Environmental and Age Effects on Methylation Changes in Human Brain and Blood Cells

by

Orit Giguzinsky

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 2014
[February 2014]

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
January 29, 2014

Signature redacted

Certified by .....
Prof. Manolis Kellis Thesis Supervisor
January 29, 2014

Signature redacted

Accepted by .....
Prof. Albert R. Meyer, ChairmanMasters of Engineering Thesis Committee
January 29, 2014

Environmental and Age Effects on Methylation Changes in Human Brain and Blood Cells

by

Orit Giguzinsky

Submitted to the

Department of Electrical Engineering and Computer Science

January 29, 2014

In partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Previous studies have shown that DNA methylation may be associated with disease, aging, the rate of aging and genetics. In this thesis, age is accurately predicted from DNA methylation in brain and blood tissues using two different algorithms, Random Forest and Linear Regression. Relationships between DNA methylation, genetics, disease and aging rate were also identified. Furthermore, the differences between the real ages and the predicted ages were found to be associated with the age of onset of aging related disease. The findings presented in this thesis take us a step further in understanding the causes of aging and age related disease and further prove the theory that methylation is related to our internal biological clock and disease.

Thesis Supervisor: Prof. Manolis Kellis
Title: Thesis Supervisor

Acknowledgments

Professor Manolis Kellis, thank you for helping me learn computational biology and allowing me the opportunity to work on such an interesting project. I have dreamed of researching aging since I was two years old. I am truly grateful. Thank you for your understanding of my medical obstacles that have made it extremely difficult to achieve this goal.

Matt, thank you for all your help and dedication. You have taught me a lot. I really appreciate everything you have done for me. I am really greatful.

Paul, thank you for your time, help and dedication. I am truly grateful for all your help.

I would like to thank Professor Leonard Guarente, Sergiy Libert, Eric Bell, Michael Bonkowski and Dr. Christin Glorioso for teaching me the basics of biology and aging and inspiring me to continue researching the field.

I would also like to thank Rachel Sealfon, Gerald Quon, Lori B Chibnik, Andreas Pfenning, Aleksandra Dmitriyevna Kudriashova, David Hendrix, Isaac Nativ, Bat Sheva Ilany and Rami Giguzinsky.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Aging is often associated with a large number of symptoms and diseases, significantly affecting the health and quality of life of older individuals. These diseases and symptoms tend to develop after a certain age, and are often difficult to diagnose, often resulting in misdiagnosis and improper treatment. Traditional medical assessment tools including blood tests, bmi, family history and even genetic testing may be very useful in early diagnosis, and treating specific diseases. However, some diseases have vague or unknown risk factors that are not as useful in risk assessment, prevention and early diagnosis.

Given that aging related symptoms tend to develop later in life and not at any age, there could be a biological mechanism that measures one's age and is also associated with the age of onset of certain diseases and symptoms. Identifying such a biological mechanism could potentially help explain why different individuals age (i.e. develop these symptoms) at different rates and provide better risk assessment tools for specific diseases. Such risk assessment tools would be based on an individual's biological age, instead of relying on one's chronological age, which often leads to prejudice and misdiagnosis. Understanding the biological mechanisms underlying aging and age related disease also has the potential to assist in the development of better drugs, medical procedures and potential cures.

In this thesis I explore DNA methylation as a biological mechanism that measures age. DNA methylation is the addition of a molecule (methyl group, CH3) that gets

added to the DNA in specific locations ([30]) and is believed to be associated with gene expression ([36]). DNA methylation tends to occure in CpG dinucleotides [9, 47]. CpGs are short regions that have a high content of CG sequences [22].

Two age predictors are created using methylation data from the prefrontal cortex of hundreds of brains and two different prediction algorithms, Random Forest, and Linear Regression. The predicted ages are relatively accurate, resulting in an error rate of about 4.5 years in both algorithms. An attempt is made to improve the accuracy by adjusting for environmental factors, but then proves to be unsuccessful. The features that were found to be highly predictive of age in each algorithm are explored and the findings suggest that they may be associated with multiple age related conditions.

Evidence is found suggesting that the differences between the predicted methylation ages and the chronological ages are related to the age of onset of aging related conditions and to one's individual biological rate of aging. These age related conditions and individual differences in the rate of aging seem to be associated with genetic factors and methylation changes that work together. These findings suggest that compared to chronological age, methylation age in combination with genetics may provide a better risk assessment mechanism for developing an aging related disease.

1.1 Organization

Chapter 2 introduces studies that have done similar research and provided background and ideas for the research presented in this thesis. The results presented in this thesis are compared to those found by these previous studies and a high level of consistency is found.

In Chapter 3, I introduce the data I worked with to predict age and explore the causes of aging. These data enable me to research the biological differences between people at different ages, environmental factors that may mask these changes and the relationship between these changes, genetics, aging and aging related disease.

Chapters 4 and 5 use the brain methylation data from Chapter 3 to predict age

from methylation. The prediction results are accurate in both prediction algorithms, which implies that methylation is so highly influenced by age that the relationship between age and methylation can be accurately modeled using different prediction algorithms. The findings presented in these two chapters suggest that the error residuals, the difference between real age and methylation age, are related to biological factors that seem to be associated with one's individual aging rate.

In Chapter 6 the data are corrected for different environmental covariates in order to try to reduce the prediction errors obtained in Chapters 4 and 5. A covariate is a continuous variable that affects the dependent variable [35, 44]. This chapter aims to correct for methylation changes caused by environmental factors that mask age related methylation changes and thus make it more difficult to accurately model the relationship between DNA methylation and age that could potentially reveal the causes of aging and aging related disease.

In Chapter 7 a set of highly predictive age features is identified in each of the two algorithms presented in Chapters 4 and 5. These features are explored for their biological functions and chromatin states. Chromatin is a combination of DNA and proteins inside the nucleolus of the cell [29]. Chromatin states are combinations of chromatin marks where each state shows specific characteristics, biological roles, functional annotations and motifs. These states enable us to annotate the genome in a way that can potentially help us better understand human disease ([13]. The results indicate that the chromatin states that are associated with aging differ from those of general methylation probes and that the biological functions associated with these features are related to key functions that deteriorate with age and may be associated with aging related conditions such as high blood pressure, Alzheimer's (AD) and heart disease.

Chapter 8 uses the predictions from Chapter 4 and 5 and the significant features identified in Chapter 7 to explore the relationship between the age of disease onset and the prediction residuals, showing that faster methylation aging is related to earlier onset of disease. These results further prove the theory that the prediction residuals reflect individual biological aging rates. The findings presented in this chapter also

show that compared to real age, methylation age is a better indicator of the risk of developing an aging related disease. These findings suggest that the age of onset of age related disease is related to one's individual age acceleration, meaning that those with an older methylation age relative to their real age are more likely to suffer from aging related conditions at an earlier age. The differences in the significance level between age acceleration and the age of disease onset across the different diseases seems to be related to the tissue type from which I got the methylation data. Out of the 8 aging related conditions whose age of onset was examined in this chapter, AD seemed to have the strongest relationship with age acceleration, probably because the methylation used to predict the age acceleration was taken from the prefrontal cortex of the brain.

In Chapter 9 I examine the relationship between methylation, genetics and aging disease and show that such a relationship exists. In this chapter I identify genetic factors or single nucleotide polymorphisms (SNPs) that are highly associated with changes in methylation in patients that suffer from a specific aging related condition. I also identify SNPs that together with individual prediction residuals identified in Chapter 8 have a strong influence on DNA methylation. This further proves that aging related conditions and aging rate differences are influenced by genetic and epigenetic factors that affect one another. Epigenetics is the study of change in gene expression or in cell related phenotype that does not change the DNA ([6, 18]).

Chapter 10 explores age prediction in blood using the strategies and findings used to predict age in the prefrontal cortex of the brain. In this chapter I use the blood methylation data that was used in [19] and show that Random Forest can predict age from blood methylation nearly as well as it predicts age from brain methylation. This further validates the results presented by the authors and further proves the idea that DNA methylation can predict age in multiple tissues. This chapter also shows that methylation probes that are highly predictive of age in brain methylation are somewhat predictive in blood.

Chapter 11 summarizes the findings and contributions presented in this thesis, and discusses potential future directions.

Chapter 2

Background

2.1 Background

Genetics is known to be associated with disease, but often genetics does not sufficiently explain the causes of disease ([39]. Twin studies have shown that non-genetic factors often play an important role in the onset of disease. In spite of their significant genetic similarities, monozygotic twins may develop different diseases [38]. These differences seem to be linked to environmental effects and DNA methylation differences. Identical twins have been shown to have similar methylation patterns at younger ages, and different methylation patterns at older ages [17].

Studies (e.g. Fernandez et al. and Maegawa et al.) have shown that DNA methylation changes with age. In recent years researchers such as Hannum et al. and Horvath have started exploring the idea of predicting age from DNA methylation.

In 2013, Hannum et al. built an age predictor from methylation in whole blood of patients aged 19-101 using Elastic Net, which is a multivariate regression method. The age predictor performed well, achieving an error rate of 4.9 years when tested on a verification cohort. The authors showed that a model based on their blood methylation age predictor demonstrated similar age prediction abilities in other tissues, such as breast, kidney, lung and skin. The authors also found evidence suggesting that differences in apparent methylomic aging rate (AMAR), which is the ratio of one methylation age to one's chronological age, are biologically based and related to

individual differences in aging rates. Furthermore, the authors revealed a connection between age methylation markers and genetic factors. The study also showed that the majority of the age related methylation markers were located close to genes that are known to be associated with aging related conditions. This study also identifies 71 markers that are highly predictive of age.

Around the same time, Hannum et al. created a multi tissue Elastic Net based age predictor that estimates one's DNA methylation age using 8000 samples and 51 healthy cell types and tissues. According to the study, age prediction is accurate in most tissues. The author also shows that our rate of aging is logarithmic before adulthood and then linear in adults. In addition, age acceleration, which is defined by Hannum et al. as the difference between DNA methylation age and chronological age, is highly heritable. The study reveals 353 CpGs that are highly predictive of age.

This thesis further validates the above results and theories presented by the previous studies mentioned above. Similarly to previous studies, such as Fernandez et al., Maegawa et al., Hannum et al. and Horvath, this thesis shows that methylation changes with age.

Unlike the studies presented by Hannum et al. and Horvath where Elastic Net was used to predict age from methylation, the age predictors presented in this thesis use Random Forest and Linear Regression to predict age. In addition, this thesis mainly focuses on age related methylation in the prefrontal cortex of the brain, which is a tissue type that has not been thoroughly explored by these studies. In spite of the differences in prediction algorithms and tissue types, the results presented in this thesis show a high level of consistency with the results presented by Hannum et al. and Horvath. This high level of consistency further proves the theory that methylation can accurately predict age in multiple tissues and that the relationship between age and methylation is relatively strong since it can be accurately modeled with different algorithms and statistical tools. Another finding that is consistent with these two studies is that the age prediction residuals are associated with biological factors, such as aging rate. This finding further confirms that these biological factors

are associated with one's individual rate of aging. The high prediction accuracy of the Linear Regression based age predictor presented in this study further confirms Horvath's claim that the rate of aging during adulthood is linear.

This thesis provides additional proof for Hannum et al.'s findings that showed that age acceleration is genetically and epigenetically related and that there is a relationship between the genetic and epigenetic components involved. The fact that I show that genetics are associated with age acceleration further proves Horvath's claim that age acceleration is highly heritable.

Like Hannum et al. and Horvath this thesis identifies a set of markers that are highly predictive of age and consistently shows that their biological functions may be associated with aging related conditions and diseases. However, the number of identified age related markers differs across both studies and this thesis. In contrast with the markers presented by Horvath that seem to be highly predictive of age across most tissues, the age markers presented by this thesis that prove to be highly predictive in brain methylation, are only slightly predictive in blood methylation and there seems to be no overlap in highly significant age probes across the two tissues.

Chapter 3

Data and Statistical Methods

3.1 Introduction to the AD dataset

The dataset contains prefrontal cortex methylation data and yearly covariate data for 723 patients between the ages 66 and 108 years from two different studies (ROS and MAP). The patients begin the study when they do not have Alzheimers Disease (AD) and are assessed on a yearly basis. During each assessment each patient fills out a questionnaire and has his or her cognitive ability tested. The questionnaire includes questions related to the patient's health, environment, quality of life, race, gender and age. For example, one of the questions the patients are asked is how often they smoke.

The prefrontal cortex methylation data is extracted when a patient passes away. I also have blood methylation data for a small number of patients and genotype data for a large number of the patients. I have blood methylation data obtained from 44 patients when they joined the study and from 45 patients when they passed away. Only 42 of these patients have associated blood methylation samples from before and after death.

3.1.1 Pre-processing and data clean-up

The methylation data is initially adjusted for plates. Plates are like cookie sheets where each plate is a different batch inserted into the machine. Different plates may have been run on different days with slight differences in the biochemistry. Therefore, adjusting for the batch enables us to adjust for environmental (i.e. temperature and humidity) and biochemical variance.

3.1.2 Co-variates

The original dataset includes over 200 covariates. Only a small subset of these co-variates is included in this paper, the ones that seemed relevant to my research. The table 3.1 lists and defines the covariates included in this paper.

Covariate Name	Actual Name	Description
cAD	AD	Does the patient have Alzheimers?
lower extremity	lower extremity pain	Does the patient have lower extremity pain?
claudication cum	claudication	Has the patient reported pain in legs while walking that included their calves since he or she joined the study?
thyroid cum	thyroid disease	Has the patient ever had (past or current) thyroid disease?
cancer cum	cancer	Has the patient ever had (past or current) cancer?
hypertension cum	hypertension	Has the patient ever suffered from (past or current) hypertension?
heart cum	heart disease	Has the patient ever been diagnosed with one of the following: coronary, or coronary thrombosis, or coronary occlusion, or myocardial infarction?
cigcyn	stroke	Cerebral infarcts gross chronic?
cimcyn	stroke2	Cerebral infarct microscopic chronic?
fu year	how long	Indicates how long a patient has been in the study
age bl	age of baseline	How old the patient was when he or she joined the study
age death	age of death	How old the patient was when he or she passed away. This is also the age at which methylation data was obtained.
msex	gender	the patient's gender
cesdsum	depression	depression level determined by asking the patient 10 questions
smoking	is smoking	has the patient ever been a smoker
pkyr bl	smoking packs	packs per year
q2smol	ever smoked	has the patient ever smoked regularly?
diabetes sr rx ever	diabetes history	has the patient ever had diabete or possible diabetes?
educ	education	number of years of education
study	study	study type ROS or MAP
dxpark	parkinsons	has the patient ever had parkinsons disease?
bmi	bmi	body mass index

Table 3.1: List of Covariates

3.1.3 Introduction to the Blood dataset

The Whole Blood methylation data [1] presented by Hannum et al. [19] was obtained from the National Center of Biotechnology Information (NCBI) and used in this thesis.

3.1.4 Pre-processing and data clean-up

The methylation data was loaded from into R and the ID_REF column was removed. No additional processing was done. Since adjusting the data for plates seemed to make predictions significantly less accurate, the data remained unadjusted.

3.1.5 Co-variates

The covariates were obtained as a soft file and processed in R using GEOquery. The data included a total of 6 covariates including age, ethnicity, gender, plate, source and tissue. Age is the age of each patient in years, where each age is an integer. Ethnicity is a string covariate representing the ethnicity of each patient, including the strings "Caucasian - European" and "Hispanic - Mexican". Gender is a single character string containing the gender information of the different participants, where the gender is represented as "M" for male or "F" for female. Source is a string covariate representing one of the following locations: "UCSD", "Utah", "Boston", "USC". Plate is a number representing the batch (as explained in the previous section). This dataset contains 9 different plate numbers. Tissue is a string representing the tissue type from which methylation was taken. All the tissue types in this data set are "whole blood".

3.2 Statistical Methods

Statistical methods are methods used to collect, interpret, summarize and analyze data and are widely used in multiple disciplines [11]. In this chapter I discuss the statistical methods used in this thesis.

3.2.1 Classification

Classification maps features to labels [32].

3.2.2 Regression

Regression is a statistical tool that investigates the relationships between variables [43]. In this paper I mainly use two different types of regression, Linear Regression and Random Forest Regression.

3.2.2.1 Linear Regression

Linear Regression defines a linear relationship between variables [37]. In this thesis I use linear regression to estimate age from DNA methylation and to explore the relationship between different variables, such as age acceleration and age of disease onset.

3.2.2.2 Random Forest Regression

Random Forest is an algorithm that was invented by Breiman in 2001 and can be used for classification and regression. The algorithm combines tree predictors that depend on random vectors that are sampled independently [7]. The trees are grown based on a random subset on the features and each tree votes for a class [42]. In Random Forest Regression, the final result is given by taking average over the votes of all the trees. In classification, the final result is determined by the majority ([27]). In this paper I use the Randon Forest Regression to predict age from DNA methylation.

3.2.2.3 Cross Validation

Cross validation provides a way to estimate the accuracy of a method on new data. This evaluation is done by taking out some of the data, training the data on the remaining portion of the data and then testing the learned model on the data I took out before, and thus testing the method on "new data" [41].

K-Fold Cross Validation In K Fold Cross Validation, I split the data into k parts (folds) and then go through each part using it as testing data and the other parts as training data [41]. In this thesis I use K-Fold Cross Validation to evaluate the accuracy of the age predictors.

3.2.3 Statistical Tests

Statistical tests enable us to compare two different datasets [24]. These tests can be done at different confidence levels (I mainly used .95 and .99). In this thesis I mainly used R to perform statistical tests. I considered a result significant if its p-value was less than one minus the confidence interval.

3.2.3.1 T-test

T-tests provide a way to compare two means, even when the exact mean and standard deviations are not known, but can be estimated from sample data [24]. I used R's t.test function to perform this computation.

3.2.3.2 F-test

I use F test to compare standard deviations [3]. I used R's var.test, passing in sample data, to do this test.

3.2.3.3 Analysis of Varience

Analysis of Varience (ANOVA) compares the means of two or more independent groups [16]. In this thesis ANOVA is used to find combinations of covariates that are associated with age.

3.2.4 Principal Component Analysis (PCA)

Principal Component Analysis is an algorithm that reduces the number of dimensions used in the data by identifying the directions of maximum data variations. These

directions are linear combinations of the original data called principal components ([40].

In this thesis PCA is used to combine multiple covariates in a way that creates a significant relationship between this combination of covariates and age. This is done by running PCA on a large number of covariates and identifying principal components (PCs) that are highly associated with age.

Chapter 4

Random Forest and Age Estimation based on DNA Methylation

Understanding the relationship between age and methylation may help us better understand the biological causes of aging related disease. In this chapter I create an age predictor that uses Random Forest Regression to predict age from DNA methylation in the prefrontal cortex of their brain. The predicted ages have a low error rate of approximately 4.5 years and very few outliers. In later chapters I find a strong relationship between the signed prediction errors (age accelerations) and the age of onset of aging related disease, which suggests that these errors may be largely caused by biological clock differences.

4.1 Overall Prediction Power

In this section I will examine the prediction accuracy and the advantages of using Random Forest compared to Linear Regression.

4.1.1 Accuracy

The accuracy varies a little, since the patients in each fold are selected at random, where the selections change in each run. Overall, the prediction error is generally around 20 $years^2$, which is within approximately 4.5 years of the real age.

4.1.2 Prediction Advantages Compared to Linear Regression

Compared to Linear Regression, Random Forest seems to predict results that have smaller errors, are more stable, and are better at assessing the risk of developing an age related disease.

Comparing the results of Random Forest Cross Validation (rfcv) to those of Linear Regression Cross Validation (lmcv), we learn that Random Forest produces more stable (lower standard diviation across runs) predictions with a smaller mean error. First, I ran Random Forest Cross Validation (rfcv) and Linear Regression Cross Validation (lmcv) on the data 5 times. In order to compare the results, I ran a t-test on the prediction errors associated with lmcv and rfcv (based on [24], t test is a good way to compare two means). The p-value was 0.03558 less than .05, meaning that these results are statistically different. The fact that the p-value is relatively close to .05 indicates that although the error rates of the two different prediction algorithms are statistically different at confidence level .95, they are statistically the same at confidence level .99 (since the p-value is greater than than .01). The mean of rfcv was 19.92725 and that of lmcv was 21.15926, which is higher than that of rfcv. I compared the standard deviation using an F test (based on [3] F tests can be used to compare standard deviations). The resulting p-value was 0.01402 less than .05, meaning that the standard deviations of the two algorithms differ significantly. The standard deviation of the results produced by rfcv is more than 4 times smaller than the standard deviation of the ages predicted by lmcv, meaning that Random Forest seems to be a more stable way to predict age compared to Linear Regression.

As we will see in chapter 8, another benefit of Random Forest compared to Linear Regression is that the ages that Random Forest estimates seem to have a much

stronger relationship with the age of onset of several aging related diseases. This implies that the Random Forest provides a better way to estimate one's biological age.

4.2 Computational Techniques

4.2.1 Original Cross Validation

I decided to use a library called randomForest that can be imported into R, since my data has been analyzed in R. The randomForest library has a cross validation function, which I used to evaluate the performance of the age predictor and the optimal number of features to use. I used R's rfcv function with 5 folds to predict the patients' ages from their methylation profiles. In order to determine whether these results were random or predicted based on the methylation profiles, I ran cross validation again after permuting the ages (meaning that age from patient y is randomly assigned to the methylation data of patient z).

