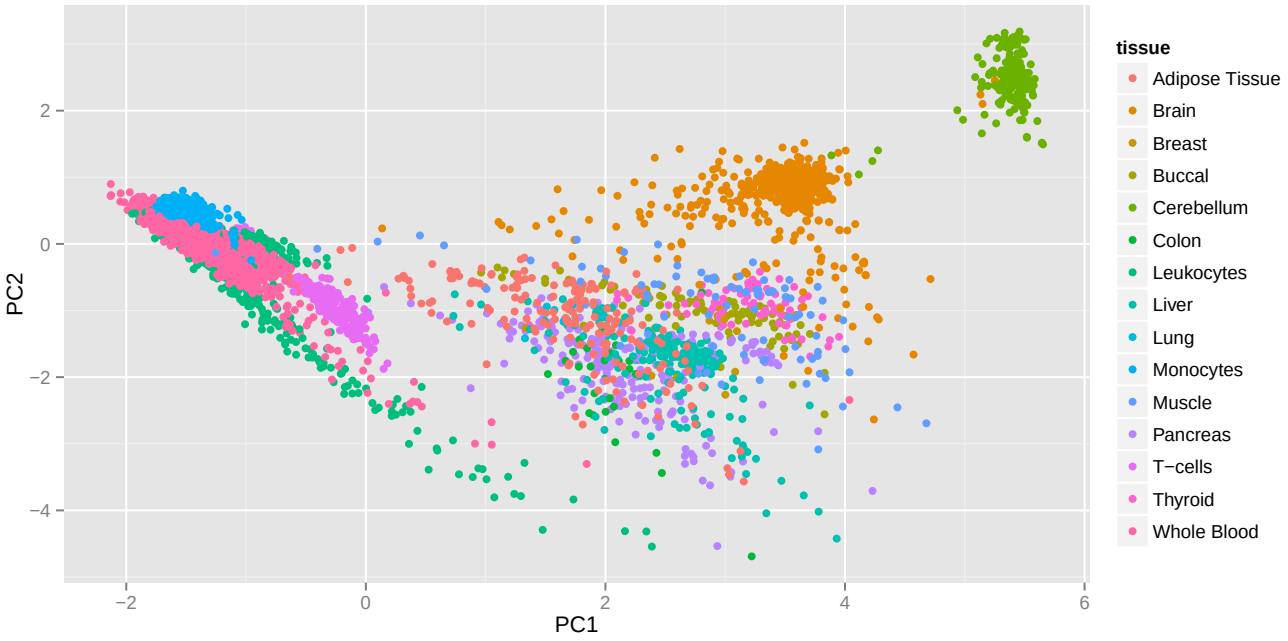


Figure 4: PCA of GEO samples colored according to tissue

Ancestry Informative Methylation Markers (AIMMs)

DNA methylation can only occur at guanines immediately preceeded by a cystine, known as CpG sites. SNPs in a CpG site will necessarily disrupt methylation. Conversely, we can use methylation at sites with known SNPs as a way to infer genotypes.

In order to create a set of ancestry informative methyaltion markers, I mapped common variants from the 1000 genomes project (>5% MAF in one population) onto the CpG sites on the Illumina 450k array. In total, I discovered 14226 methylation sites with a common SNP in the CpG site.[4]

[4] The ids for these sites are stored in ./data/aimms.Rdata in the project directory.

Training the ancestry model

Data from the Human Variation Panel was used for training the model. This dataset contains 96 samples from each of three ancestral background: European (EUR), African (AFR) and Asian (ASN). All samples were collected in the United States. Importantly, the samples are lymphoblastic cell lines. Because of this, the samples appear to be over 100 years-old according to the epigenetic clock.

Elastic net regression was used to build a predictive model for ancestry.[5] The model selected 40 probe sites for inclusion in the model.

[5] See ./ancestry_prediction/aimm_model.R for script for training the model

Testing the ancestry model

Four GEO series, whose samples were annotated with ancestry information, were used to test the ancestry predictive model. The accuracy on these sets ranged from 0.95 to 1. We see that the ancestry model performs well on all four of these datasets.