4.2.2 Unbiased Cross Validation

I changed the cross validation function to run on the 30, 50, 70, 80, 90, 100, 150, 180, 200, 250, 300 ,500, and 1000 most significant probes. I simply copied the code for the rfcv function in the randomForest R library and modified it to find the significant probes for each training set, and used those probes as features for training and testing. Thus, for each fold, the algorithm computes the significant probes, based on the training data only. In contrast, the biased predictor, used the predetermined set of probes associated with age that were computed based on the training and testing data, to predict age from methylation.

I used a small numbers of probes for the unbiased CV, because the original (biased) CV results showed that the predictor performed best for 100 probes. The top probes are picked by first running Random Forest on the training data, using all the significant probes as features and then taking the highest ranked probes returned by

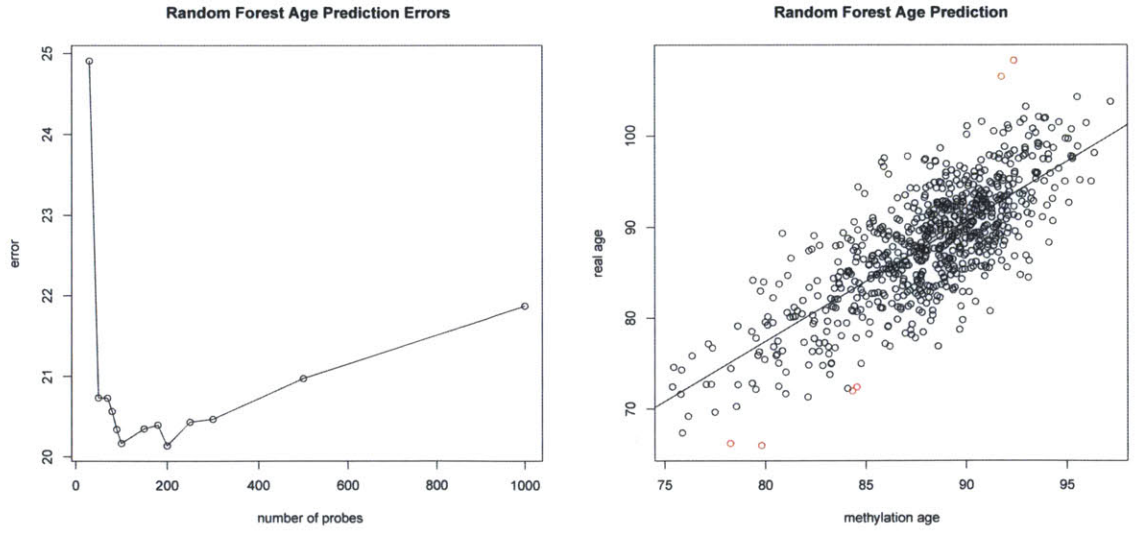
Rs importance function. In cases where the number of significant probes is too low, the top 11500 probes (with the lowest p-value) are used for the initial (initial for each fold) randomForest run whose result in combination with the importance function helps determine the desired features for each subsequent randomForest run on the same training set with less features.

4.2.2.1 Why Implement an Unbiased Version

In the original cross validation method, the significant probes are determined by the data from all the experiments rather than the experiments used for training in each fold. Due to this bias problem I implemented an unbiased cross validation function.

4.3 Results and Outliers

In this section I consider the results and the outliers and discuss how they were computed and what they mean.



(a) Age Prediction Errors based on Methylation using Random Forest

(b) Age Prediction based on Methylation using Random Forest

Figure 4.3.1: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms

Figure 4.3.1 shows the prediction error rates and the predicted ages compared

to the real ages. Figure 4.3.1a shows the error rates of the age predictions based on methylation. Figure 4.3.1b shows the predicted ages vs. the real ages, where the predicted ages were taken from the computation that uses 200 highly significant probes (since it had the lowest error rate). The outliers are colored in red and the error rate is 20.13697 $years^2$.

4.3.1 Results

Based on the results presented above, there seems to be a strong relationship between methylation and age that can be predicted relatively accurately (within 4.49 years) based on methylation in the prefrontal cortex of the brain.

In order to determine the accuracy of the results, I created another plot that shows the relationship between the age predictions and randomly ordered ages, meaning that the predictor is trained on random methylation profiles and ages that are not associated with these methylation profiles. This random plot is shown in Figure B.0.4b in the Appendix. To permute the ages I simply used R's sample function and with patients' real ages as an argument. Figure B.0.4 in the Appendix compares the plot of methylation ages vs. real ages to that of the randomly ordered ages. Looking at the differences between the two plots, it is obvious that there is a relationship between age and methylation and that only predictors that are trained and tested on real methylation profiles and their associated ages predict age accurately.

4.3.2 Outliers

Overall there are 6 outliers. The real age of 2 of the patients is significantly older than their methylation age and the real age of 4 of the patients is significantly younger than their predicted age. The outliers seem to mainly include the oldest and youngest patients in the study. One possible explanation is that the age prediction computation may not be as accurate for patients around the edges, since there are very few patients who are around the same age in the study. In fact, the outliers in the plot of randomly ordered ages (B.0.4b) are all around the edges. However, unlike the outliers

in the randomly ordered plot which seem to have random methylation ages (mainly located in the middle of the methylation age range), the outliers in Figure 4.3.1b have methylation ages that are significantly closer to the real ages, where the outlier methylation ages are also close to the edges. This difference suggests that there may be a biological cause for the outlier points.

Age acceleration is another possible cause for the outlier points. Based on the plot, the chronologically younger outlier patients had high age accelerations (their methylation ages was very high compared to their real ages) and the chronologically older outlier patients had low age accelerations (their methylation ages was very low compared to their real ages). This theory may also help explain why the chronologically older outlier patients lived over 100 years and longer than the other patients in the plot.

4.3.2.1 Outlier Computation

The outliers were computed in R, taking any value that is greater than the third quartile plus 1.5 times the interquartile range and any value that is smaller than the first quartile minus 1.5 times the interquartile range. The interquartile is the difference between the third quartile and the first quartile. The quartiles were computed in R using the quantile function [15, 34]. The first quartile is the sample or number that is smaller than 75% of the samples and the third quartile is the sample or number that is larger than 75% if the samples [28]

4.4 What we learn biologically

Biologically we learn that age can be estimated relatively accurately based on methylation data when using Random Forest. These results are consistent with previous studies such as Hannum et al. [19] and Horvath [21], where age was predicted based on methylation .

In Chapter 8, we will see that the estimated methylation ages produced by this algorithm can potentially help determine the risk of the onset of aging related disease.

These estimated ages seem to provide a better way to assess the risk of developing an age related disease than real age. These results may potentially have some significant applications in biology and medicine.

Chapter 5

Linear Regression and Age Estimation based on DNA Methylation

In order to try to improve the results produced by Random Forest, I used Linear Regression (lm) to estimate the ages of the research participants based on the methylation in the prefrontal cortex of their brains. The accuracy of the ages predicted using Linear Regression is similar to that of Random Forest, approximately 4.5 years. Later on we will see that similarly to the signed prediction errors (age accelerations) produced by Random Forest, the signed errors of the ages predicted by Linear Regression have a relationship with the age of onset of age related disease. However, this relationship is weaker than that of the age accelerations produced by Random Forest. Another fact that we learn in later chapters is that a relatively large portion (2.5%) of meQTL SNPs is associated with age accelerations predicted by Linear Regression, which implies that these age accelerations have strong genetic influences.

5.1 Overall Prediction Power

In this subsection, I look at the high accuracy of the algorithm and present the algorithm's advantages compared to Random Forest.

5.1.1 Accuracy

Based on the results presented here, the prediction accuracy is relatively high. I computed the error rate by considering the result corresponding to the number of features with the lowest error rate (150 probes). The error rate is 21.83377 $years^2$, which is approximately 4.67 years.

5.1.2 Prediction Advantages Compared to Random Forest

The prediction performance of Linear Regression is similar to that of Random Forest, but still has significantly larger errors. However, compared to Random Forest the age accelerations produced by Linear Regression are associated with a much larger fraction of meQTL SNPs.

5.2 Computational Techniques

5.2.1 Unbiased Cross Validation

In order to avoid the bias described in the previous chapter, I implemented an unbiased cross validation function for Linear Regression. The implementation is very similar to that of the unbiased version Random Forest Cross Validation.

5.2.2 Implementation

I implemented an unbiased Linear Regression cross validation method in R by tweaking the existing unbiased Random Forest cross validation method described in the previous chapter. For each fold, the function finds the significant probes in age prediction from the testing data and uses the predictor to predict the ages in the testing data. The most predictive probes in each fold will be intersected with the ones in the other folds and returned.

5.3 Issues

My implementation of Linear Regression Cross Validation (lmcv) fails to compute the error for the lowest and the heights number of probes (lowest=0, highest=10000000). Therefore, the error corresponding to 30 probes and the error rate corresponding to 1000 probes are not presented, analyzed or considered here.

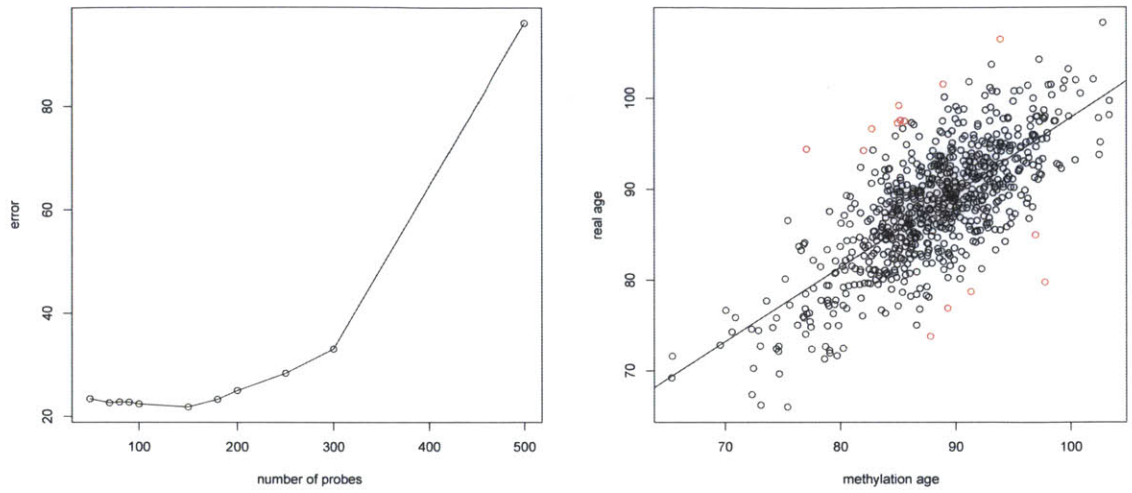
5.4 Outliers and Results

I created a few plots that show the prediction results produced by Linear Regression, the error rates and the outliers. These results are depicted in Figure 5.4.1, where Figure 5.4.1b shows the relationship between the real ages and the predicted ages and Figure 5.4.1a shows the error rates based on the number of features used. A random age plot that predicts age based on methylation and permuted ages, B.0.5a, can be found in the Appendix. The outliers were computed in the same way as they were computed in Chapter 4 as described in 4.3.2.1.

Above, the outliers have a wide range of chronological ages and mid to high methylation ages. Unlike the results produce by Random Forest whose outliers seemed to be located around the edges, the ages predicted by Linear Regression are more scattered and they also make up a relatively large fraction of the predicted ages. Overall there are 14 outliers, approximately 1.94% of the total number of predicted ages, which is more than 2 fold higher than the number of outlier ages produced by Random Forest.

There are several possible explanations for the outliers, including prediction inaccuracy, a biological cause, such as biological clock differences or a combination of the two. The fact that the methylation ages of the outliers are affected by the real ages implies that they are more than just computation errors. However, similarly to the random plot, the outliers here are located in the center of the methylation age range meaning that there is probably also a significant error component associated with these results. Unlike the Random Forest outliers, the polarity of the outliers'

chronological age is generally not the same as that of their methylation ages.



(a) Age Prediction Errors based on Methylation using Linear Regression

(b) Age Prediction based on Methylation using Linear Regression

Figure 5.4.1: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms

5.5 What we learn biologically

Here I show once again that there is a clear relationship between age and methylation, a relationship that was shown to exist in previous studies, such as Hannum et al. [19] and Horvath [21]. The fact that Linear Regression accurately predicts age from methylation provides further evidence for Horvath's theory that age increases linearly with age in adults. Similarly to Random Forest, given patient brain methylation data, this algorithm seems to estimate age relatively well, with an error rate of about 4.67 years. Approximately 2% of the results are outliers, which is relatively large compared to Random Forest. The outliers seem to be caused by computational errors and biological causes.

Later on we will see that the meQTL SNPs associated with age accelerations of the results that were presented here make up a large portion of SNPs, which implies that the age accelerations computed by Linear Regression have a strong genetic component. We will also see that these age accelerations have a stronger relationship with the onset of an aging related disease than real age does, however, this relationship

seems to be significantly weaker than the relationship between the methylation ages produced by Random Forest and the age of onset of aging related disease.

Chapter 6

Adjusting for Covariates and Environmental Factors to Improve Age Estimation based on Methylation

Often environmental factors that affect methylation can hide the methylation changes caused by the disease, making it more difficult to correctly identify disease specific methylation patterns. I previously examined such environmental factors and their effects on DNA methylation in the Dorsolateral Prefrontal Cortex of the brains of Alzheimers disease (AD) patients and controls. The patient data was analyzed while adjusting for these environmental effects, which made it easier to identify disease specific contributions to methylation in some cases. For example, correcting for covariates such as parkinson's, apoe4n (a gene that seems to be related to AD), and smoking, seemed to help predict AD.

In this chapter I try to improve the age prediction accuracy by correcting for environmental covariates, such as smoking. Although I try a large number of covariates, it seems that none of the covariates improves the accuracy of the predictions. In fact, correcting for single covariates generally makes predictions worse and no combination

of pairs covariates seems to be significant enough to try to correct for any of the pairs.

6.1 Covariate Combinations

In this section I identify covariates and combinations of covariates that are associated with age. This is done in order to determine which covariates are worth trying to correct for in order to reduce the error rate associated with age prediction. First I identify single covariates that are associated with age and then I identify multiple covariates, pairs and groups, that are related to age.

6.1.1 Single Covariate and its Relationship with Age

In this subsection, I identify covariates that have a significant relationship with age in order to try to reduce the error rate of age prediction later on. I identify these covariates by computing the linear regression on each covariate with age and looking at their p-values. I convert ordinal covariates into factors (an R type) before passing them into lm, while numeric covariates remain the same. The significance of relationship of each covariate with age is depicted in Figure 6.1.3.

6.1.2 How Relationship between Multi Covariate Combinations Relate to Age

In this subsection I identify pairs and groups of covariates that are associated with age. Then, I correct the methylation data for these groups of covariates in order to try to improve the prediction errors of age from methylation.

6.1.2.1 Identifying Pairs of Covariates that are Associated with Age

I used ANOVA and linear regression to try to find meaningful combinations of covariates, where 2 covariates make up the model and age is the response. The pseudocode is depicted in Figure 6.1.1. I used the algorithm described below on multiple covariates, one pair at a time. First I use linear regression to construct two different

models, m1 and m2, where m1 uses two covariates seperately and m2 combines them. The result of ANOVA on m1 and m2 is stored in m3. The p-value is obtained from m3. The p-value determines whether a combination of the two covariates is more strongly associated with age than using them separately. This is done in order to determine which pairs of covariates have strong associations with age when the two covariates are work together. The goal is to adjust for significant pairs of covariates in order to reduce the prediction error. Unfortunately, none of the combinations seem meaningful (p value of less than .05 after adjusting for the number of tests).

```
m1=lm(outarg ~ inarg1+inarg2)
m2=lm(outarg ~ inarg1*inarg2)
m3=anova(m1,m2,test="LRT")
```

Figure 6.1.1: Computation of the Relationship between Multi Covariate Combinations and Age

6.1.2.2 Using PCA to Find Groups of Covariates that are Associated with Age

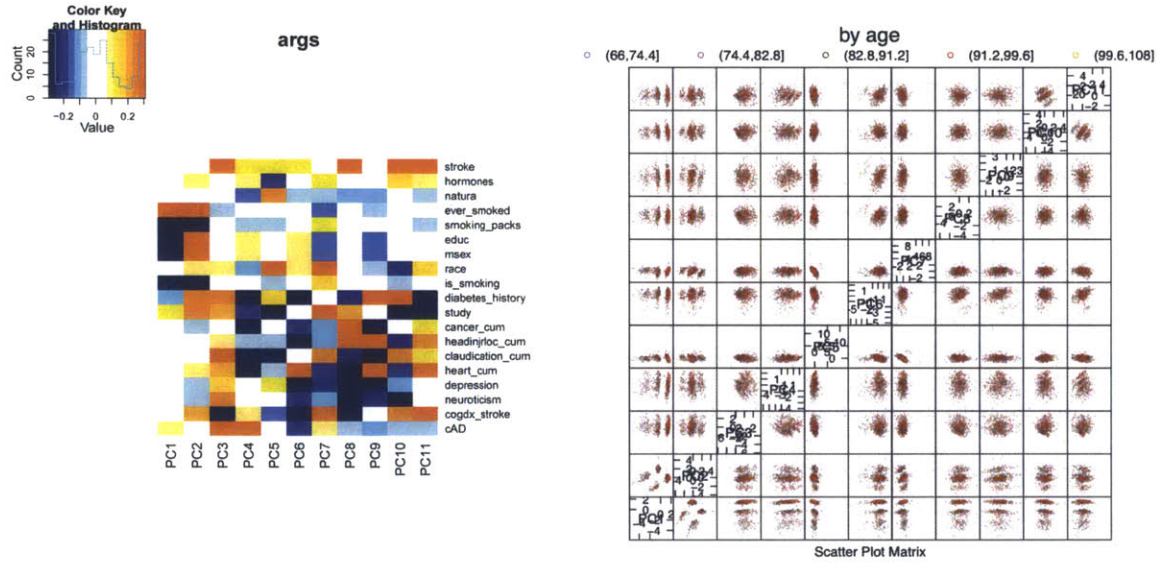
PCA was used to identify combinations of covariates as principal components (PCs) that are strongly associated with age. The goal is to correct for these significant PCs when predicting age from methylation in order to increase the accuracy of the predictions. PCA was computed using R's pca method in the pcaMethods package with completeObs. CompleteObs tells R to replace NAs or unknown values with estimates. Overall 19 covariates were reduced to 11 PCs, PC1 through PC11.

The loadings, which are the correlations between the principal components and the original variables [2], are displayed in Figure 6.1.2a. Each entry in the heat map is color-coded by the magnitude of the correlation between the corresponding PC and the corresponding covariate. Darker colors are associated with stronger values of absolute correlations, for example dark blue indicates a larger absolute correlation value than light blue. Yellow-orange values represent positive correlations, blue values are associated with negative correlations and white values indicate little to no correlation.

The scores, the projections in PC space [33], are displayed in Figure 6.1.2b and

color coded by 5 different age groups. This plot is symmetric along the bottom left to top right diagonal. Each of the subplots displays a scatterplot of the scores projected onto the PCx PCy plane where x is the column number from the left and y is the row number from the bottom. For example the scatter plot in row 1 column 2 shows a scatter plot of PC1 vs PC2.

In order to identify PCs that are associated with age I look for groups of points that are separated by color or age group in Figure 6.1.2b. The covariates that are strongly associated with these PCs can be found by looking at the entries corresponding to these PCs in Figure 6.1.2a and identifying the covariates with the dark colors.



(a) Heatmap of covariate combination

(b) Scatter plot colored by age

Figure 6.1.2: Heatmap and Scatterplot showing how the different covariates are associated with age

PC1 and PC2 look interesting. PC1 is largely associated with smoking (all 3 smoking covariates), education and gender. Its points in Figure 6.1.2b are divided into 3 clusters in 9 out of its 10 sub plots. These sub plots are located in row 1 from the bottom columns 3-11 from the left. In the remaining sub plot (PC1 vs. PC2), which is located in row 1 column 2, the points are divided into 4 clusters. In most of the PC1 subplots the older points are located on the right side of each plot. The plots associated with PC2 look very similar to those of PC1 and are located in the 2nd row from the bottom. In PC2 the majority of the points corresponding to the

older patients seem to be located in the middle cluster whereas in PC1 they seem to be mainly located in the upper cluster.

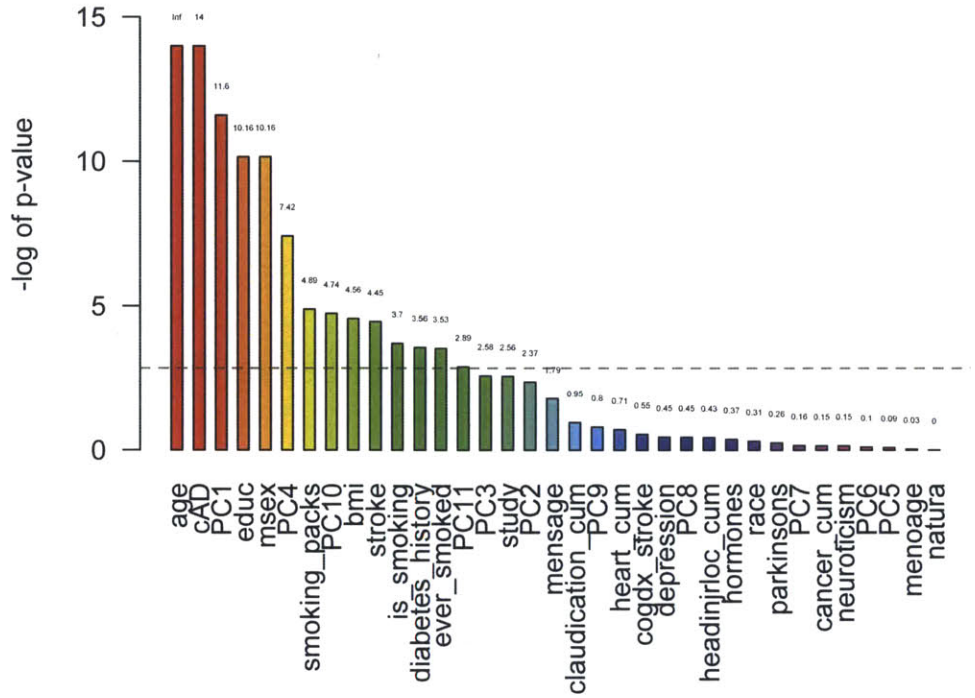


Figure 6.1.3: P-values of a covariate as the model and age as the response variable

Figure 6.1.3 shows the significance of the relationships PCs and covariates have with age. This Figure shows the log value of the p-value of the linear regression. All the values on the left side of the plot above are greater (meaning that their p-values are smaller) than the significance limit, meaning that they are significant. The dashed line shows the p-value above which a value is considered significant (all the columns that are taller than this lines correspond to covariates that are significantly associated with age).

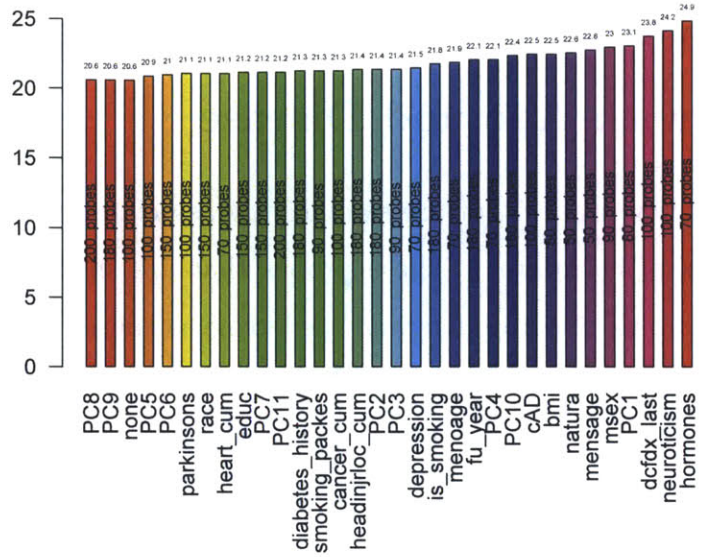
PC2, like PC1, is strongly associated with smoking, education and gender. How-

ever some of its smoking components are inversely associated with age compared to PC1. Furthermore, PC2 is also moderately associated with the covariates diabetes and study, whereas PC1 is only mildly associated with these two covariates.

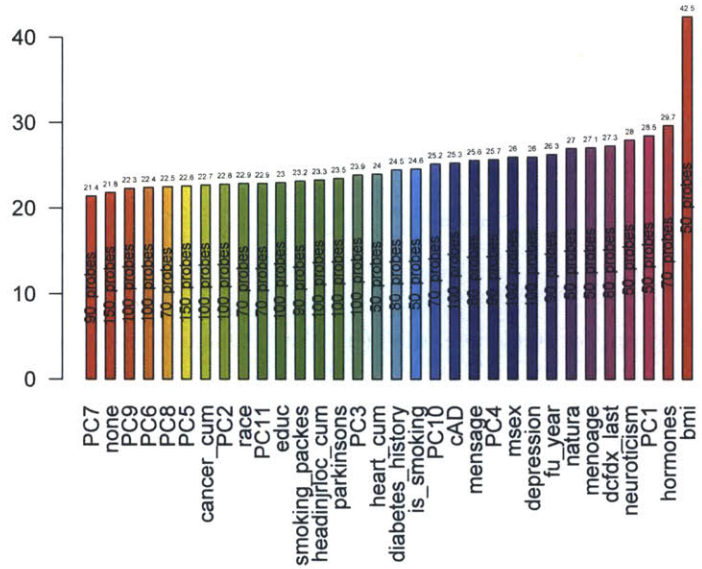
6.2 Prediction Error when Correcting for Covariates

Adjusting for covariates generally does relatively little when using Random Forest to predict age. The prediction error rates are generally very similar, 20.6-24.9 $years^2$, with the three values with the largest error having relatively larger errors (23.8-24.9 $years^2$), where the values with the largest errors are dcfdx_last, neuroticism, and hormones. Based on Figure 6.2.1, adjusting for covariates either increases or does not change the prediction error, since not correcting yields in one of the smallest errors.

Linear regression seems to be slightly more affected by covariate correction. Similarly to Random Forest, linear regression does not seem to predict better when the data is corrected for different covariates. In fact, it seems to predict worse in most cases. Here, the errors caused by adjusting for covariates seem to be larger than those of Random Forest. The largest prediction error is 42.5, which occurs when correcting for bmi. Correcting for bmi seems to cause a large increase in the prediction error, possibly because a relatively large portion of the bmi data is unknown, which makes the function run 5 fold cross validation on a relatively small number of experiments compared to data corrected for other covariates.



(a) Prediction error by covariate corrected for error - rf



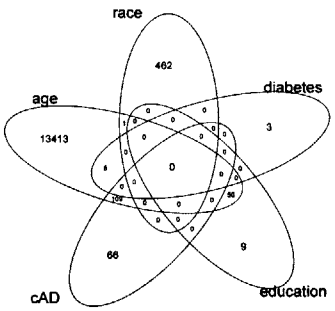
(b) Prediction error by covariate corrected for error - lm

Figure 6.2.1: Prediction error by covariate corrected for error - lm and rf

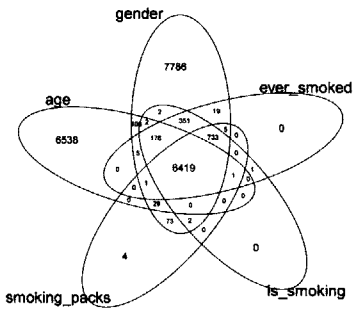
Figure 6.2.1 shows the prediction error when adjusting for different covariates. The results are shown for two different prediction algorithms Random Forest (6.2.1a) and Linear Regression (6.2.1b). The covariate "none" indicates that I did not correct for any covariate (except for plates, as explained in Chapter 3).

6.3 Identifying Methylation Probes that are Associated with Age and Disease

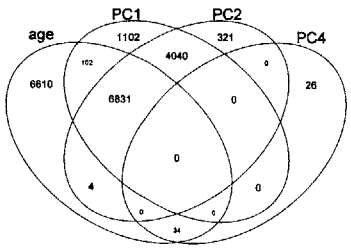
In this section I identify probes that are significantly associated with age and other covariates or groups of covariates. These probes that are shared across different covariates and age help us better understand the biological causes of the relationships between age and covariates that I explored before. Furthermore, I will later try to use these shared probes to try to reduce the error rates of the age prediction. The Venn Diagrams below show which significant probes are shared across different covariates.



(a) Venn Diagram for age, race, diabetes, education and cAD (AD)



(b) Venn Diagram for age, gender, ever smoked, is smoking and smoking packs (number of smoking packs)

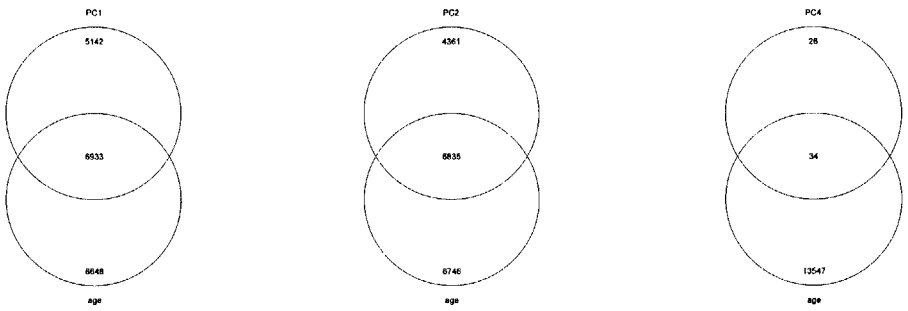


(c) Venn Diagrams for age, PC1, PC2 and PC4

Figure 6.3.1: Venn Diagrams for Different Covariates

Figure 6.3.1 shows the number of methylation probes that are shared across dif-

ferent covariates, PCs and age. Based on Figure 6.3.1a there seems to be a relatively large number of probes shared between age, smoking related covariates (is_smoking, smoking_packs and ever_smoked) and gender (msex in Figure 6.1.3). Age seems to share a smaller number of probes with cAD (AD) and education, as shown in Figure 6.3.1b. Some of the PCs seem to share a small number of significant probes with other PCs and some PCs share no probes with the other PCs (these PCs are not shown here). Figure 6.3.1c shows 3 PCs that share a large number of significant probes with age in a 4 variable Venn Diagram. Figure 6.3.1c also shows that some of these PCs (e.g. PC1 and PC2) also share a large set of significant probes with each other and with age. Figure 6.3.2 below shows the number of probes each of these PCs shares with age in 2 circle venn diagrams, focusing on the shared probes between these PCs and age. The vast majority of the covariates and PCs that share a non negligable number of significant probes with age are also shown to be significantly associated with age in Figure 6.1.3. This helps explain one of the biological causes of significant relationships between covariates or PCs and age.



(a) Age and PC1
Here, PC1 shares 6933 significant probes with age.

(b) Age and PC2
PC2 shares 6835 significant probes with age

(c) Age and PC4
Here, 34 probes are shared between PC4 and age

Figure 6.3.2: Venn Diagrams for Age and PCs that Share Significant Probes with Age

Figure 6.3.2 shows the 3 PCs (PC1, PC2 and PC4) that share significant probes with age. Both PC1 and PC2 (mainly associated with education, gender, and all 3 smoking covariates) have a relatively large number of significant probes shared with age. Based on Figure 6.3.2a, 6.3.2b, and Figure 6.3.1c, PC1 and PC2 share a large number of significant probes with each other and with age. This is not surprising

given that PC1 and PC2 are associated with roughly the same covariates, based on Figure 6.1.2a, and are also associated with age based on Figure 6.1.2b.

PC4, which is mainly associated with heart cum (heart cumulative history), claudication cum (claudication cumulative history), cancer cum (cancer cumulative history) and diabetes history, has a significantly smaller number of significant probes shared with age compared to PC1 and PC2. This may help explain why Figure 6.1.2b shows no association between PC4 and age.

Most of the other PCs (PC3, PC5-PC11) share no significant probes with age. In fact, other than PC9 that shares 1 significant probe with age, no other PC out of the PC5-PC11 group shares any significant probes with age.

6.4 Causes of Increase in Error after Adjusting for Covariates

Earlier I saw that adjusting for covariates tends to increase (or in some cases has no effect on) the prediction error. There are two possible causes for the increase in error after correcting for covariates, one possible cause is having a large number of unknowns in the covariate data and the other is the fact that adjusting for covariates that are highly associated with age may correct for age and thus reduce its predictability.

6.4.1 Adjusting for Covariates with Lots of Unknowns Increases Prediction Errors

One cause seems to be the number of unknowns in the data for each covariate, which is related to the number of experiments that are passed into rfcv and lmcv. When adjusting for a specific covariate I only adjust experiments that have data for this covariate. Therefore covariates that have a lot of unknowns result in less experiments being returned from the data correction function and thus less experiments being passed into rfcv and lmcv. Running 5 fold cross validation on less data may result

in larger error rates. The high prediction error rate of lmcv when adjusting for bmi, illustrates this concept. The bmi covariate has a large number of unknowns which causes a very low number of experiments to be used in rfcv and lmcv and thus a low number of experiments are used in each fold. The error rate is 42.5, which is much higher than the prediction error rate for data adjusted for any other covariate.

There are two ways to test this theory, one is to run cross validation with less folds when correcting for covariates with a large number of unknowns. Another possible solution could be not to adjust experiments that are associated with unknown data instead of dropping them. In my opinion the former solution is more correct because it does not adjust some experiments and and not adjust other experiments, it simply drops all the experiments that cannot be corrected for a specific covariate and adjusts the number of folds accordingly, thus insuring that the number of experiments in each fold is sufficient.

I tried to solve this problem by adjusting the number of cross validation folds based on the number of experiments I use to perform the the cross validation. I reduced the number of folds for data with less experiments. However, the resulting errors were larger because the training sets were smaller.

Next, I tried to use more folds in order to increase the number of experiments that are used for training in each fold. Unfortunately, this solution did not seem to reduce the prediction error. It seems that the only way to improve the errors related to unknown covariate data is to add additional experiments with known data. Also, in general, increasing the number of experiments in the study may reduce the prediction errors for most or all covariates.

6.4.1.1 Adjusting for Covariates that are Correlated with Age

I looked at the relationship between age and covariate and compared it to the prediction error that results from adjusting for this covariate. This was done in order to determine whether I adjust for important age data when adjusting for covariates that are highly correlated with age, and thus increase the prediction error. The result plots were compared to results with permuted ordering to make sure the perceived

effects of the age related component in each covariate are not random.

Based on the results there does not seem to be a relationship between age prediction error and the relationship between age and the covariate I am adjusting for (For more information see Figure B.0.6 in the Appendix). One possible reason for this is that the error caused by having a large portion of unknown covariate data and thus a small number of experiments makes this relationship more difficult to find.

In order to check if adjusting for covariates without adjusting for their age components would work better, I tried to adjust the data in 3 different ways. I tested these potential solutions on some of the covariates, using Random Forest Cross Validation to predict age and evaluate the results. First I tried to identify principal components that were not correlated with age, adjust for them and use Random Forest Cross Validation to evaluate the age prediction error. However, this did not help reduce the error. I also tried to solve the problem by reducing the number of methylation probes by running PCA on the methylation matrix, but this did not reduce the error rate. Another solution I tried was to only adjust for probes that were not predictive of age when correcting for a specific covariate. Unfortunately, this solution failed to improve the prediction results.

Chapter 7

Methylation Probes that are Highly Predictive of Age

The following chapter explores probes that are highly predictive of age, meaning that they are significant in at least one cross validation fold. In each fold the 100 most important probes are selected (in Random Forest Cross Validation this is determined by R's importance function and in Linear Regression Cross Validation it is determined by p-values). I pick the most important 100 probes because multiple runs of Random Forest Cross Validation and Linear Regression Cross Validation have shown that the error tends to be relatively low (though not always the lowest) when using 100 probes to predict age from methylation. Next I look at the chromatin states and GO terms of probes that are predictive of age in several folds (2+ folds or 5 folds) and examine the similarities and differences between the two prediction algorithms. GO terms are consistent definitions of gene products [46] and will be further discussed in Section 7.3. Compared to general methylation probes, highly predictive age related probes seem to be associated with different chromatin states. I then look at the biological processes associated with the methylation probes in each chromatin state. The resulting terms suggest that age related probes may be associated with heart related functions, arterial blood pressure, memory, learning, cognition and processes related to neurons and axons and thus imply that aging related probes may be involved in aging related conditions such as memory deterioration, high blood pressure and heart

disease.

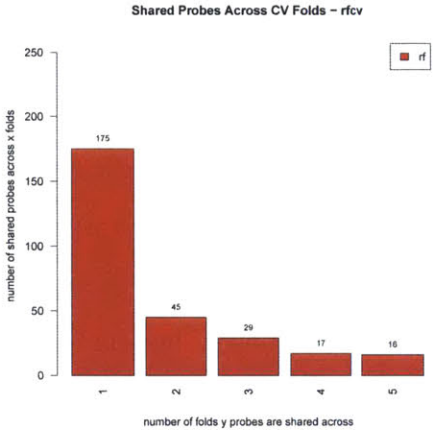
In my analysis I mainly focus on age probes that are shared across all 5 folds or at least 2 folds. Probes that are shared across all 5 folds are consistently predictive of age across multiple tests, which indicates that they are highly affected by age. Methylation probes that are shared across more than one fold show some level of consistency, as they were used to predict age more than once, meaning that they were probably not selected by mistake or by chance. Although probes that are shared across 5 folds show a higher level of dependence on age, age probes that are shared across 2 or more folds may also be important since they provide a larger set of age related probes that possibly include missed age related probes and methylation probes that change with age only in certain groups of individuals.

7.1 Shared Probes Across CV Folds in RF and LM

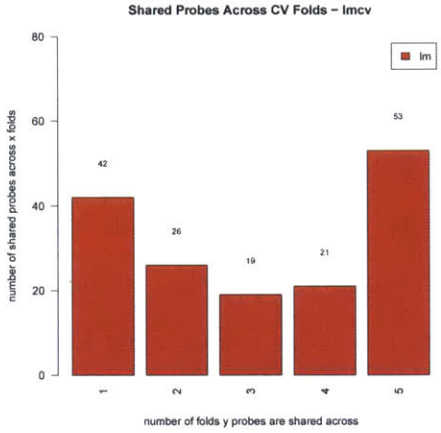
I explore the highly significant probes that are shared across different folds in Random Forest Cross Validation and Linear Regression Cross Validation in order to identify probes that are consistently highly predictive of age.

Figure 7.1.1 shows the number of highly predictive probes that are shared across x folds, where x is between 1 and 5. The results for two different algorithms are compared here, Random Forest and Linear Regression. Figure 7.1.1a shows the results for Random Forest. Here, 16% (16 out of 100) of the highly predictive probes in each fold are shared across all 5 folds. Figure 7.1.1b depicts the results for Linear Regression. In this plot, 53 (>50%) of the highly predictive probes in each fold are shared across all 5 folds. Figure 7.1.1c shows the results for both algorithms stacked on top of each other. In Figure 7.1.1d, we see the intersection of the highly predictive age probes found by each of the two prediction algorithms. The probes are intersected based on the number of folds they are shared across. For example, the 29 probes that are shared across 3 folds in Random Forest Cross Validation are intersected with the 19 probes that are shared across the same number of folds in Linear Regression Cross Validation. Only two of the probes that are shared across three folds in each

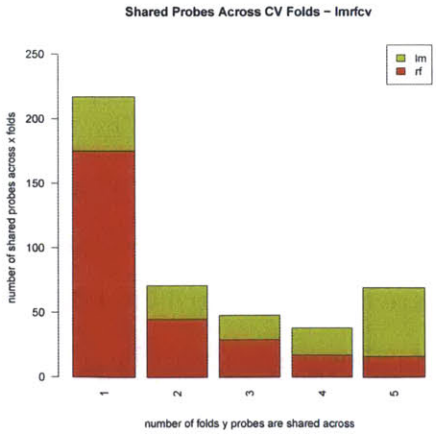
of the algorithms are shared across three folds in both algorithms. It is interesting to note that 13 probes (81.25% of the folds that are shared across all 5 folds in rfcv) are highly predictive of age in all 5 folds of both algorithms. This high level of consistency suggests that these probes are highly affected by our biological age, which also implies that these 13 probes are probably related to age related changes in the prefrontal cortex after age 64. In the next two sections I will explore the chromatin states and biological functionalities of these probes and try to determine how they are related to human aging and aging related disease.