Table 2: Accuracy of Ancestry Model

GEO Series	Tissue	Average Age	N	Accuracy
GSE36064	Leukocytes	4.58	51	1.00
GSE40279	Whole Blood	64.04	425	0.99
GSE50759	Buccal	9.03	88	0.95
GSE53740	Whole Blood	67.78	358	0.98

Hispanic samples represent an interesting test case for the ancestry model both because this population was not included in the training set and because the hispanic population is admixed. We see that hispanic samples are predicted to be either European or Asian.

Furthermore, if we look at the probabilities for each ancestry for each of the hispanic samples, we see that the model predicts some hispanic samples to be almost entirely Asian, while others are pre-

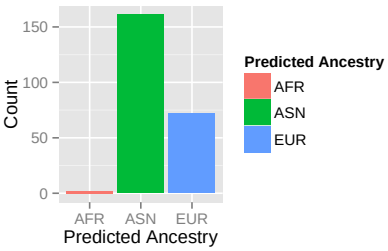


Figure 5: Most likely ancestry for Hispanic samples

dicted to be almost entirely European.

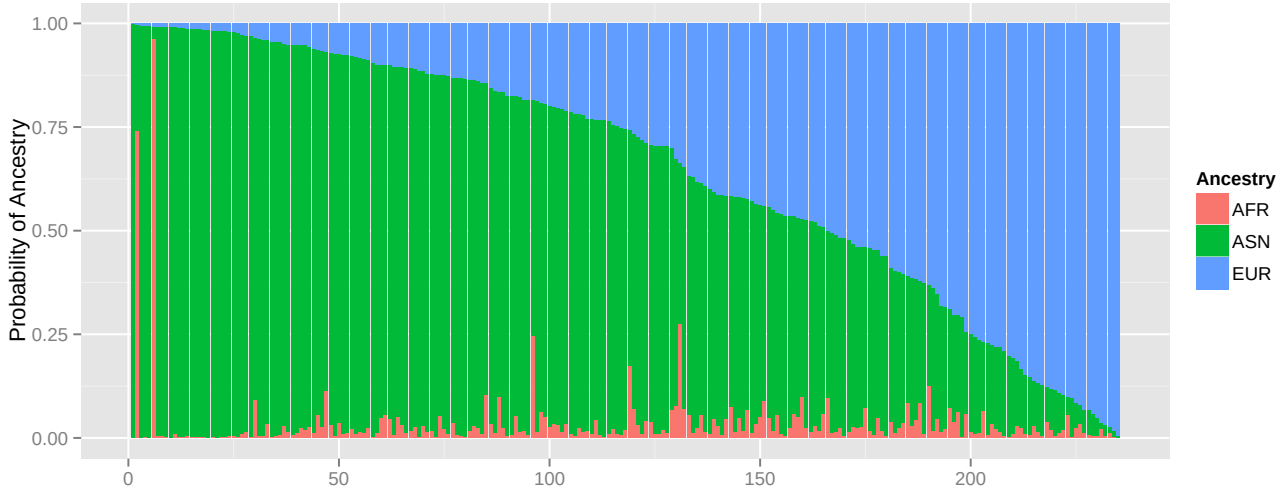


Figure 6: Predicted Ancestral Probabilities of Hispanic Samples

It would be interesting to correlated these predicted ancestry probabilities with actual admixture estimates. It is possible that these predicted probabilities correlated well with admixture estimates for the proportion of European and Asian (that is Native American) ancestry in these hispanic samples.

Applied ancestry model to entire GEO dataset

I then applied the ancestry model to the entire GEO dataset. Unsurprisingly, we find that most of the samples in GEO are of European origin, followed by African and then Asian samples as the most rare.

Table 3: Counts for each tissue-ancestry pairing

Tissue	AFR	ASN	EUR	Total
Adipose Tissue	2	0	172	174
Brain	26	11	793	830
Breast	6	0	8	14
Buccal	17	1	78	96
Colon	2	0	26	28
Leukocytes	63	7	849	919
Liver	3	0	228	231
Lung	2	0	9	11
Lymphoblasts	96	96	96	288
Monocytes	306	95	774	1175
Muscle	2	0	84	86
Pancreas	1	10	210	221

Tissue	AFR	ASN	EUR	Total
T-cells	61	15	186	262
Thyroid	27	6	49	82
Whole Blood	34	230	1642	1906
Total	648	471	5204	6323

Promoter-level analysis

Next, I used linear modeling to determine the effects of age, tissue and ancestry on methylation for each promoter of 16758 genes. These models look for broad effects. For example, measurement of the effect of aging is an effect with is stable across multiple probe sites, multiple tissues and multiple ethnicity. The creation of a dataset that is this large allows us to measure these effects in a highly accurate manner.