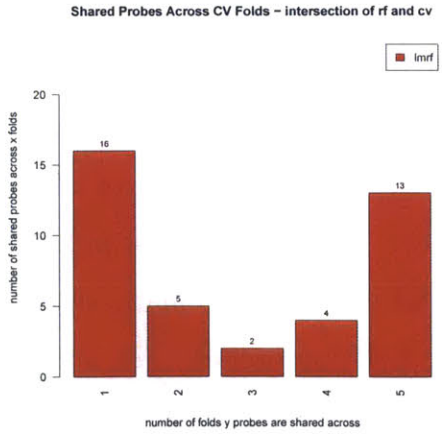
(a) **Random Forest** Highly Predictive Age Probes and the Number of Folds they are Shared Across



(b) **Linear Regression** Highly Predictive Age Probes and the Number of Folds they are Shared Across



(c) **Random Forest and Linear Regression** Intersection of Highly Predictive Age Probes



(d) **Random Forest and Linear Regression** Highly Predictive Age Probes and the Number of Folds they are Shared Across

Figure 7.1.1: Shared Probes Across CV Folds

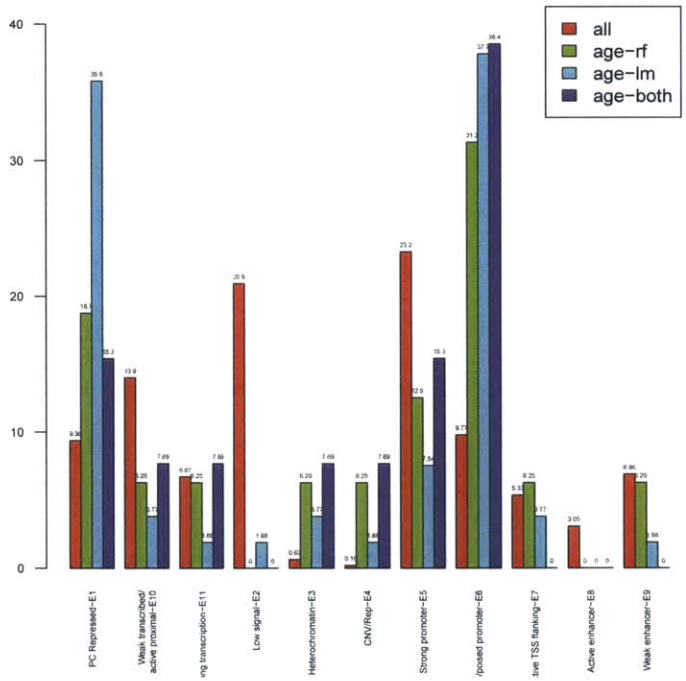
7.2 Chromatin States of Highly Predictive Probes

In the previous section I identified sets of significant probes that were consistently predictive of age across multiple tests (folds). In this section I examine the chromatin states of these highly predictive age probes and compare them to the chromatin states of general methylation probes. Here I find that the distribution of the age predictive probes across the chromatin states differs from that of the general probes. Unlike general methylation probes that tend to be strongly associated with Low Signal (E2) and Strong Promoter(E5) states, age probes are mostly related to Inactive/Poised Promoter (E6) and PC Repressed(E1) states, only moderately associated with Strong Promoter(E5), and very weakly represented in Low Signal (E2) states.

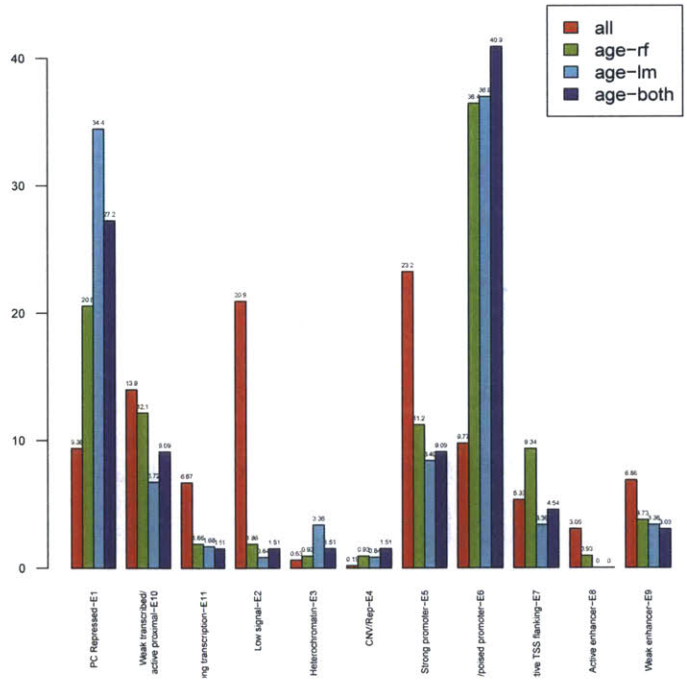
Figures 7.2.1a and 7.2.1b show the chromatin states corresponding to the highly predictive probes found across all 5 folds and more than one fold respectively. The states of the highly predictive probes found using Random Forest (labeled "age-rf") are compared to those found using Linear Regression (labeled "age-lm"). In addition, the states are also compared to the general distribution of all 480000 probes across the different states, which is labeled "all", and to the intersection of the highly significant probes found using Random Forest with those found using Linear Regression (labeled age-both).

Overall, the chromatin state distribution of the probes is relatively consistent across the two algorithms and the intersections of their results. The distribution of the results also differs from that of all the existing probes (most of which are not predictive of age). This implies that probes that are highly predictive of age are strongly associated with specific chromatin states and are therefore distributed differently than general methylation probes. The above charts show that a very large percentage of the highly predictive age related probes are PC Repressed (E1) and poised promoters (E6) and a moderate number of them are located in weak transcribed/ active proximal (E10), Active TSS flanking (E7) and Strong Promoter(E5) states. In contrast, general methylation probes are mostly low signal (E2) and Strong Promoters(E5) and only a moderate percentage of them (slightly less than 10% of them in each group)

are PC Repressed (E1) and poised promoters (E6).



(a) States of highly predictive probes that are shared across all 5 folds



(b) States of highly predictive probes that are shared across two folds or more

Figure 7.2.1: States of highly predictive probes that are shared multiple folds

7.3 Biological Functions of Highly Predictive Probes

In this section I explore the biological functions of methylation probes that are highly predictive of age. I do this by finding the Gene Ontology (GO) terms associated with the highly predictive age probes separately for each chromatin state and prediction algorithm.

Gene Ontology is a project that aims to consistently describe gene products. These gene products are described by 3 different vocabularies: biological processes, cellular components and molecular functions. Biological processes are sequences of molecular functions. Cellular components are cell components, such as anatomical structures like the nucleus. Molecular functions are activities performed by one or multiple gene products [46].

I used Rs hyperGTest to find the GO terms overrepresented in the highly predictive age probes. I mainly focused on finding the associated biological processes.

7.3.1 GO Categories of Highly Predictive Probes

In this section I present the GO terms associated with age probes for chromatin states that I found to be highly or moderately associated with age related probes, such as PC Repressed, Poised promoter and Active TSS flankin. For the full results see A.1.1 and A.2.1 in the Appendix. Some terms are associated with Random Forest (rf) only, Linear Regression (lm) only or the interaction of their age related probes.

7.3.1.1 E1 - PC Repressed

Here I examine the biological functions associated with 2+ fold and 5 fold age related probes that are associated with the PC Repressed state. Probes that are shared across 2+ folds seemed to be associated with metabolic processes (rf and lm only), morphogenesis (rf and lm only), development, biosynthetic processes (rf and lm only), nervous system development (rf and lm only), axon related processes (rf and interaction only) and neuron differentiation (rf and interaction only) and generation of neurons (lm only). The processes found in the 5 fold age probes were similar to those

found in the interaction of the 2+ fold age probes.

7.3.1.2 E6 - Poised promoter

In this subsection look at the biological functions associated with 2+ fold and 5 fold age related probes that are associated with Poised promoters.

The 2+ fold age probes are involved in regulation of metabolic processes and biosynthesis processes (rf and lm only), ossification (rf and lm only), development (including forebrain development), forebrain neuron related processes. The 5 fold probes seemed to be mostly related to development and neuron related processes.

7.3.1.3 E7 - Active TSS flanking

Here I explore some of the interesting biological functions that are associated with 2+ fold and 5 fold age related probes that are related to Active TSS flanking. The lm-rf interaction age probes are not associated with any GO terms, so I only present biological functions that are associated with significant age probes in Random Forest and Linear Regresion.

The 2+ fold probes are associated with cognition and memory (lm), learning and memory (rf), heart related processes, such as regulation of heart contraction, cardiac muscle related development, heart rate regulation, fear response, muscle related processes and positive regulation of systemic arterial blood pressure. The 5 fold age probes the Linear Regression results were similar to those of the 2+ fold probes. However, the only 2 GO terms that were associated with the Random Forest age probes, were neuroblast proliferation and B cell differentiation. The interaction probes had no associated GO terms.

7.3.2 Summary of Age Probe Biological Functions

Interestingly, the results presented in this section suggest that age related probes could be associated with heart related functions, arterial blood pressure, memory, learning, cognition and processes related to neurons and axons. These are systems that often

cause problems as we age. These results may help explain why older individuals are more likely to develop conditions such as high blood pressure, memory problems (AD, dementia, or just general deterioration) and heart disease.

Chapter 8

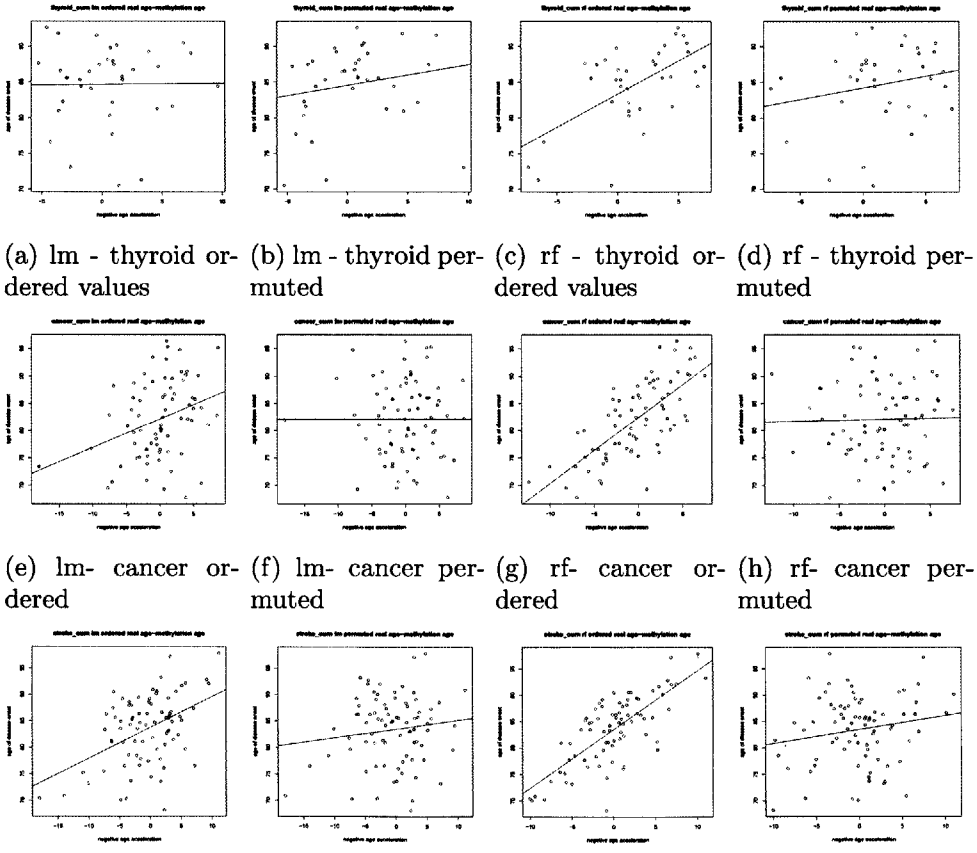
Age Acceleration and Disease Onset

Age acceleration has previously been linked to the ticking rate of one's biological clock and shown to change significantly in cancer cells [21]. In this chapter I examine the relationship between the age of disease onset and age acceleration. Each patient's methylation age at the time of death is used to approximate her methylation age at the time of disease onset. I conclude that compared to real age, methylation age seems to provide a better way to assess one's risk of developing an aging related disease. This may hold the key to better understanding the mechanisms involved in the onset of aging related disease, which could potentially help us create better risk assessment, diagnosis and prevention tools in the future.

8.1 Relationship between Age of Disease Onset and Age Acceleration

In this section, I provide convincing evidence that shows that disease generally seems to start sooner for those with larger age accelerations. I computed the correlation between the age of disease onset and negative age acceleration using R's cor.test function. Figures 8.1.1 and 8.1.2 show the relationship between negative age accel-

erations and age. Condensed plots showing the correlation values associated with Random Forest and Linear Regression across the different age related conditions can be found in Figures B.0.2 and B.0.1 in the Appendix. The p-values were adjusted for the number of tests performed by dividing it by the number of covariates tested in this set. The results were compared to the correlations between permuted onset ages and negative age accelerations. The correlations of all the permuted ages were insignificant, since their p-values were too high in prediction algorithms. Conversely, the correlation results for 7 out of 8 age related conditions were significant in both algorithms. Thyroid disease was the only condition whose onset did not seem to be related to methylation age, and it seemed not to be correlated only in one of the two algorithms.

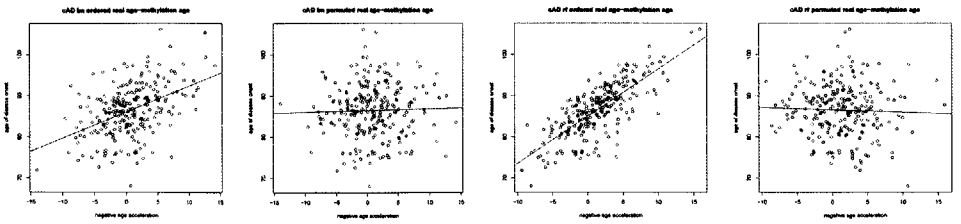


(a) lm - thyroid ordered values (b) lm - thyroid permuted (c) rf - thyroid ordered values (d) rf - thyroid permuted

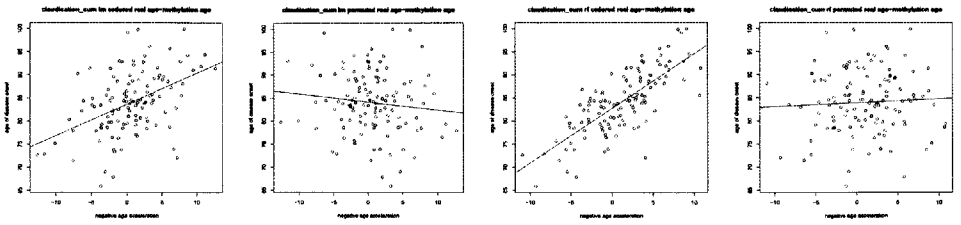
(e) lm- cancer ordered (f) lm- cancer permuted (g) rf- cancer ordered (h) rf- cancer permuted

(i) lm- stroke ordered (j) lm- stroke permuted (k) rf- stroke ordered (l) rf- stroke permuted

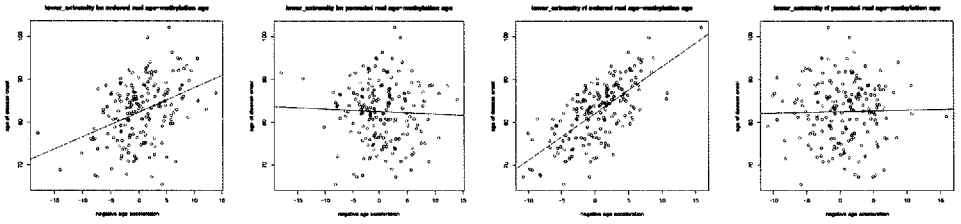
Figure 8.1.1: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms



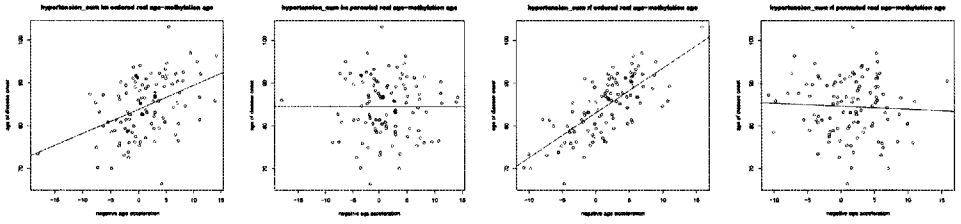
(a) lm- AD ordered (b) lm- AD permuted (c) rf- AD ordered (d) rf- AD permuted



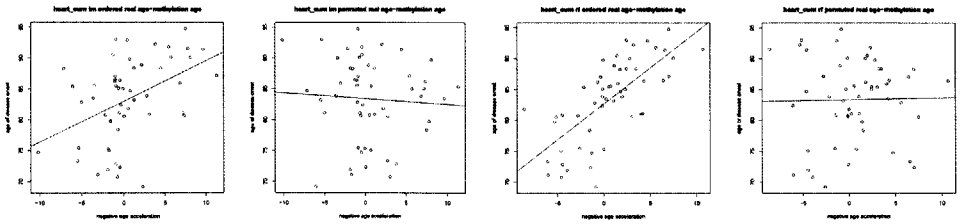
(e) lm- claudication ordered (f) lm- claudication permuted (g) rf- claudication ordered (h) rf- claudication permuted



(i) lm- l. extremity ordered (j) lm- l. extremity permuted (k) rf- l. extremity ordered (l) rf- l. extremity permuted



(m) lm- hypertension ordered (n) lm- hypertension permuted (o) rf- hypertension ordered (p) rf- hypertension permuted



(q) lm- heart ordered (r) lm- heart permuted (s) rf- heart ordered (t) rf- heart permuted

Figure 8.1.2: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms

Figures 8.1.1 and 8.1.2 show the relationship between negative age acceleration (real age - methylation age) and the age of onset of different aging diseases.
Here we see relatively clearly that the larger (more negative) the age acceleration is, the older the age of disease onset. A larger (more positive) value in the plot indicates a slower (more negative) age acceleration.

8.1.1 Correlation Differences across the Different Diseases

Although there seems to be a significant relationship between the age of disease onset and disease acceleration in the majority of the diseases presented above, this relationship and its level of significance vary across the different diseases. Here I examine the possible causes of this variation.

The age of onset of AD seems to be highly correlated with age acceleration in both algorithms. Claudication (leg pain while walking), lower extremity pain and hypertension also seem to be highly correlated, but less correlated than AD when using linear regression to predict age. Heart disease and cancer are still correlated, but their correlation values are significantly lower, particularly when using linear regression. Thyroid disease appears to only be significantly correlated when using Random Forest.

The results make sense given that AD is known to be an age related disease that is associated with the prefrontal cortex (Kashani et al. found impairments in the vesicular glutamate systems of the prefrontal dorsolateral cortex of AD patients and these impairments are associated with cognitive decline).

Looking at methylation age in other tissues may result in higher correlations in the other conditions. For example, heart disease may be associated with methylation taken from heart cells. Thyroid disease may be associated with thyroid tissues or immune cell methylation, since thyroid problems are often caused by an autoimmune condition e.g. Graves disease, based on MedlinePlus [26]. Another possible explanation for the high thyroid p-value is that certain types of thyroid disease appear at a relatively young age. For example, Graves disease is known to mainly affect people between age 20 and 40 [8]. In fact, I know several people who've a thyroid condition

this at an early age, e.g. 20's. However, hypothyroidism is often associated with more mature populations (e.g women over 60, based on WebMD [45]).

Based on Cunha [10], hypertension is associated with the stiffening of the arterioles. Perhaps looking at the methylation age in the arterioles would be more useful than prefrontal cortex methylation age in evaluating the effects of methylation age on blood pressure.

There are notable differences in correlation across the two algorithms. The correlation p-values associated with random forest seem to be significantly higher (more significant) than related to linear regression.

8.1.2 Plot Similarities across the Different Diseases

Some of the plots look relatively similar across different diseases. Here I examine these similarities.

In Figure 8.1.2, the plots for AD and claudication and lower extremety look very similar. The Venn diagrams in Figure 9.1.4b show that AD and claudication share a relatively large number of significant probes, which may be the reason why their plots look so similar.

Random Forest seems to produce better results than Linear Regression does in this case. The relationship between the methylation ages estimated by rf and the age of disease onset seems to be more direct than that produced by lm. This difference between lm and rf is reflected in 8.1.1 above, in Figure B.0.3 (and in Figures B.0.1 and B.0.2 in the Appendix).

8.2 Disease Onset Risk Assessment - Real Age vs. Approximated Methylation Age

In this section, I show that the standard deviation of the real ages of disease onset is larger than that of the methylation ages of disease onset, meaning that the methylation age is a better predictor of disease than real age. I attempted to estimate the

patients' methylation age of onset by subtracting the differentce between their age of death and their age of onset from their methylation age of death. I was unable to compute their exact methylation age of onset since I only have methylation data from the time of death of each patient. In most cases the standard deviation of the estimated methylation ages of onset is smaller than the standard deviation of the real ages of onset.

Figure B.0.3 in the Appendix compares the mean and standard deviation of the estimated methylation onset ages and the real disease onset ages. These values are compared across 8 different aging related diseases and 2 prediction algorithms, random forest (on the right) and linear regression (on the left). Similarly to the correlation results (shown in Figure B.0.1), the random forest yields larger differences than linear regression. These results suggest that one's methylation age may provide a more accurate way to assess one's risk of developing an aging related disease than real age. These results are consistent with those presented in subsection 8.1 above.

Chapter 9

SNPs, DNA Methylation and Aging Disease

So far I have looked at methylation changes associated with aging and age related disease. In this chapter I look at the genotypes responsible for these changes and examine the relationship between genotype, methylation, aging phenotypes, such as disease and age acceleration.

To identify genetic variations I use at SNPs. Single nucleotide polymorphism (SNP) is a genetic variation in a single DNA building block. Some of these genetic variations or SNPs are associated with disease [31].

9.1 Disease cis-meQTLs

In order to better understand the relationship between methylation, genetics and age related disease, I computed the cis-meQTLs for eight different aging related conditions(lower extremety pain, claudication, thyroid disease, alzheimer's (AD), hypertension, heart disease, stroke and cancer) and the age accelerations associated with the ages predicted by Random Forest and Linear Regression. Methylation quantitive trait loci (meQTLs or mQTLs) are variations in methylation that are associated with genetic variations or SNPs [4, 5, 20]. MeQTLs that are cis are methylation-genotype associations where the CpGs that are located near the associated SNPs [4]. In this

section I identify such methylation-genotype associations that are also associated with aging related conditions, where variations in methylation are associated with speicific SNPs and aging related conditions..

9.1.1 SNP file format

For each SNP I have the chromosome (chr), id(rid), the position(pos), the minor allele(A1), the major allele(A2) and the number of minor alleles (A2) each patient has (0, 1 or 2).

9.1.2 Disease cis-meQTL Computation

For each SNP_i in the set of SNPs, the closest CpGs (within distance 1000000) are found and then Likelihood Ratio Testing is done (using a script written by Matthew Eaton) to compare two linear models, where one model associates SNPs with changes in DNA methylation and the other model associates a combination of SNPs and disease with changes in DNA methylation. This computation is further described below and is also very similar to the formula presented in the first cis-meQTLs slide of the 6.047 Lecture 22 (Personalized Genomics) notes [12].

The goal here is to determine whether SNP and disease explain changes in methylation better than SNP_i alone. I do this by using F distribution to compare a model that estimates methylation using SNP_i to a model that predicts methylation using SNP_i and a specific aging related disease, <aging disease>, where both models correct for gender. I assume that the two models are the same as the null hypothesis. In order to determine whether our assumption is correct I compute the p-value. If it is low enough, it contradicts our assumption that the two models are the same, indicating that methylation is better estimated by both SNP_i and <aging disease>. The pseudo-code for this computation is shown in Figure 9.1.1 below.

LM1: methylayion $\sim SNP_i$ + gender
LM2: methylayion $\sim SNP_i$*<aging disease> + gender
q: # parameters in LM1
p: # parameters in LM2
n: # samples (patients I think)
This is computed as follows:
f:$(RSS_{LM1} - RSS_{LM2}/(q-p))/(RSS_{LM2}/(n-q))$ distributed as F(q-p,n-q)
R's pf function is used to compute its p-values
This computation is done for each aging disease seperately.

Figure 9.1.1: Computing meQTLs where Methytaion is Associated with SNP and Disease

9.1.3 Issues

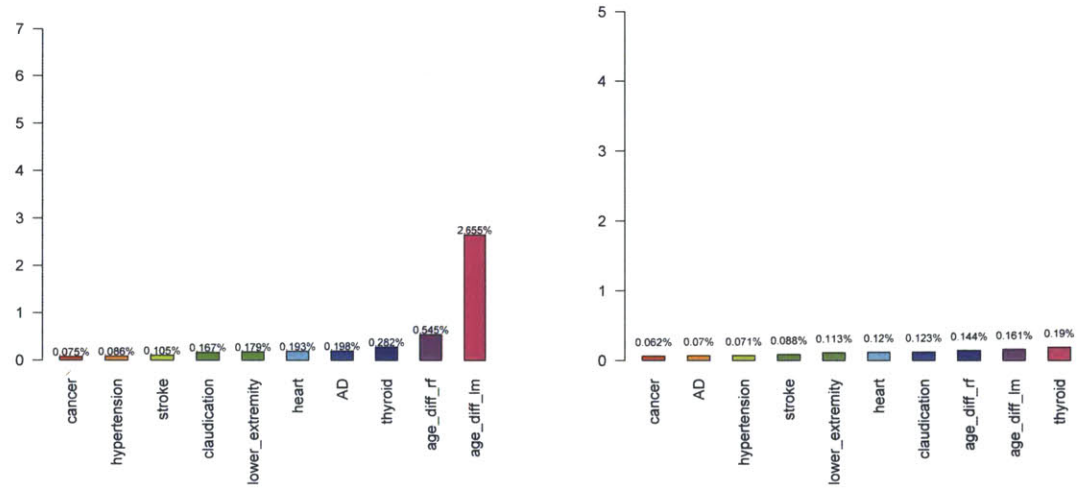
Chromosome Y is not accounted for because I'm not sure which number in the table is chromosome Y. The SNP table has all the chromosomes named 0 through 26. In contrast, probes.df has chromosomes 1 through 22, x and y. X and Y could map to 0, 23 (probably not 23), 25 or 26.

9.1.4 Significant Aging and Disease Probes

Determining the epigenetic components that are associated with aging, genetics and aging related disease may help us better understand the causes of aging and aging related disease. In order to identify such methylation probes, I intersected the meQTL related methylation probes I described above with the highly significant age related probes I computed in Chapters 4, 5 and 7. The existence of such probes would imply that the some of the age related methylation changes are also associated with genetically related aging diseases. For each disease, I intersected the meQTL associated probes with the highly significant aging probes that were shared across 1 fold or more in the cross validation of each age prediction algorithm. I used two age prediction algorithms Random Forest and Linear regression. There seems to be no overlap between the previously computed age related methylation probes and the methylation probes that are associated with the meQTLs above.

9.1.5 Fraction of meQTL SNPs and related Methylation Probes Associated with Disease

In this subsection I examine the meQTL SNPs that are associated with different aging related conditions. I look at the fraction of meQTL SNPs associated with each disease and their locations in the human genome. The results indicate that a large portion of SNPs is associated with age acceleration related methylation changes, especially when using Linear Regression to predict age from methylation. This implies that age acceleration has strong genetic influences that are associated with methylation.



(a) Fraction of meQTL SNPs Associated with Disease

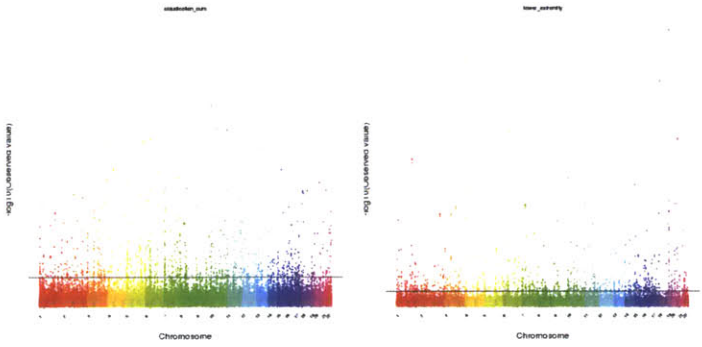
(b) Fraction of meQTL Methylation Probes Associated with Disease

Figure 9.1.2: Fraction of meQTL SNPs and meQTL Methylation Probes Associated with Disease

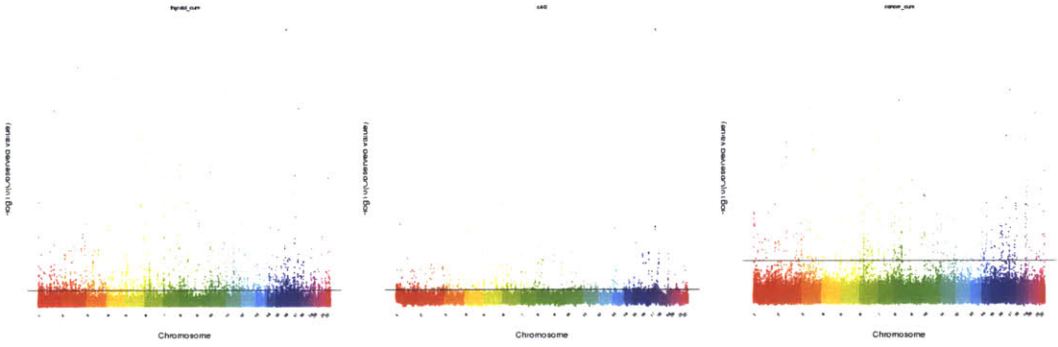
Figure 9.1.2a shows the percentage of cis-meQTLs SNPs associated with disease. Figure 9.1.2b shows the fraction of cis-meQTLs related methylation probes that are associated with Disease. Figure 9.1.3 shows a Manhattan plot for meQTL SNPs associated with each disease. The x-axis indicates is the location of the SNP and the y-axis shows its p-value. The black line in each plot shows the significance threshold.

The investigated diseases seem to have a relatively small fraction of the meQTL SNPs associated with them. However, there seems to be a larger portion of meQTL SNPs associated with age acceleration. There are at least twice as many meQTL SNPs associated with negative age acceleration as there are meQTL SNPs associated

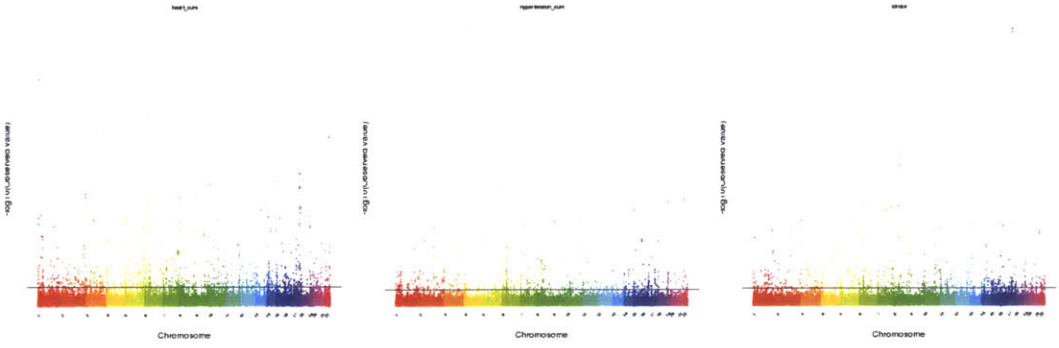
with each of the aging related diseases shown in the plot above. There is also a very large difference between the fraction of SNPs associated with negative age acceleration across the different age prediction algorithms (rf and lm). Significantly more SNPs seem to be associated with age acceleration when the methylation age is predicted using LM compared to when using RF. All investigated diseases and the age accelerations generated by both prediction algorithms seem to have a small portion of meQTL related methylation probes.



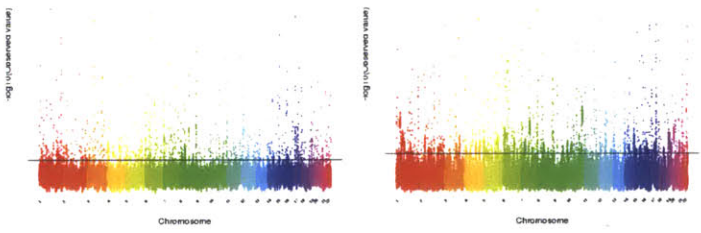
(a) Claudication (b) Lower Extremity Pain



(c) Thyroid Disease (d) AD (e) Cancer



(f) Heart Disease (g) Hypertension (h) Stroke

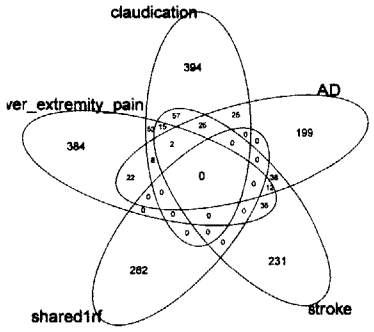


(i) Age Acceleration - RF (j) Age Acceleration - LM

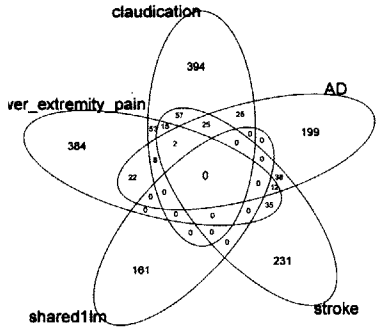
Figure 9.1.3: Manhattan Plots for the Disease Related meQTLs

9.1.6 Venn Diagrams of Aging and Disease Methylation Probes

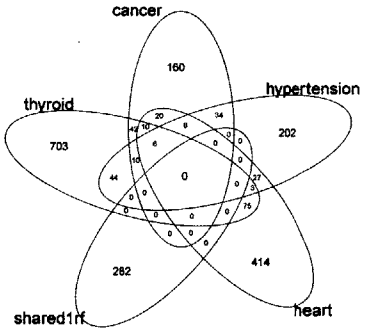
In this subsection I look at the highly predictive age probes that are shared across different aging related conditions. Since aging related conditions are more common in older individuals, it is worth investigating whether these conditions share probes with other conditions and age.



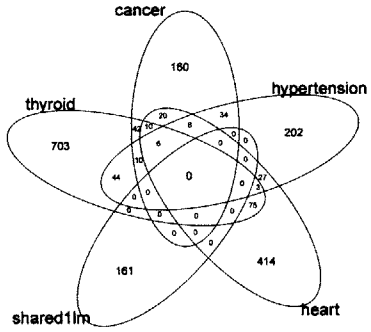
(a) Venn diagrams of aging probes (sharedrf) that are shared across 1 fold or more (of rf) and disease probes (the other nodes)



(b) Venn diagrams of aging probes (sharedlm) that are shared across 1 fold or more (of lm) and disease probes (the other nodes)



(c) Venn diagrams of aging probes (sharedrf) that are shared across 1 fold or more (of rf) and disease probes (the other nodes)



(d) Venn diagrams of aging probes (sharedlm) that are shared across 1 fold or more (of lm) and disease probes (the other nodes)

Figure 9.1.4: Venn Diagrams of shared meQTL Probes across Different Aging Related Diseases

The Venn diagrams above show how many aging methylation probes (sharedlm or sharedrf) that are shared across x folds or more are shared with disease specific methylation probes (the other nodes).

It seems like none of the highly significant aging methylation probes are associated with any of the aging related diseases listed above (no probes are shared between aging probes and disease probes). However, some of the aging related diseases seem to share a large number of methylation probes. For example, thyroid disease and heart disease share 94 meQTL related methylation probes. AD and Claudication share 60 probes. Claudication, AD and stroke share 27 methylation probes. Claudication and lower extremity pain have 78 probes in common. 44 probes are shared across AD and lower extremity.

9.1.6.1 AD and Claudication

Given that AD and claudication share a large number of methylation probes and due to the fact that claudication may prevents patients from walking, I checked the correlation between claudication and AD. A significant correlation value between the two would imply that there may be a link between lack of exercise and the onset of AD. Based on the results, there is no significant correlation between AD and claudication since the p-value is 0.2417 ($> .05$) and the correlation value is 0.04720556.

Chapter 10

Age Prediction from Methylation in Blood

In this chapter age is predicted in blood methylation. The predictions are done in two different datasets where one measures methylation in whole blood and the other measures methylation in CD4+ cells. I also determine the age predictability of significant brain age probes in whole blood and CD4+ blood cells. The results indicate that age is highly predictable in whole blood and somewhat predictable in CD4+ cells. However, predictions associated with CD4+ cells were difficult due to the small number of available samples, meaning that CD4+ cells may be strongly associated with age. Based on the results, significant brain based age probes are somewhat predictive of age in whole blood and CD4+ cells. However, it is difficult to estimate how well brain based age probes predict age in CD4+ cells due to the small number of CD4+ samples provided.

10.1 Age Prediction in CD4+ Blood Cells

Our immune system's ability to protect against infectious disease decreases as we age. This seems to be linked to immune system changes related to age (Kovaiou and Grubeck- Loebenstein, and Hale et. al). In this section I check whether there is a relationship specific to CD4+ cells between methylation, aging and age related

disease. According to Kovaiou and Grubeck-Loebenstein, understanding aged CD4+ has important clinical applications that can potentially promote healthy aging and improved quality of life.

In this section I examine the relationship between methylation and age in blood CD4+ cells. The CD4+ data were obtained from the brain methylation study from which the brain methylation data was obtained. I have a very small set of blood samples obtained at the time at which each patient joined the study (baseline samples) and at the time of death.

10.1.1 General Age Prediction in CD4+ Blood Cells

Ages were predicted in 42 baseline samples alone, 44 post death samples alone and in 86 samples consisting of both sets. Although the vast majority of the samples in each set were obtained from the same patients these samples were treated independently when the two sets were combined as if they were collected from different patients. The idea behind combining the two sets was to double the number of available samples, assuming that the methylation profiles of patients that had both a baseline sample and a post death sample were different enough to be considered as samples obtained from two different patients. Prediction accuracies were evaluated using 3 or 4 fold cross validation since the sample size was too small for 5 fold cross validation.

The ages were predicted for each of the sets in two different ways. One way, the brain based implementation, uses significant brain methylation age probes as significant blood probes and R's Random Forest Cross Validation function (rfcv). The other implementation, the blood based implementation, uses the unbiased rfcv I presented in Chapter 4, which selects significant probes in each fold out of all $\tilde{4}8{,}000$ available probes and does not rely on significant probes obtained from the brain.

The results imply that there seems to be a relationship between CD4+ methylation and age, which suggests that age related immune system deterioration may be associated with changes in CD4+ methylation. However, the small number of samples and the narrow range of ages make predictions difficult and in some cases nearly random. Larger datasets with a wider range of ages would be required in order to

properly test the relationship between CD4+ methylation.

10.1.1.1 Brain Based Implementation

The results suggest that brain based age methylation probes are also indicative of age in CD4+ cells. However, generally, predictions were unstable and not very accurate compared to random predictions. The brain based predictions in CD4+ methylation seemed to largely depend on the difference in age ranges of the participants. The results were evaluated based on how the errors of the real predictions compared to the errors of a predictor with random or permuted ages and based p-values of the linear regression of the chronological ages and the methylation ages.

The death age based sample predictor performed so poorly, that the predictions resembled those of a random predictor. The difference between the oldest patient and the youngest patient in this dataset was about 20 years and the difference between the oldest methylation age and the youngest methylation age was about 6-7 years. It seems like this predictor failed because of the narrow age range of the experiments in the dataset and perhaps also due to the small number of samples available.

The baseline age predictions were slightly better, but very inconsistent. Most of the time the predictions of the real predictor seemed to perform better than the random predictor and create a significant relationship between the real ages and the predicted ages. The samples in this data had a wider age range than the death age dataset. The difference between the oldest participant and the youngest participant was approximately 30 years and the age difference between the oldest prediction and the youngest prediction was about 10 years. This predictor was not very successful probably due to the relatively narrow age range and the small number of samples in the dataset.

The combined age predictor performed better than the other two, consistently predicting ages that had a significant relationship with the chronological ages and with a higher level of accuracy than the random predictor. The prediction error of this age predictor was about 6.28 years. The predictions associated with this dataset were better because it had more samples and its participants had a wider range of

ages. The dataset associated with this predictor had twice as many samples as the other two datasets and a wider age range. The age difference between the youngest and oldest participants was about 34 years and the difference between the youngest and oldest predicted ages was about 12 years.

10.1.1.2 Blood Based Implementation

The blood based age predictors performed very poorly, predicting ages in a way that resembled random prediction and within a very narrow range of ages. The poor performance of this predictor was probably largely associated with the small number of samples that was used to train the data.

10.1.2 Individual Age Prediction in CD4+ Blood Cells

The biological age progression and the related symptoms or conditions seems to vary across different individuals. This implies that age prediction accuracy may vary depending on the tested individual when the predictor is trained on a random group of patients. In this section I try to determine the accuracy of an age predictor that was trained on one's own blood (or a dataset that includes one's own blood methylation) methylation at a younger age. Possible applications include individualized biological age assessment that is more tailored to one's body. This would potentially enable individuals to monitor the changes in their own aging rate based on environmental factors, work on ways to slow it down (e.g. by eating healthier, reducing stress and exercising) and monitor their progress. This type of individual monitoring may also potentially enable researchers to determine the effects of environmental factors on biological aging and the rate at which we age.

I used patient blood methylation data from CD4+ blood cells to predict age in 44 patients, training the predictor on methylation data collected when each patient joined the study and testing it on methylation data from the time the patient passed away. My goal was to check the accuracy of age prediction when the predictor trained on methylation data from specific individuals and then tested on the same individuals

when they are older. The results indicated a clear relationship between their real age and the predicted age, however the error rates were relatively high and the age range of the predicted ages was very narrow. One possible cause is the small number of samples used to train the predictor (approximately 42 samples). I had blood methylation data for a very small set of patients. Another possible reason is that the predictor was trained on a younger set of patients (roughly the same patients when they were younger), meaning that it was trained to predict younger ages. An additional possible cause for prediction inacuracy is that the predictions here are based on methylation in CD4+ blood methylation, which may be less affected by age than brain methylation and thus cause less accurate predictions. Another possible cause of error is that the individualized predictor was trained on methylation data taken from a set of patients. Perhaps training each individualized age predictor for each patient, on her own methylation data alone (instead of training it on methylation data from a set of patients that includes this patient) result in better prediction accuracy for that specific patient at a later age.

10.2 Prediction in Whole Blood using Data from Existing Papers

In this section I use Random Forest Cross Validation to predict age in whole blood methylation [1] and assess the prediction accuracy. The data was generated by Hannum et al. [19], where the authors used a different algorithm (Elastic Net) to predict age from methylation in whole blood. First I predict age from whole blood methylation and produce results that show a relationship between predicted age and real age, but are are less accurate than those in the original paper. Next, I use methylation probes that were found to be highly predictive of age in brain methylation to predict age from blood methylation. First, I try to predict age in patients between 19 and 101 and then I predict age in patients over 65 because the brain data in which I computed the significant age probes was collected from patients who were 65 or older.