Preprocessing probes

Before running the linear models, I discarded probe sites which have a common SNP in the sequence of the probe, multiple matching probes and probes on the X and Y chromosome. I defined the promoter region as the segment within 1,500 base pairs of the transcription start site (TSS). For each gene, if there were multiple transcription start sites, I took the TSS with the greatest number of methylation probes. Finally, I discarded genes with fewer than 4 sites in their promoter. In total, this left 16758 genes.[6]

[6] See ./linear_modeling/util.R for the script which conducts this probe filtering

Preprocessing samples

Before running the model, I discarded samples from *GSE*56105 as this GEO series showed a batch effect on principle component analysis. Additionally, I applied Horvath's epigenetic clock to the dataset and discarded 20 samples with an age greater than 120 as these are likely cancerous samples.

Preprocessing tissue annotations

I decided to group the tissues into two main categories: solid and liquid tissues. These two main categories correspond to the two main clusters shown in the PCA in figure 2. In the linear model, I measure da fixed effect, which corresponds to the methylation differences between these two tissue groups, and I controled for the substructure within each of these groups using random effects.

Fitting models

For each of these promoters, I ran a separate linear mixed effects model[7] for each promoter.[8] Each model calculated the effect of of age, tissue group and ancestry on methylation across all probes in the promoter.

[7] See ./linear_modeling/stratified_model_funcs.R for the function which fits one of these linear mixed effects models

[8] See ./linear_modeling/stratified_models.R for a wrapper to distribute this computation on BIC

Results

I defined significant differentially methylated promoters (DMPs) as those that showed both statistical significance (FDR p-value < 0.05) and biological significant ($\Delta\beta > 0.05$). This resulted in 349 tissue-DMPs, 131 age-DMPs and 32 ancestry-DMPs.

Figure 7: Venn Diagram of differentially methylated promoters (DMPs)

Table 4: Top Tissue DMPs

Gene	Estimate	$\Delta\beta$	q value
NCKAP1L	-2.42	0.80	0
BIN2	-2.43	0.79	0
MIR142	-2.44	0.71	0
PTPN22	-1.97	0.70	0
CST7	-2.22	0.69	0
RIN2	1.69	-0.69	0
SPN	-2.10	0.67	0
LINC00426	-1.67	0.67	0
PTPN6	-1.97	0.66	0
CD28	-1.72	0.66	0

Table 5: Top Age DMPs

Gene	Estimate	$\Delta\beta$	q value
DMBT1	0.01	0.13	0
ELOVL2-AS1	0.01	0.11	0
ZNF577	0.01	0.10	0
SAMD9	0.01	0.10	0
DDO	-0.01	-0.10	0
NWD1	-0.01	-0.10	0
LDHD	-0.01	-0.10	0
HLA-DQA1	0.01	0.09	0
CCDC88C	0.01	0.09	0
RBM12B-AS1	0.01	0.09	0

Table 6: Top Ancestry DMPs

Gene	Estimate	$\Delta\beta$	q value
PM20D1	0.79	0.19	0
OR2L13	0.57	0.14	0
SPATC1L	-0.61	-0.10	0
FMOD	-0.52	-0.10	0
DUSP22	-0.45	-0.10	0
TNNT1	-0.46	-0.10	0
LOC100131289	0.51	0.10	0
LRIT2	0.42	0.09	0
EIF4E3	0.34	0.08	0
HLA-DPA1	0.39	0.08	0

Recalibrating the epigenetic clock

To recalibrate the epigenetic clock, I used the top 10 age-DMPs as defined by the previous section and use the probes in those promoters to build a predictive model for age based on methylation. I decided to use an ensemble approach for this modeling. I used 80% of the data to train 11 different predictive models using 11 different methods. Then, I used 10% of the data to build an ensemble from these models. In the ensemble, a final prediction is made by taking the weighted average of the predictions from each of the 10 models used as the basis[9] for the training script.