10.2.1 Five Fold Cross Validation on All Patients

I used (unbiased) Random Forest Cross Validation to predict age from methylation in whole blood that was collected from 656 people aged 19-101. I predicted age in whole blood methylation in the same way as I predicted age in brain methylation, but without adjusting for plates. The prediction results were relatively accurate, 7.32 years, which implies that methylation changes with age at any age, even for people as young as 19. These results are similar to (but less accurate than) those obtained by the authors of [19] (where the data came from). The lowest prediction rate, 7.32, was obtained with 50 features, but a comparable prediction rate of 7.3682820250042 was obtained with 100 features. The results with the lowest prediction error are displayed in Figure 10.2.1 where the x-axis shows the predicted ages and the y-axis shows the real ages.

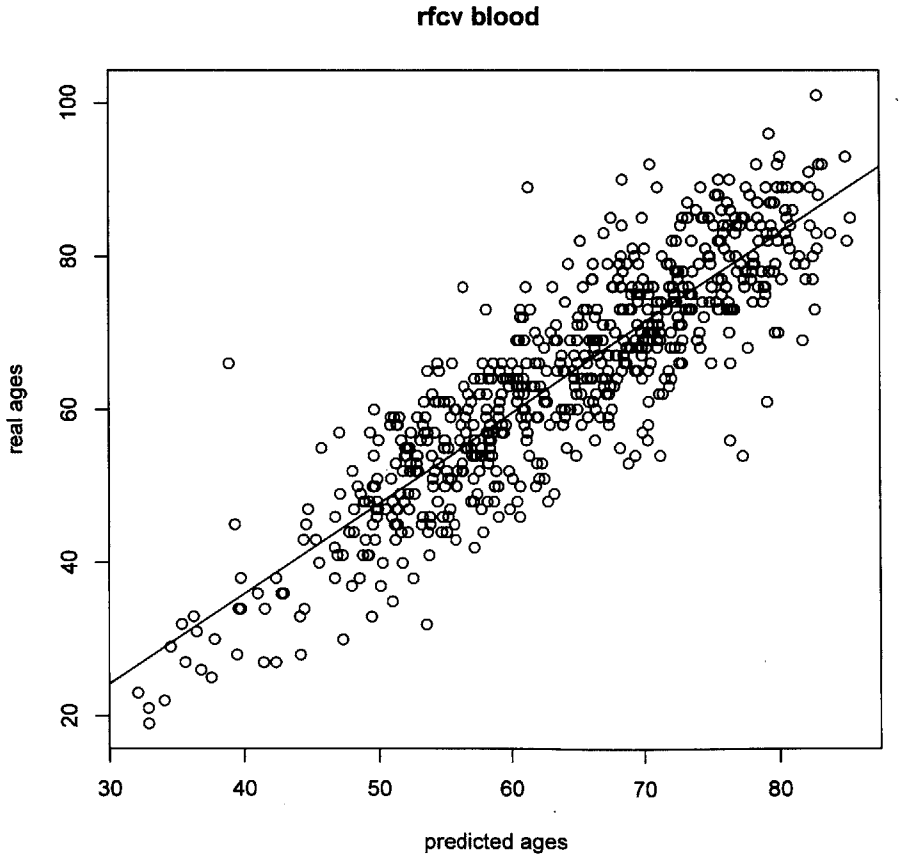


Figure 10.2.1: Age Prediction based on Blood Methylation using Random Forest

The prediction error presented here is worse but still comparable to that obtained from brain methylation. One possible reason for the difference in prediction accuracy between blood methylation and brain methylation is the large difference in patient age range between experiments associated with these tissues. Unlike the brain related experiment data that came from a relatively narrow range of ages (65-108) the blood methylation experiment data was obtained from patients of a wider age range (19-101). A wider age range is more likely to result in larger errors, since predictions can be anywhere between 19 and 101. The errors seem especially large along the edges. Since the number of experiments along the edges is so small, the corresponding ages do not seem to be represented in the predictions and the resulting prediction errors along the edges are large.

An attempt was made to adjust for environmental covariates, such as race, to reduce the prediction error. However, adjusting for different covariates, seemed to make predictions worse. Surprisingly, even adjusting for plates increased the prediction error rate.

The most significant age probes that were previously found by Random Forest in brain methylation were compared to most significant age probes found in blood. These probes were selected by intersecting the 100 most significant probes in each cross validation fold and picking the probes that were highly significant in 2 or more folds. There seemed to be no overlap in highly significant probes across the two tissues. An attempt was made to predict age in blood based on highly significant age methylation probes that were identified in the brain by the Random Forest predictor (results that were consistent across 2 cross validation folds or more). This was done using R's rfcv function, since there was no bias problem in this case. The prediction results were slightly better than those obtained by a random predictor but not nearly as accurate as those obtained by the most highly predictive probes in blood methylation that are presented above. These results contradict the results presented by Horvath that suggested that age can be accuratly predicted by the same CpGs across most tissues.

Chapter 11

Conclusion and Future Directions

11.1 Conclusion

In this thesis, I predicted age from methylation in the prefrontal cortex of the brain using two different algorithms, Random Forest and Cross Validation. I then identified a set of methylation probes that were highly predictive of age and investigated their chromatin states and biological functions. The findings suggested that age probes may be involved in age related conditions such as memory and memory and learning problems (e.g. AD or Dimensia), high blood pressure, heart disease. I also found that the age accelerations that were computed by random forest were associated with the age of onset age of age related disease, suggesting that high age accelerations are associated with earlier onset of disease.

I also predicted age in whole blood methylation and showed that age can be accuratly predicted in whole blood. Surprisingly, probes that were found highly predictive in the brain were only somewhat predictive of age in whole blood and there was no overlap between probes that were highly predictive in brain methylation and probes that were highly predictive of age predictive in blood methylation.

An attempt was made to predict age in a small set of CD4+ methylation samples. The results indicated that CD4+ methylation is predictive of age and that highly predictive brain methylation probes are also predictive in CD4+ cells. These results imply that age related immune system changes, such as immune system weakening,

may be associated with age related methylation changes in CD4+ cells. However, the level of significance and age prediction accuracy associated with the relationship between CD4+ cell methylation and age was difficult to estimate, due to the small number of available samples.

The findings presented in this thesis take us a step further in understanding the causes of aging and age related disease and further prove the theory that methylation is related to our internal biological clock and disease which has previously been suggested by Horvath and Hannum et al..

11.2 Future Directions

I would like to continue investigating the relationship between age methylation and the age of disease onset. I want to check if this relationship is as strong in blood as it is in the brain, since disease risk assessment based on blood methylation would be significantly more practical than risk assessment that requires brain tissues. I also wish to continue investigating the effects of environmental covariates on age prediction and aging. This can be done by exploring additional covariates, adjusting for their effects and trying to improve the prediction error rates. I also want to look at genes that are associated with meQTL SNPs that are related to age accelerations and research their previously identified functions and how they relate to the findings presented in this thesis. In addition, I would like to explore the relationship between CD4+ cell methylation and age in a larger set of experiments and a wider range of ages, and use the results to find the biological functions associated with these changes and how they are associated with age related immune system changes and weakening.

Appendix A

Tables

A.1 GO Categories of Highly Predictive Probes - across all 2 folds or more

A.1.1 BF - Biological Process

A.1.1.1 E1 - PC Repressed

Looking at the repressed processing, it seems that these terms mostly include tissue development and reulation, where develepment is often known to slow down as we age. These results may help explain why. These results are consistent across multiple algorithms (Random Forest, Linear Regression and the intersection of the resulting probes).

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:2000594	6.557377e-03	Inf	1	1	positive regulation of metanephric DCT cell differentiation
GO:0072289	6.557377e-03	Inf	1	1	metanephric nephron tubule formation
GO:0021633	6.557377e-03	Inf	1	1	optic nerve structural organization
GO:0042488	6.557377e-03	Inf	1	1	positive regulation of odontogenesis of dentine-containing tooth
GO:0021650	6.557377e-03	Inf	1	1	vestibulocochlear nerve formation
GO:0050770	7.083967e-03	19.472222	2	20	regulation of axonogenesis
GO:0048522	7.242644e-03	4.334001	8	507	positive regulation of cellular process
GO:0003002	7.288509e-03	6.554098	4	126	regionalization
GO:0048589	7.385615e-03	9.210134	3	64	developmental growth
GO:0048675	7.811361e-03	18.438596	2	21	axon extension
GO:0048936	6.557377e-03	Inf	1	1	peripheral nervous system neuron axonogenesis
GO:0072069	6.557377e-03	Inf	1	1	DCT cell differentiation
GO:0072173	6.557377e-03	Inf	1	1	metanephric tubule morphogenesis
GO:0009058	1.339917e-03	6.094003	10	627	biosynthetic process
GO:0006355	1.505255e-03	8.431894	6	307	regulation of transcription, DNA-dependent
GO:0034641	1.620504e-03	5.903328	10	641	cellular nitrogen compound metabolic process
GO:0050767	2.226913e-03	9.351724	4	91	regulation of neurogenesis
GO:0048568	6.100634e-04	9.965278	5	117	embryonic organ development
GO:0007420	1.142362e-03	8.578811	5	134	brain development
GO:0045595	1.224383e-03	7.125000	6	208	regulation of cell differentiation
GO:0010720	1.298984e-03	17.803977	3	35	positive regulation of cell development
GO:0035799	6.557377e-03	Inf	1	1	ureter maturation
GO:0039003	6.557377e-03	Inf	1	1	pronephric field specification
GO:0031293	6.557377e-03	Inf	1	1	membrane protein intracellular domain proteolysis
GO:0035566	6.557377e-03	Inf	1	1	regulation of metanephros size
GO:0042471	2.374226e-03	14.188636	3	43	ear morphogenesis
GO:2000026	2.785198e-03	5.962025	6	243	regulation of multicellular organismal development
GO:0045665	6.408555e-03	20.627451	2	19	negative regulation of neuron differentiation
GO:0021524	6.557377e-03	Inf	1	1	visceral motor neuron differentiation
GO:0010467	1.950287e-05	12.204762	11	501	gene expression
GO:0032774	6.154983e-05	9.717742	10	444	RNA biosynthetic process
GO:0001657	4.801744e-05	27.880000	4	34	ureteric bud development
GO:0048812	1.037049e-04	11.775591	6	133	neuron projection morphogenesis
GO:0010556	9.767622e-05	9.102845	10	467	regulation of macromolecule biosynthetic process
GO:0034645	1.326662e-04	8.710359	10	483	cellular macromolecule biosynthetic process
GO:0048667	1.082007e-04	11.677734	6	134	cell morphogenesis involved in neuron differentiation
GO:0090304	1.456389e-04	8.593096	10	488	nucleic acid metabolic process
GO:0032990	1.385849e-04	11.121269	6	140	cell part morphogenesis
GO:0061360	6.557377e-03	Inf	1	1	optic chiasma development
GO:0048856	6.156160e-03	5.267640	10	832	anatomical structure development
GO:0045892	8.596770e-03	6.228125	4	132	negative regulation of transcription, DNA-dependent
GO:0051093	6.121942e-03	6.913793	4	120	negative regulation of developmental process
GO:0001763	7.066929e-03	9.368182	3	63	morphogenesis of a branching structure
GO:0030182	6.775001e-03	5.791589	5	236	neuron differentiation
GO:0021545	1.730144e-04	38.290909	3	18	cranial nerve development
GO:0007399	1.712950e-04	8.246842	9	389	nervous system development
GO:0019219	2.325389e-04	8.020833	10	514	regulation of nucleobase-containing compound metabolic process
GO:0031326	2.092549e-04	8.147590	10	508	regulation of cellular biosynthetic process
GO:0001658	4.192566e-04	27.272727	3	24	branching involved in ureteric bud morphogenesis
GO:0001709	2.787297e-04	31.863636	3	21	cell fate determination
GO:0019222	4.703536e-04	7.837771	11	687	regulation of metabolic process
GO:0030030	4.649138e-04	8.718750	6	174	cell projection organization
GO:0048880	5.902607e-04	88.208333	2	6	sensory system development
GO:0009887	5.489817e-04	7.484000	7	257	organ morphogenesis
GO:0048598	2.926501e-03	6.809028	5	165	embryonic morphogenesis
GO:0007411	2.821245e-03	8.722581	4	97	axon guidance
GO:0051216	5.069539e-03	10.641509	3	56	cartilage development
GO:0071841	4.475035e-03	4.957865	7	363	cellular component organization or biogenesis at cellular level
GO:0022604	3.669331e-03	12.034816	3	50	regulation of cell morphogenesis
GO:0001755	3.474967e-03	29.291667	2	14	neural crest cell migration
GO:0044281	3.463812e-03	5.173661	10	701	small molecule metabolic process
GO:0072305	6.557377e-03	Inf	1	1	negative regulation of mesenchymal stem cell apoptosis involved in metanephric nephron morphogenesis
GO:0048762	3.071308e-03	12.873967	3	47	mesenchymal cell differentiation
GO:2000597	6.557377e-03	Inf	1	1	positive regulation of optic nerve formation
GO:0051799	6.557377e-03	Inf	1	1	negative regulation of hair follicle development

Table A.1: E1 - PC Repressed - rf

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0035850	2.989355e-03	33.428571	2	8	epithelial cell differentiation involved in kidney development
GO:0048729	2.641136e-03	5.294118	6	138	tissue morphogenesis
GO:0051093	1.275498e-03	6.185759	6	120	negative regulation of developmental process
GO:0043009	1.648355e-03	5.858824	6	126	chordate embryonic development
GO:0048762	1.344102e-03	10.129743	4	47	mesenchymal cell differentiation
GO:0043583	3.347030e-03	7.729323	4	60	ear development
GO:0060231	4.741656e-03	25.047619	2	10	mesenchymal to epithelial transition
GO:0072273	4.741656e-03	25.047619	2	10	metanephric nephron morphogenesis
GO:0048699	4.759035e-03	3.832558	8	266	generation of neurons
GO:0050794	5.231970e-03	6.693198	10	701	regulation of cellular process
GO:0030154	1.804941e-03	4.496522	11	545	cell differentiation
GO:0042472	5.711413e-03	9.750000	3	35	inner ear morphogenesis
GO:0032835	6.863811e-03	20.019048	2	12	glomerulus development
GO:0060429	6.879253e-03	4.276946	6	167	epithelium development
GO:0001654	7.141583e-03	6.141353	4	74	eye development
GO:0010770	2.256793e-03	40.133333	2	7	positive regulation of cell morphogenesis involved in differentiation
GO:0060173	2.266308e-03	8.682105	4	54	limb development
GO:0009890	1.236530e-04	7.129252	8	155	negative regulation of biosynthetic process
GO:0048513	1.985188e-04	5.361257	14	587	organ development
GO:0045944	2.014368e-04	6.595781	8	166	positive regulation of transcription from RNA polymerase II promoter
GO:0031324	2.382902e-04	6.419753	8	170	negative regulation of cellular metabolic process
GO:0009058	6.724635e-05	8.971800	11	472	biosynthetic process
GO:2000113	7.274048e-05	7.749020	8	144	negative regulation of cellular macromolecule biosynthetic process
GO:0048562	7.642385e-05	11.112329	6	79	embryonic organ morphogenesis
GO:0001658	9.444331e-05	22.021053	4	24	branching involved in ureteric bud morphogenesis
GO:0048880	1.622625e-03	50.190476	2	6	sensory system development
GO:0072207	1.622625e-03	50.190476	2	6	metanephric epithelium development
GO:0035136	1.472660e-03	16.523684	3	22	forelimb morphogenesis
GO:0009952	1.608817e-03	7.148383	5	84	anterior/posterior pattern specification
GO:0051254	2.416009e-04	5.822449	9	219	positive regulation of RNA metabolic process
GO:0072298	3.309936e-04	201.047619	2	3	regulation of metanephric glomerulus development
GO:0060284	1.017843e-03	6.485699	6	115	regulation of cell development
GO:0072189	1.088887e-03	66.952381	2	5	ureter development
GO:0032774	3.065188e-10	18.854706	19	444	RNA biosynthetic process
GO:0019219	4.458309e-09	15.516667	19	514	regulation of nucleobase-containing compound metabolic process
GO:0090304	1.732943e-09	16.640192	19	488	nucleic acid metabolic process
GO:0034641	2.336240e-07	11.378617	19	641	cellular nitrogen compound metabolic process
GO:0006355	1.335007e-08	17.336066	15	381	regulation of transcription, DNA-dependent
GO:0019222	7.906312e-07	10.267964	19	687	regulation of metabolic process
GO:0010556	3.807535e-07	18.823529	10	231	regulation of macromolecule biosynthetic process
GO:0044281	1.124686e-06	9.359677	19	701	small molecule metabolic process
GO:0031326	1.113711e-06	16.493902	10	256	regulation of cellular biosynthetic process
GO:0007422	1.908415e-03	14.935714	3	24	peripheral nervous system development
GO:0009653	1.013240e-03	6.007326	8	320	anatomical structure morphogenesis
GO:0001655	5.562080e-03	6.629960	4	69	urogenital system development
GO:0048522	9.653621e-03	2.086559	11	507	positive regulation of cellular process
GO:0031076	9.340545e-03	16.666667	2	14	embryonic camera-type eye development
GO:0048705	9.018032e-04	8.224638	5	74	skeletal system morphogenesis
GO:0048333	8.058793e-03	18.190476	2	13	mesodermal cell differentiation
GO:0048048	8.058793e-03	18.190476	2	13	embryonic eye morphogenesis
GO:0001708	7.753839e-03	8.650000	3	39	cell fate specification
GO:2000027	7.207887e-03	8.901429	3	38	regulation of organ morphogenesis
GO:0065007	3.051087e-03	6.012286	21	1364	biological regulation
GO:0060900	2.989355e-03	33.428571	2	8	embryonic camera-type eye formation
GO:0010467	7.221642e-06	16.464646	8	206	gene expression
GO:0034645	2.150952e-06	13.817073	11	339	cellular macromolecule biosynthetic process
GO:0000122	2.112818e-05	11.112500	7	87	negative regulation of transcription from RNA polymerase II promoter
GO:0001657	2.072009e-05	19.952107	5	34	ureteric bud development
GO:0051253	4.072409e-05	8.477867	8	133	negative regulation of RNA metabolic process
GO:0007389	2.520671e-05	8.060440	9	165	pattern specification process
GO:0009790	6.046347e-05	6.548857	10	232	embryo development
GO:0010629	5.625185e-05	8.065140	8	139	negative regulation of gene expression
GO:0060562	6.505326e-06	11.294118	6	70	epithelial tube morphogenesis
GO:0051172	6.242300e-05	7.935840	8	141	negative regulation of nitrogen compound metabolic process
GO:0001763	4.256890e-04	9.837165	5	63	morphogenesis of a branching structure
GO:0010628	4.158434e-04	5.384728	9	235	positive regulation of gene expression
GO:0035137	8.026296e-04	20.970000	3	18	hindlimb morphogenesis
GO:0021545	8.026296e-04	20.970000	3	18	cranial nerve development
GO:0031328	7.057698e-04	4.944444	9	252	positive regulation of cellular biosynthetic process
GO:0051173	4.729417e-04	5.260248	9	239	positive regulation of nitrogen compound metabolic process
GO:0044238	4.690956e-04	5.434772	19	1004	primary metabolic process
GO:0051090	2.954258e-03	8.023392	4	58	regulation of sequence-specific DNA binding transcription factor activity
GO:0001755	9.340545e-03	16.666667	2	14	neural crest cell migration
GO:0010557	4.295455e-04	5.338263	9	236	positive regulation of macromolecule biosynthetic process
GO:0048935	2.989355e-03	33.428571	2	8	peripheral nervous system neuron development
GO:0035107	1.829837e-03	9.249720	4	51	appendage morphogenesis

Table A.2: E1 - PC Repressed - lm

Intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0021524	0.0009367681	Inf	1	1	visceral motor neuron differentiation
GO:0031103	0.0028089875	1065.5000	1	3	axon regeneration
GO:0048936	0.0009367681	Inf	1	1	peripheral nervous system neuron axonogenesis
GO:0031290	0.0037444388	710.0000	1	4	retinal ganglion cell axon guidance
GO:0021559	0.0037444388	710.0000	1	4	trigeminal nerve development
GO:0048880	0.0056140243	425.6000	1	6	sensory system development
GO:0060384	0.0046794510	532.2500	1	5	innervation
GO:0048934	0.0074818540	303.7143	1	8	peripheral nervous system neuron differentiation

Table A.3: E1 - PC Repressed - intersection

Mainly developmental, regulation and metabolic processes.

A.1.1.2 E2 - Low signal

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0007218	0.008901061	Inf	1	52	neuropeptide signaling pathway

Table A.4: E2 - Low signal - rf

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0007218	0.008901061	Inf	1	52	neuropeptide signaling pathway

Table A.5: E2 - Low signal - lm

Intersection

No highly significant age prediction probes are associated with this state. (check)

A.1.1.3 E3 - Heterochromatin

Random Forest

No lumi ids??.

Linear Regression

"No results met the specified criteria" (R's error message when running hyperGTest).

Intersection

...

A.1.1.4 E4 - CNV/Rep

Random Forest

"No results met the specified criteria" (R's error message when running hyperGTest).

Linear Regression

"No results met the specified criteria" (R's error message when running hyperGTest).

Intersection

"No results met the specified criteria" (R's error message when running hyperGTest).

A.1.1.5 E5 - Strong promoter

These results show that the processes involved here are related to muscle contraction and cell death. This may provide an explanation for why older people have a harder time walking and exercising. The results below are consistent across multiple algorithms.

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0034587	0.001470498	1630.2000	1	2	piRNA metabolic process
GO:0030049	0.008065459	162.8400	1	11	muscle filament sliding
GO:0001510	0.005871184	232.7143	1	8	RNA methylation
GO:0070252	0.008795987	148.0182	1	12	actin-mediated cell contraction

Table A.6: E5-Strong promoter-rf

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0034587	0.001470498	1630.2000	1	2	piRNA metabolic process
GO:0030049	0.008065459	162.8400	1	11	muscle filament sliding
GO:0001510	0.005871184	232.7143	1	8	RNA methylation
GO:0070252	0.008795987	148.0182	1	12	actin-mediated cell contraction

Table A.7: E5-Strong promoter-lm

Intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0030049	0.001348370	Inf	1	11	muscle filament sliding
GO:0030048	0.002574160	Inf	1	21	actin filament-based movement
GO:0070252	0.001470949	Inf	1	12	actin-mediated cell contraction
GO:0006921	0.005883795	Inf	1	48	cellular component disassembly involved in apoptosis

Table A.8: E5-Strong promoter-intersection

Muscle related.

A.1.1.6 E6 - Poised promoter

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0060849	0.0071576339	293.333333	1	2	regulation of transcription involved in lymphatic endothelial cell fate commitment
GO:0070172	0.0071576339	293.333333	1	2	positive regulation of tooth mineralization
GO:0071281	0.0035849057	Inf	1	1	cellular response to iron ion
GO:0030900	0.0048717251	10.297730	3	110	forebrain development
GO:0045935	0.0052633206	4.669798	6	481	positive regulation of nucleobase-containing compound metabolic process
GO:0045892	0.0061723683	5.190126	5	345	negative regulation of transcription, DNA-dependent
GO:0021623	0.0035849057	Inf	1	1	oculomotor nerve formation
GO:0031394	0.0035849057	Inf	1	1	positive regulation of prostaglandin biosynthetic process
GO:0042631	0.0035849057	Inf	1	1	cellular response to water deprivation
GO:0071105	0.0035849057	Inf	1	1	response to interleukin-11
GO:0010558	0.0093660906	4.659005	5	381	negative regulation of macromolecule biosynthetic process
GO:0060537	0.0098341580	7.797883	3	127	muscle tissue development
GO:0071104	0.0071576339	293.333333	1	2	response to interleukin-9
GO:0045934	0.0081955332	4.824372	5	369	negative regulation of nucleobase-containing compound metabolic process
GO:0021761	0.0063799601	19.297794	2	34	limbic system development
GO:0045165	0.0066164227	9.066589	3	110	cell fate commitment
GO:0060836	0.0071576339	293.333333	1	2	lymphatic endothelial cell differentiation
GO:0060839	0.0071576339	293.333333	1	2	endothelial cell fate commitment
GO:0048513	0.0001588439	6.042135	11	990	organ development
GO:0007417	0.0009559532	6.728387	6	345	central nervous system development
GO:0002027	0.0009277419	56.363636	2	13	regulation of heart rate
GO:0001503	0.0013630151	10.012652	4	141	ossification
GO:0021879	0.0010800534	51.656863	2	14	forebrain neuron differentiation
GO:0060173	0.0021562778	13.758803	3	74	limb development
GO:0009891	0.0013978265	5.480807	7	515	positive regulation of biosynthetic process
GO:0007519	0.0023278503	13.376712	3	76	skeletal muscle tissue development
GO:0048701	0.0022263204	34.398693	2	20	embryonic cranial skeleton morphogenesis
GO:0032849	0.0071576339	293.333333	1	2	positive regulation of cellular pH reduction
GO:0021985	0.0071576339	293.333333	1	2	neurohypophysis development
GO:0043627	0.0030984427	12.037037	3	84	response to estrogen stimulus
GO:0045893	0.0027905740	5.369525	6	424	positive regulation of transcription, DNA-dependent
GO:0002125	0.0035849057	Inf	1	1	maternal aggressive behavior
GO:0001501	0.0034097913	7.689642	4	181	skeletal system development
GO:0003409	0.0035849057	Inf	1	1	optic cup structural organization
GO:0003404	0.0035849057	Inf	1	1	optic vesicle morphogenesis
GO:0007192	0.0035849057	Inf	1	1	activation of adenylate cyclase activity by serotonin receptor signaling pathway
GO:0006419	0.0035849057	Inf	1	1	alanyl-tRNA aminoacylation
GO:0021506	0.0035849057	Inf	1	1	anterior neuropore closure
GO:0009956	0.0035849057	Inf	1	1	radial pattern formation
GO:0060349	0.0067522098	18.709447	2	35	bone morphogenesis
GO:0001764	0.0067522098	18.709447	2	35	neuron migration
GO:0021979	0.0071576339	293.333333	1	2	hypothalamus cell differentiation
GO:0021893	0.0071576339	293.333333	1	2	cerebral cortex GABAergic interneuron fate commitment
GO:0021882	0.0071576339	293.333333	1	2	regulation of transcription from RNA polymerase II promoter involved in forebrain neuron fate commitment
GO:0021557	0.0071576339	293.333333	1	2	oculomotor nerve development
GO:0003334	0.0071576339	293.333333	1	2	keratinocyte development
GO:0030855	0.0071293099	8.814205	3	113	epithelial cell differentiation

Table A.9: E6 - Poised promoter - rf

rf-The results above imply that poised promoters are associated with generation, development and fate commitment of different tissues in the brain.

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0021882	7.533639e-03	277.842105	1	2	regulation of transcription from RNA polymerase II promoter involved in forebrain neuron fate commitment
GO:0014706	9.964464e-03	7.719840	3	121	striated muscle tissue development
GO:0003334	7.533639e-03	277.842105	1	2	keratinocyte development
GO:0071281	3.773585e-03	Inf	1	1	cellular response to iron ion
GO:0030500	6.270668e-03	19.444444	2	32	regulation of bone mineralization
GO:0071639	3.773585e-03	Inf	1	1	positive regulation of monocyte chemotactic protein-1 production
GO:0032849	7.533639e-03	277.842105	1	2	positive regulation of cellular pH reduction
GO:0042518	7.533639e-03	277.842105	1	2	negative regulation of tyrosine phosphorylation of Stat3 protein
GO:0042538	7.533639e-03	277.842105	1	2	hyperosmotic salinity response
GO:0060836	7.533639e-03	277.842105	1	2	lymphatic endothelial cell differentiation
GO:0060839	7.533639e-03	277.842105	1	2	endothelial cell fate commitment
GO:0000122	6.354999e-03	6.350000	4	204	negative regulation of transcription from RNA polymerase II promoter
GO:0070172	7.533639e-03	277.842105	1	2	positive regulation of tooth mineralization
GO:2000679	7.533639e-03	277.842105	1	2	positive regulation of transcription regulatory region DNA binding
GO:0010563	7.893174e-03	17.143791	2	36	negative regulation of phosphorus metabolic process
GO:0042127	7.995012e-03	4.198948	6	495	regulation of cell proliferation
GO:0060562	7.102987e-03	8.782805	3	107	epithelial tube morphogenesis
GO:0042326	7.471506e-03	17.666667	2	35	negative regulation of phosphorylation
GO:0008015	1.459860e-03	9.750000	4	136	blood circulation
GO:0001503	1.668910e-03	9.385036	4	141	ossification
GO:0045777	1.778671e-03	39.000000	2	17	positive regulation of blood pressure
GO:0048048	1.996483e-03	36.555556	2	18	embryonic eye morphogenesis
GO:0045923	1.029398e-03	53.222222	2	13	positive regulation of fatty acid metabolic process
GO:0048706	1.064015e-03	17.742081	3	55	embryonic skeletal system development
GO:0021879	1.198249e-03	48.777778	2	14	forebrain neuron differentiation
GO:0035113	1.371257e-03	16.170279	3	60	embryonic appendage morphogenesis
GO:0048522	5.439158e-03	3.646030	10	1243	positive regulation of cellular process
GO:0001843	5.524370e-03	20.841270	2	30	neural tube closure
GO:0051094	4.306475e-03	5.632768	5	300	positive regulation of developmental process
GO:0022008	4.488027e-03	4.304940	7	594	neurogenesis
GO:0035108	2.141347e-03	13.730465	3	70	limb morphogenesis
GO:0031076	2.468100e-03	32.481481	2	20	embryonic camera-type eye development
GO:0033128	3.773585e-03	Inf	1	1	negative regulation of histone phosphorylation
GO:0042631	3.773585e-03	Inf	1	1	cellular response to water deprivation
GO:0048593	2.371159e-05	30.447674	4	47	camera-type eye morphogenesis
GO:0032355	5.491653e-05	24.194444	4	58	response to estradiol stimulus
GO:0048701	5.030235e-05	54.633218	3	20	embryonic cranial skeleton morphogenesis
GO:0048736	1.432599e-04	18.607143	4	74	appendage development
GO:0061029	8.081568e-05	293.222222	2	4	eyelid development in camera-type eye
GO:0009887	1.723221e-04	7.873009	7	345	organ morphogenesis
GO:0007417	1.723221e-04	7.873009	7	345	central nervous system development
GO:0060363	2.011253e-04	146.555556	2	6	cranial suture morphogenesis
GO:0048545	1.987633e-04	11.639456	5	152	response to steroid hormone stimulus
GO:0021761	7.060480e-03	18.222222	2	34	limbic system development
GO:0031394	3.773585e-03	Inf	1	1	positive regulation of prostaglandin biosynthetic process
GO:0060849	7.533639e-03	277.842105	1	2	regulation of transcription involved in lymphatic endothelial cell fate commitment
GO:0060429	9.962931e-03	5.539474	4	232	epithelium development
GO:0009790	9.314080e-03	4.624413	5	360	embryo development
GO:0021623	3.773585e-03	Inf	1	1	oculomotor nerve formation
GO:0009719	8.995788e-03	4.666667	5	357	response to endogenous stimulus
GO:0001841	8.768128e-03	16.185185	2	38	neural tube formation
GO:0051253	8.482499e-03	4.738713	5	352	negative regulation of RNA metabolic process
GO:0009892	8.312555e-03	4.161402	6	499	negative regulation of metabolic process
GO:0021985	7.533639e-03	277.842105	1	2	neurohypophysis development
GO:0021979	7.533639e-03	277.842105	1	2	hypothalamus cell differentiation
GO:0060349	2.785469e-04	28.941176	3	35	bone morphogenesis
GO:0001764	2.785469e-04	28.941176	3	35	neuron migration
GO:0060900	3.737362e-04	97.666667	2	8	embryonic camera-type eye formation
GO:0010460	2.809380e-04	117.222222	2	7	positive regulation of heart rate
GO:0008016	5.515598e-04	22.549498	3	44	regulation of heart contraction
GO:0030900	4.400320e-04	13.859838	4	110	forebrain development
GO:0045165	6.585647e-04	12.202830	4	110	cell fate commitment
GO:0051240	5.748949e-04	9.129032	5	191	positive regulation of multicellular organismal process
GO:0001654	1.003373e-03	10.842437	4	123	eye development
GO:0003015	8.527789e-04	19.235294	3	51	heart process
GO:0051216	3.365696e-03	11.618019	3	82	cartilage development
GO:0035115	2.721738e-03	30.766082	2	21	embryonic forelimb morphogenesis
GO:0021506	3.773585e-03	Inf	1	1	anterior neuropore closure
GO:0009956	3.773585e-03	Inf	1	1	radial pattern formation
GO:0007192	3.773585e-03	Inf	1	1	activation of adenylate cyclase activity by serotonin receptor signaling pathway
GO:0003409	3.773585e-03	Inf	1	1	optic cup structural organization
GO:0003404	3.773585e-03	Inf	1	1	optic vesicle morphogenesis
GO:0021557	7.533639e-03	277.842105	1	2	oculomotor nerve development
GO:0031328	9.233773e-03	4.061224	6	510	positive regulation of cellular biosynthetic process
GO:0002125	3.773585e-03	Inf	1	1	maternal aggressive behavior
GO:0021893	7.533639e-03	277.842105	1	2	cerebral cortex GABAergic interneuron fate commitment
GO:2000026	6.472886e-03	4.406593	6	474	regulation of multicellular organismal development

Table A.10: E6 - Poised promoter - lm

Intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0014014	0.0071576339	293.3333	1	19	negative regulation of gliogenesis
GO:0021872	0.0067815572	310.6471	1	18	forebrain generation of neurons
GO:0021766	0.0090369489	229.3478	1	24	hippocampus development
GO:0048701	0.0075336393	277.8421	1	20	embryonic cranial skeleton morphogenesis
GO:0051216	0.0002364989	Inf	2	82	cartilage development
GO:0021893	0.0007546458	5297.0000	1	2	cerebral cortex GABAergic interneuron fate commitment
GO:0021882	0.0007546458	5297.0000	1	2	regulation of transcription from RNA polymerase II promoter involved in forebrain neuron fate commitment
GO:0021877	0.0018860803	1323.5000	1	5	forebrain neuron fate commitment
GO:0048755	0.0015090067	1765.0000	1	4	branching morphogenesis of a nerve
GO:0009954	0.0049001058	440.5000	1	13	proximal/distal pattern formation
GO:0048715	0.0037703803	587.6667	1	10	negative regulation of oligodendrocyte differentiation
GO:0021772	0.0064054094	330.1250	1	17	olfactory bulb development
GO:0021895	0.0049001058	440.5000	1	13	cerebral cortex neuron differentiation

Table A.11: E6 - Poised promoter - intersection

Development and formation.

A.1.1.7 E7 - Active TSS flankin

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0030819	0.005590743	267.3000	1	5	positive regulation of cAMP biosynthetic process
GO:0042596	0.005590743	267.3000	1	5	fear response
GO:0048025	0.007819735	178.1333	1	7	negative regulation of nuclear mRNA splicing, via spliceosome
GO:0002027	0.008932668	152.6571	1	8	regulation of heart rate
GO:0003084	0.004474683	356.4667	1	4	positive regulation of systemic arterial blood pressure
GO:0055021	0.004474683	356.4667	1	4	regulation of cardiac muscle tissue growth
GO:0060080	0.004474683	356.4667	1	4	regulation of inhibitory postsynaptic membrane potential
GO:0014897	0.005590743	267.3000	1	5	striated muscle hypertrophy
GO:0030810	0.008932668	152.6571	1	8	positive regulation of nucleotide biosynthetic process
GO:0032411	0.008932668	152.6571	1	8	positive regulation of transporter activity
GO:0007611	0.001013532	61.7093	2	45	learning or memory
GO:0046878	0.001120239	Inf	1	1	positive regulation of saliva secretion
GO:0003061	0.001120239	Inf	1	1	positive regulation of the force of heart contraction by norepinephrine
GO:0060421	0.001120239	Inf	1	1	positive regulation of heart growth
GO:0055025	0.001120239	Inf	1	1	positive regulation of cardiac muscle tissue development
GO:0072365	0.001120239	Inf	1	1	regulation of cellular ketone metabolic process by negative regulation of transcription from RNA polymerase II promoter
GO:0061051	0.001120239	Inf	1	1	positive regulation of cell growth involved in cardiac muscle cell development
GO:0001996	0.002239432	1069.8000	1	2	positive regulation of heart rate by epinephrine-norepinephrine
GO:2000725	0.001120239	Inf	1	1	regulation of cardiac muscle cell differentiation
GO:0003301	0.002239432	1069.8000	1	2	physiological cardiac muscle hypertrophy
GO:0002025	0.002239432	1069.8000	1	2	vasodilation by norepinephrine-epinephrine involved in regulation of systemic arterial blood pressure
GO:0035811	0.002239432	1069.8000	1	2	negative regulation of urine volume
GO:0014742	0.002239432	1069.8000	1	2	positive regulation of muscle hypertrophy
GO:0010611	0.003357580	534.8000	1	3	regulation of cardiac muscle hypertrophy
GO:0003057	0.003357580	534.8000	1	3	regulation of the force of heart contraction by chemical signal
GO:0045986	0.003357580	534.8000	1	3	negative regulation of smooth muscle contraction
GO:0044058	0.003357580	534.8000	1	3	regulation of digestive system process
GO:0051929	0.003357580	534.8000	1	3	positive regulation of calcium ion transport via voltage-gated calcium channel activity
GO:0048636	0.003357580	534.8000	1	3	positive regulation of muscle organ development

Table A.12: E7 - Active TSS flankin - rf

rf-Neuron development and immune system - suppressed and weakens with age.

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0007189	0.0081921002	178.06667	1	11	activation of adenylate cyclase activity by G-protein signaling pathway
GO:0010578	0.0081921002	178.06667	1	11	regulation of adenylate cyclase activity involved in G-protein signaling pathway
GO:0007610	0.0035013458	40.16923	2	132	behavior
GO:0014897	0.0037299476	445.66667	1	5	striated muscle hypertrophy
GO:0030819	0.0037299476	445.66667	1	5	positive regulation of cAMP biosynthetic process
GO:0042596	0.0037299476	445.66667	1	5	fear response
GO:0003084	0.0029847942	594.33333	1	4	positive regulation of systemic arterial blood pressure
GO:0055021	0.0029847942	594.33333	1	4	regulation of cardiac muscle tissue growth
GO:0060080	0.0029847942	594.33333	1	4	regulation of inhibitory postsynaptic membrane potential
GO:0030097	0.0031939521	42.16129	2	126	hemopoiesis
GO:0007613	0.0089343325	161.84848	1	12	memory
GO:0003014	0.0089343325	161.84848	1	12	renal system process
GO:0005980	0.0089343325	161.84848	1	12	glycogen catabolic process
GO:0002520	0.0039879438	37.50360	2	141	immune system development
GO:0045321	0.0045049558	35.16216	2	150	leukocyte activation
GO:0046620	0.0074494511	197.88889	1	10	regulation of organ growth
GO:0003044	0.0081921002	178.06667	1	11	regulation of systemic arterial blood pressure mediated by a chemical signal
GO:0050890	0.0004665545	115.34783	2	48	cognition
GO:0003061	0.0007468260	Inf	1	1	positive regulation of the force of heart contraction by norepinephrine
GO:0030098	0.0005269595	108.22449	2	51	lymphocyte differentiation
GO:0055025	0.0007468260	Inf	1	1	positive regulation of cardiac muscle tissue development
GO:0046878	0.0007468260	Inf	1	1	positive regulation of saliva secretion
GO:0061051	0.0007468260	Inf	1	1	positive regulation of cell growth involved in cardiac muscle cell development
GO:0060421	0.0007468260	Inf	1	1	positive regulation of heart growth
GO:0001996	0.0014932336	1783.66667	1	2	positive regulation of heart rate by epinephrine-norepinephrine
GO:2000725	0.0007468260	Inf	1	1	regulation of cardiac muscle cell differentiation
GO:0043500	0.0074494511	197.88889	1	10	muscle adaptation
GO:0048513	0.0068883371	21.29047	3	664	organ development
GO:0003301	0.0014932336	1783.66667	1	2	physiological cardiac muscle hypertrophy
GO:0002025	0.0014932336	1783.66667	1	2	vasodilation by norepinephrine-epinephrine involved in regulation of systemic arterial blood pressure
GO:0035811	0.0014932336	1783.66667	1	2	negative regulation of urine volume
GO:0014742	0.0014932336	1783.66667	1	2	positive regulation of muscle hypertrophy
GO:0010611	0.0022392230	891.66667	1	3	regulation of cardiac muscle hypertrophy
GO:0003057	0.0022392230	891.66667	1	3	regulation of the force of heart contraction by chemical signal
GO:0045986	0.0022392230	891.66667	1	3	negative regulation of smooth muscle contraction
GO:0044058	0.0022392230	891.66667	1	3	regulation of digestive system process
GO:0051929	0.0022392230	891.66667	1	3	positive regulation of calcium ion transport via voltage-gated calcium channel activity
GO:0048636	0.0022392230	891.66667	1	3	positive regulation of muscle organ development
GO:0002027	0.0059629017	254.52381	1	8	regulation of heart rate
GO:0048025	0.0052190012	297.00000	1	7	negative regulation of nuclear mRNA splicing, via spliceosome
GO:0055013	0.0067063850	222.66667	1	9	cardiac muscle cell development
GO:0045823	0.0067063850	222.66667	1	9	positive regulation of heart contraction
GO:0045776	0.0067063850	222.66667	1	9	negative regulation of blood pressure
GO:0030801	0.0067063850	222.66667	1	9	positive regulation of cyclic nucleotide metabolic process
GO:0032411	0.0059629017	254.52381	1	8	positive regulation of transporter activity
GO:0030810	0.0059629017	254.52381	1	8	positive regulation of nucleotide biosynthetic process

Table A.13: E7 - Active TSS flankin - lm

Intersection

No highly significant age prediction probes are associated with this state.

Heart learning and memory.