[9] See ./model_building/training.R

Table 7: Predictive modeling methods used in ensemble model

Abbreviation	Name	Family
gbm	Stochastic Gradient Boosting	Boosting
rf	Random Forest	Boosting
cubist	Cubist Regression Tree	Boosting
svmLinear	Support Vector Machine - Linear Kernel	Kernel
rvmLinear	Revelevance Vector Machine - Linear Kernel	Kernel
rvmRadial	Reveleance Vector Machine - Radial Kernel	Kernel
svmRadialCost	Support Vector Machine - Radial Kernel	Kernel
gaussprRadial	Gaussian Process	Kernel
ppr	Projection Pursuit Regression	Regression
pls	Partial Least Squares	Regression
pcr	Principle Component Regression	Regression

Training ensemble model

The model was trained using an 80-10-10 split of the data. 80% of the data was used for training the base models in the ensemble, 10% of the data was used for finding the weights of the models in the ensemble and the final 10% of the data was used for testing.

Many of the base models contain parameters, which need to be optimized for good performance. Thus, for each of the models, I conducted a 10-fold cross-validation with a grid search over reasonable parameters. The set of parameters which showed the best performance for each model during this cross-validation was then used.[10]

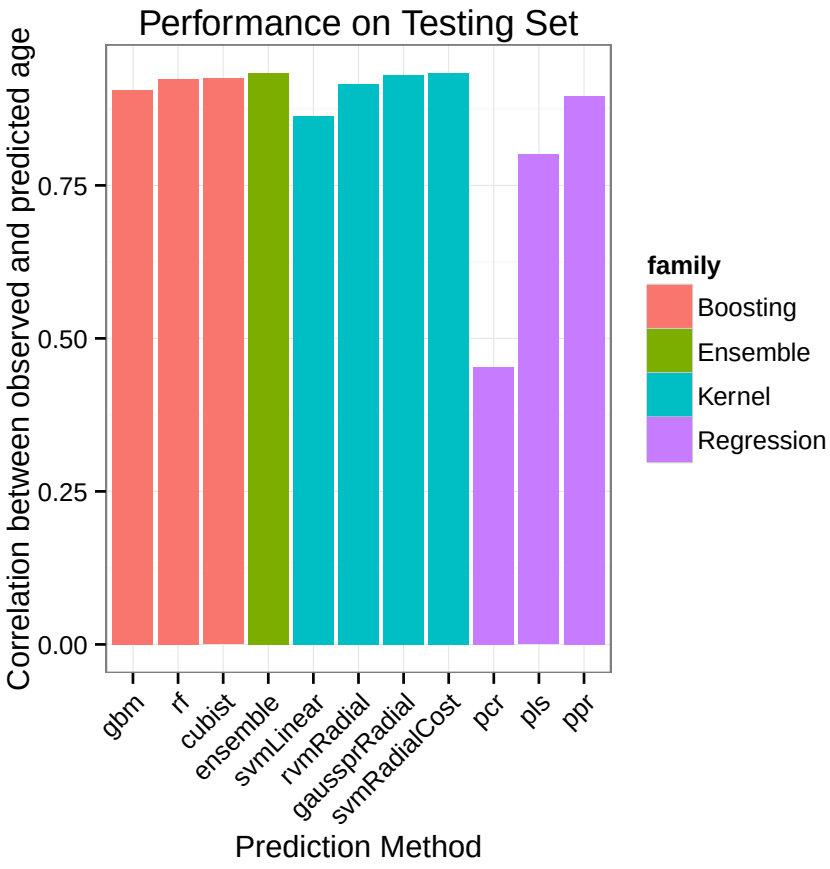
[10] See ./model_building/get_params.R for a script which has reasonable parameter spaces for each of the models used in the ensemble

The ensemble was then trained by using the predictions from the base models as the input features for a linear regression. The resulting ensemble predictions are then a weighted sum of the predictions of the base models.

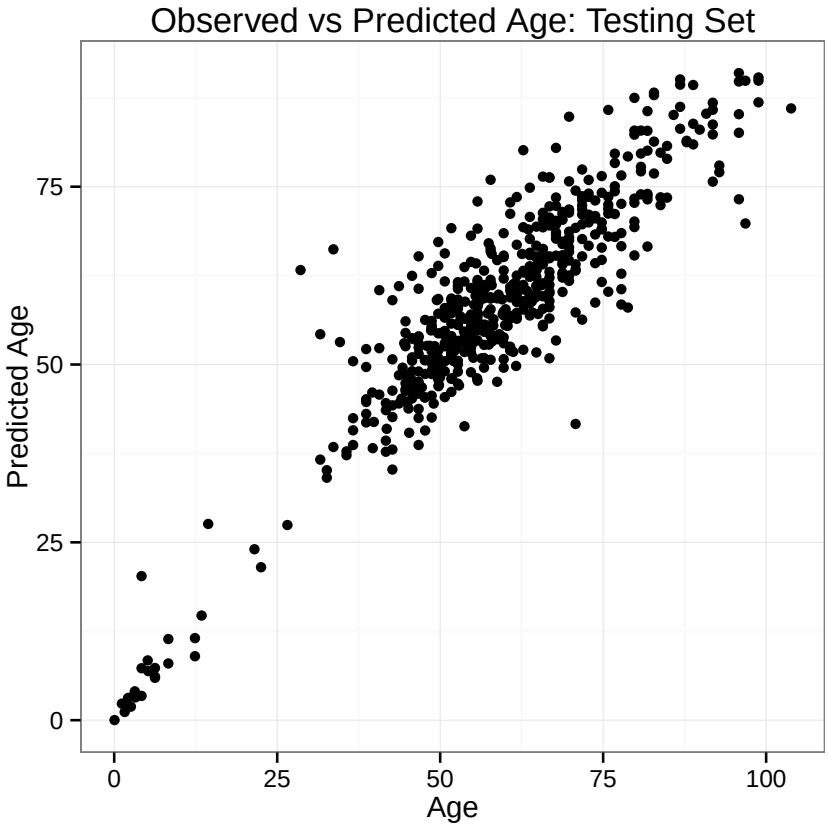
Test final result

To apply the ensemble model to the testing set, we first make predictions using each of the base models. Then, these predictions are used as another feature set for input into the ensemble model. We can then test the performance of the individual base models as well as the performane of the ensemble models.

We see the that ensemble model performs the better than any of the individual models in the ensemble. This is in concordance with literature of predictive modeling, where an ensemble of models typically outpreforms an individual model.Additionall we see that many of the commonly used methods, such as random forests (rf) and support vector machines with a radial kernel (svmRadialCost),


Performance on Testing Set
Correlation between observed and predicted age
0.00
0.25
0.50
0.75
gbm
rf
cubist
ensemble
svmLinear
rvmRadial
gausssprRadial
svmRadialCost
pcr
pls
ppr
Prediction Method
family
Boosting
Ensemble
Kernel
Regression

show quite similar performances.



Next Steps

To finish this project, we need to finalize the set of DMPs, finalize the recalibrated epigenetic clock and release and R package so that other groups can apply both the recalibrated epigenetic clock and the predictive model for ancestry.

To finalize the DMPs, we should continue to use both a biological and statistical criteria for significants. Different variations for the linear model used to measure the effect of aging, tissue group and ancestry can be used, and these variation can change the number of DMPs that are detected.

To finalize the recalibrated epigenetic clock, we should retrain the clock using all probe sites in the age DMPs. The training script in the project directory can do this.

The recalibrated epigenetic clock should be validated one ancestry and one tissue not included in the training set. For this, I suggest excluding the cerebellar samples, since they are outliers on the PCA plot, for the entirely analysis. Then, the cerebellar samples can be

used as a testing set. Additionally, I would use pygmy samples to validate the clock performance on an ancestral group not included in the training set.

Finally, we should create an R package that can apply the recalibrarte epigenetic clock and ancestral model to novel datasets. This package will essentially provide wrapper functions for the prediction methods of the models. The training scripts for the recalibrated epigenetic clock and the ancestry model both have examples of how to apply the model to a new dataset.

References