A.1.1.8 E8 - Active enhancer

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0007567	0.0007130125	Inf	1	2	parturition

Table A.14: E8 - Active enhancer - rf

Linear Regression

XXX

Intersection

No highly significant age prediction probes are associated with this state.

A.1.1.9 E9 - Weak enhancer

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0006200	0.009414481	220.0870	1	24	ATP catabolic process
GO:0006636	0.008631642	241.1429	1	22	unsaturated fatty acid biosynthetic process
GO:0042908	0.000393159	Inf	1	1	xenobiotic transport
GO:0015911	0.001572172	1694.0000	1	4	plasma membrane long-chain fatty acid transport
GO:0046618	0.000393159	Inf	1	1	drug export
GO:0043526	0.003535648	634.6250	1	9	neuroprotection
GO:0033700	0.001572172	1694.0000	1	4	phospholipid efflux
GO:0043449	0.004320498	507.5000	1	11	cellular alkene metabolic process
GO:0019370	0.003928112	564.0000	1	10	leukotriene biosynthetic process
GO:0015908	0.007065035	298.1176	1	18	fatty acid transport
GO:0006692	0.006281268	338.0000	1	16	prostanoid metabolic process

Table A.15: E9 - Weak enhancer - rf

rf-Metabolic processes - metabolism slows with age.

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0043045	0.0003931590	Inf	1	1	DNA methylation involved in embryo development
GO:0044030	0.0003931590	Inf	1	1	regulation of DNA methylation
GO:0044027	0.0003931590	Inf	1	1	hypermethylation of CpG island
GO:0090116	0.0007862408	5084.0000	1	2	C-5 methylation of cytosine
GO:0043046	0.0007862408	5084.0000	1	2	DNA methylation involved in gamete generation
GO:0006349	0.0011792452	2541.5000	1	3	regulation of gene expression by genetic imprinting
GO:0006346	0.0011792452	2541.5000	1	3	methylation-dependent chromatin silencing
GO:0045814	0.0047128064	461.2727	1	12	negative regulation of gene expression, epigenetic
GO:0006305	0.0043204977	507.5000	1	11	DNA alkylation

Table A.16: E9 - Weak enhancer - lm

Intersection

No highly significant age prediction probes are associated with this state.

A.1.1.10 E10 - Weak transcribed/ active proximal

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0035329	0.0094978962	140.88462	1	14	hippo signaling cascade
GO:0050709	0.0088218778	152.64583	1	13	negative regulation of protein secretion
GO:0043314	0.0006808279	Inf	1	1	negative regulation of neutrophil degranulation
GO:0051051	0.0018228978	48.75421	2	101	negative regulation of transport
GO:0002283	0.0013612849	1834.50000	1	2	neutrophil activation involved in immune response
GO:0002446	0.0047580122	305.54167	1	7	neutrophil mediated immunity
GO:0045920	0.0040794073	366.70000	1	6	negative regulation of exocytosis
GO:0050766	0.0047580122	305.54167	1	7	positive regulation of phagocytosis
GO:0043300	0.0047580122	305.54167	1	7	regulation of leukocyte degranulation
GO:0051046	0.0055061830	27.29143	2	177	regulation of secretion
GO:0002886	0.0054362472	261.85714	1	8	regulation of myeloid leukocyte mediated immunity

Table A.17: E10 - Weak transcribed/ active proximal - rf

rf-No lumi probes were found.

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0050709	0.005301783	305.3750	1	13	negative regulation of protein secretion
GO:0035329	0.005708835	281.8462	1	14	hippo signaling cascade

Table A.18: E10 - Weak transcribed/ active proximal - lm

Intersection

XXX

No lumi probes were found.

Regulation and signaling.

A.1.1.11 E11 - Strong transcription

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0030163	0.001675411	Inf	2	192	protein catabolic process
GO:0006508	0.003908744	Inf	2	293	proteolysis
GO:0008089	0.001709219	1558.0000	1	4	anterograde axon cargo transport
GO:0010970	0.008105769	258.8333	1	19	microtubule-based transport

Table A.19: E11 - Strong transcription - rf

Intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0008089	0.0008548835	Inf	1	4	anterograde axon cargo transport
GO:0010970	0.0040606967	Inf	1	19	microtubule-based transport

Table A.20: E11 - Strong transcription - intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0008089	0.0008548835	Inf	1	4	anterograde axon cargo transport
GO:0010970	0.0040606967	Inf	1	19	microtubule-based transport

Table A.21: E11 - Strong transcription - lm

Linear Regression

Transport.

A.2 GO Categories of Highly Predictive Probes - across all 5 folds

A.2.1 BF - Biological Process

A.2.1.1 E1 - PC Repressed

Looking at the repressed processing, it seems that these terms mostly include tissue development and regeneration, which are processes that are often slowed down as we age. These results may help explain why. These results are consistent across multiple algorithms (Random Forest, Linear Regression and the intersection of the resulting probes).

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0021524	0.0009367681	Inf	1	1	visceral motor neuron differentiation
GO:0048936	0.0009367681	Inf	1	1	peripheral nervous system neuron axonogenesis
GO:0031103	0.0028089875	1065.5	1	3	axon regeneration
GO:0021559	0.0037444388	710.0	1	4	trigeminal nerve development
GO:0031290	0.0037444388	710.0	1	4	retinal ganglion cell axon guidance
GO:0060384	0.0046794510	532.25	1	5	innervation
GO:0048880	0.0056140243	425.60	1	6	sensory system development
GO:0048934	0.0074818540	303.7143	1	8	peripheral nervous system neuron differentiation

Table A.22: E1 - PC Repressed - rf

Linear Regression

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0021524	4.683841e-03	Inf	0.004683841	1	1	visceral motor neuron differentiation
GO:0046671	4.683841e-03	Inf	0.004683841	1	1	negative regulation of retinal cell programmed cell death
GO:0048050	4.683841e-03	Inf	0.004683841	1	1	post-embryonic eye morphogenesis
GO:0048936	4.683841e-03	Inf	0.004683841	1	1	peripheral nervous system neuron axonogenesis
GO:0044281	2.991514e-03	8.265512	3.283372365	8	701	small molecule metabolic process
GO:0045665	3.237110e-03	31.000000	0.088992974	2	19	negative regulation of neuron differentiation
GO:0048731	3.488219e-03	13.138167	2.796778778	7	738	system development
GO:0032502	4.172218e-03	11.238095	4.468384075	9	954	developmental process
GO:0045944	5.104748e-03	8.078189	0.777517564	4	166	positive regulation of transcription from RNA polymerase II promoter
GO:0001658	5.159773e-03	23.897727	0.112412178	2	24	branching involved in ureteric bud morphogenesis
GO:0006355	9.189975e-05	15.721578	2.056206089	8	439	regulation of transcription, DNA-dependent
GO:0010556	1.470170e-04	14.518519	2.187353630	8	467	regulation of macromolecule biosynthetic process
GO:0032774	1.001770e-04	15.495413	2.079625293	8	444	RNA biosynthetic process
GO:0090304	2.050564e-04	13.708333	2.285714286	8	488	nucleic acid metabolic process
GO:0034645	1.897101e-04	13.894737	2.262295082	8	483	cellular macromolecule biosynthetic process
GO:0031326	2.775370e-04	13.000000	2.379391101	8	508	regulation of cellular biosynthetic process
GO:0010467	2.500167e-04	13.241379	2.346604215	8	501	gene expression
GO:0019219	3.031459e-04	12.798419	2.407494145	8	514	regulation of nucleobase-containing compound metabolic process
GO:0048880	2.933541e-04	132.562500	0.028103044	2	6	sensory system development
GO:0001657	4.106227e-04	28.949309	0.159250585	3	34	ureteric bud development
GO:0009887	3.967047e-04	11.199203	1.203747073	6	257	organ morphogenesis
GO:0060284	1.303912e-03	12.096096	0.538641686	4	115	regulation of cell development
GO:0030154	9.599426e-04	10.358108	2.810304450	8	600	cell differentiation
GO:0034641	1.560678e-03	9.428120	3.002341920	8	641	cellular nitrogen compound metabolic process
GO:0009058	1.327411e-03	9.731826	2.936768150	8	627	biosynthetic process
GO:0007420	2.316701e-03	10.230769	0.627634660	4	134	brain development
GO:0001755	1.744383e-03	44.020833	0.065573770	2	14	neural crest cell migration
GO:0021545	2.903626e-03	32.953125	0.084309133	2	18	cranial nerve development
GO:0019222	2.585359e-03	8.518409	3.217798595	8	687	regulation of metabolic process
GO:0014033	5.159773e-03	23.897727	0.112412178	2	24	neural crest cell differentiation
GO:0007422	5.159773e-03	23.897727	0.112412178	2	24	peripheral nervous system development
GO:0021658	9.347928e-03	236.000000	0.009367681	1	2	rhombomere 3 morphogenesis
GO:0021568	9.347928e-03	236.000000	0.009367681	1	2	rhombomere 2 development
GO:0007411	8.644579e-03	9.259878	0.454332553	3	97	axon guidance
GO:0048666	7.124463e-03	7.292135	0.852459016	4	182	neuron development
GO:0000122	6.372593e-03	10.413265	0.407494145	3	87	negative regulation of transcription from RNA polymerase II promoter

Table A.23: E1 - PC Repressed - lm

Intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0021524	0.0009367681	Inf	1	1	visceral motor neuron differentiation
GO:0031103	0.0028089875	1065.5000	1	3	axon regeneration
GO:0048936	0.0009367681	Inf	1	1	peripheral nervous system neuron axonogenesis
GO:0031290	0.0037444388	710.0000	1	4	retinal ganglion cell axon guidance
GO:0021559	0.0037444388	710.0000	1	4	trigeminal nerve development
GO:0048880	0.0056140243	425.6000	1	6	sensory system development
GO:0060384	0.0046794510	532.2500	1	5	innervation
GO:0048934	0.0074818540	303.7143	1	8	peripheral nervous system neuron differentiation

Table A.24: E1 - PC Repressed - intersection

A.2.1.2 E2 - Low signal

Random Forest

No highly significant age prediction probes are associated with this state.

Linear Regression

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0007218	0.008901061	Inf	0.008901061	1	52	neuropeptide signaling pathway

Table A.25: E2 - Low signal - lm

Intersection

No highly significant age prediction probes are associated with this state. (check)

A.2.1.3 E3 - Heterochromatin

Random Forest

No lumi probes were found (converting the probe id to lumi id resulted in NA).

Linear Regression

No lumi probes were found (converting the probe id to lumi id resulted in NA).

Intersection

No lumi probes were found (converting the probe id to lumi id resulted in NA).

A.2.1.4 E4 - CNV/Rep

Random Forest

No lumi probes were found.

Linear Regression

No lumi probes were found.

Intersection

"No results met the specified criteria" (R's error message when running hyperGTest).

A.2.1.5 E5 - Strong promoter

These results show that the processes involved here are related to muscle contraction and cell death. This may provide an explanation for why older people have a harder time walking and exercising. The results below are consistent across multiple algorithms.

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0030049	0.001348370	Inf	1	11	muscle filament sliding
GO:0070252	0.001470949	Inf	1	12	actin-mediated cell contraction
GO:0030048	0.002574160	Inf	1	21	actin filament-based movement
GO:0006921	0.005883795	Inf	1	48	cellular component disassembly involved in apoptosis

Table A.26: E5-Strong promoter-rf

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0034587	0.0007353842	4077.0000	1	2	piRNA metabolic process
GO:0030049	0.0040401519	407.2500	1	11	muscle filament sliding
GO:0001510	0.0029393735	582.0000	1	8	RNA methylation
GO:0030048	0.0077035611	203.3750	1	21	actin filament-based movement
GO:0070252	0.0044068978	370.1818	1	12	actin-mediated cell contraction
GO:0031047	0.0091664057	169.3958	1	25	gene silencing by RNA

Table A.27: E5-Strong promoter-lm

Intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0030049	0.001348370	Inf	1	11	muscle filament sliding
GO:0030048	0.002574160	Inf	1	21	actin filament-based movement
GO:0070252	0.001470949	Inf	1	12	actin-mediated cell contraction
GO:0006921	0.005883795	Inf	1	48	cellular component disassembly involved in apoptosis

Table A.28: E5-Strong promoter-intersection

A.2.1.6 E6 - Poised promoter

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0014014	0.0071576339	293.3333	1	19	negative regulation of gliogenesis
GO:0021872	0.0067815572	310.6471	1	18	forebrain generation of neurons
GO:0021766	0.0090369489	229.3478	1	24	hippocampus development
GO:0048701	0.0075336393	277.8421	1	20	embryonic cranial skeleton morphogenesis
GO:0051216	0.0002364989	Inf	2	82	cartilage development
GO:0021893	0.0007546458	5297.0000	1	2	cerebral cortex GABAergic interneuron fate commitment
GO:0021882	0.0007546458	5297.0000	1	2	regulation of transcription from RNA polymerase II promoter involved in forebrain neuron fate commitment
GO:0021877	0.0018860803	1323.5000	1	5	forebrain neuron fate commitment
GO:0048755	0.0015090067	1765.0000	1	4	branching morphogenesis of a nerve
GO:0009954	0.0049001058	440.5000	1	13	proximal/distal pattern formation
GO:0048715	0.0037703803	587.6667	1	10	negative regulation of oligodendrocyte differentiation
GO:0021772	0.0064054094	330.1250	1	17	olfactory bulb development
GO:0021895	0.0049001058	440.5000	1	13	cerebral cortex neuron differentiation

Table A.29: E6 - Poised promoter - rf

rf-The results above imply that poised promoters are associated with generation, development and fate commitment of different tissues in the brain.

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0021953	0.0062823709	20.397661	2	59	central nervous system neuron differentiation
GO:0014049	0.0082784131	176.200000	1	4	positive regulation of glutamate secretion
GO:0030497	0.0082784131	176.200000	1	4	fatty acid elongation
GO:0042711	0.0082784131	176.200000	1	4	maternal behavior
GO:0007621	0.0062146716	264.350000	1	3	negative regulation of female receptivity
GO:0007625	0.0062146716	264.350000	1	3	grooming behavior
GO:0034626	0.0062146716	264.350000	1	3	fatty acid elongation, polyunsaturated fatty acid
GO:0060180	0.0062146716	264.350000	1	3	female mating behavior
GO:0042761	0.0082784131	176.200000	1	4	very long-chain fatty acid biosynthetic process
GO:0048755	0.0082784131	176.200000	1	4	branching morphogenesis of a nerve
GO:0030900	0.0001486878	21.016327	4	144	forebrain development
GO:0048513	0.0012704592	7.665819	7	990	organ development
GO:0002027	0.0003017201	106.626263	2	13	regulation of heart rate
GO:0002125	0.0020754717	Inf	1	1	maternal aggressive behavior
GO:0045165	0.0012713668	18.161215	3	110	cell fate commitment
GO:0009956	0.0020754717	Inf	1	1	radial pattern formation
GO:0007192	0.0020754717	Inf	1	1	activation of adenylate cyclase activity by serotonin receptor signaling pathway
GO:0042631	0.0020754717	Inf	1	1	cellular response to water deprivation
GO:0031394	0.0020754717	Inf	1	1	positive regulation of prostaglandin biosynthetic process
GO:0001764	0.0022449168	35.393939	2	35	neuron migration
GO:0009888	0.0021665332	7.999332	5	504	tissue development
GO:0021882	0.0041470267	528.800000	1	2	regulation of transcription from RNA polymerase II promoter involved in forebrain neuron fate commitment
GO:0003013	0.0023447433	14.537594	3	136	circulatory system process
GO:0032849	0.0041470267	528.800000	1	2	positive regulation of cellular pH reduction
GO:0021893	0.0041470267	528.800000	1	2	cerebral cortex GABAergic interneuron fate commitment
GO:0060839	0.0041470267	528.800000	1	2	endothelial cell fate commitment
GO:0060836	0.0041470267	528.800000	1	2	lymphatic endothelial cell differentiation
GO:0060047	0.0047241120	23.764172	2	51	heart contraction
GO:0060849	0.0041470267	528.800000	1	2	regulation of transcription involved in lymphatic endothelial cell fate commitment
GO:0060040	0.0082784131	176.200000	1	4	retinal bipolar neuron differentiation

Table A.30: E6 - Poised promoter - lm

Intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0014014	0.0071576339	293.3333	1	19	negative regulation of gliogenesis
GO:0021872	0.0067815572	310.6471	1	18	forebrain generation of neurons
GO:0021766	0.0090369489	229.3478	1	24	hippocampus development
GO:0048701	0.0075336393	277.8421	1	20	embryonic cranial skeleton morphogenesis
GO:0051216	0.0002364989	Inf	2	82	cartilage development
GO:0021893	0.0007546458	5297.0000	1	2	cerebral cortex GABAergic interneuron fate commitment
GO:0021882	0.0007546458	5297.0000	1	2	regulation of transcription from RNA polymerase II promoter involved in forebrain neuron fate commitment
GO:0021877	0.0018860803	1323.5000	1	5	forebrain neuron fate commitment
GO:0048755	0.0015090067	1765.0000	1	4	branching morphogenesis of a nerve
GO:0009954	0.0049001058	440.5000	1	13	proximal/distal pattern formation
GO:0048715	0.0037703803	587.6667	1	10	negative regulation of oligodendrocyte differentiation
GO:0021772	0.0064054094	330.1250	1	17	olfactory bulb development
GO:0021895	0.0049001058	440.5000	1	13	cerebral cortex neuron differentiation

Table A.31: E6 - Poised promoter - intersection

A.2.1.7 E7 - Active TSS flankin

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0007405	0.004339623	Inf	1	23	neuroblast proliferation
GO:0030183	0.005471698	Inf	1	29	B cell differentiation

Table A.32: E7 - Active TSS flankin - rf

rf-Neuron development and immune system - suppressed and weakens with age.

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0031279	0.0096860751	213.1600	1	26	regulation of cyclase activity
GO:0045762	0.0055938731	381.4286	1	15	positive regulation of adenylate cyclase activity
GO:0016202	0.0085708568	242.3636	1	23	regulation of striated muscle tissue development
GO:0051149	0.0059662401	355.9333	1	16	positive regulation of muscle cell differentiation
GO:0035051	0.0085708568	242.3636	1	23	cardiac cell differentiation
GO:0006937	0.0093144054	222.0833	1	25	regulation of muscle contraction
GO:0048588	0.0093144054	222.0833	1	25	developmental cell growth
GO:0048025	0.0026124266	891.3333	1	7	negative regulation of nuclear mRNA splicing, via spliceosome
GO:0002027	0.0029853515	763.8571	1	8	regulation of heart rate
GO:0030810	0.0029853515	763.8571	1	8	positive regulation of nucleotide biosynthetic process
GO:0032411	0.0029853515	763.8571	1	8	positive regulation of transporter activity
GO:0060080	0.0014932336	1783.6667	1	4	regulation of inhibitory postsynaptic membrane potential
GO:0014897	0.0018663677	1337.5000	1	5	striated muscle hypertrophy
GO:0030819	0.0018663677	1337.5000	1	5	positive regulation of cAMP biosynthetic process
GO:0042596	0.0018663677	1337.5000	1	5	fear response
GO:0050684	0.0074550109	280.7895	1	20	regulation of mRNA processing
GO:0007589	0.0081989779	253.9524	1	22	body fluid secretion
GO:0000272	0.0067107650	313.9412	1	18	polysaccharide catabolic process
GO:0043270	0.0074550109	280.7895	1	20	positive regulation of ion transport
GO:0030801	0.0033582066	668.2500	1	9	positive regulation of cyclic nucleotide metabolic process
GO:0045776	0.0033582066	668.2500	1	9	negative regulation of blood pressure
GO:0007586	0.0052214363	410.8462	1	14	digestion
GO:0019233	0.0052214363	410.8462	1	14	sensory perception of pain
GO:0003061	0.0003734130	Inf	1	1	positive regulation of the force of heart contraction by norepinephrine
GO:0055025	0.0003734130	Inf	1	1	positive regulation of cardiac muscle tissue development
GO:0046878	0.0003734130	Inf	1	1	positive regulation of saliva secretion
GO:0061051	0.0003734130	Inf	1	1	positive regulation of cell growth involved in cardiac muscle cell development
GO:0060421	0.0003734130	Inf	1	1	positive regulation of heart growth
GO:0001996	0.0007467563	5353.0000	1	2	positive regulation of heart rate by epinephrine-norepinephrine
GO:2000725	0.0003734130	Inf	1	1	regulation of cardiac muscle cell differentiation
GO:0003301	0.0007467563	5353.0000	1	2	physiological cardiac muscle hypertrophy
GO:0002025	0.0007467563	5353.0000	1	2	vasodilation by norepinephrine-epinephrine involved in regulation of systemic arterial blood pressure
GO:0006369	0.0093144054	222.0833	1	25	termination of RNA polymerase II transcription
GO:0007613	0.0044763536	485.7273	1	12	memory
GO:0005980	0.0044763536	485.7273	1	12	glycogen catabolic process
GO:0035811	0.0007467563	5353.0000	1	2	negative regulation of urine volume
GO:0014742	0.0007467563	5353.0000	1	2	positive regulation of muscle hypertrophy
GO:0010611	0.0011200298	2676.0000	1	3	regulation of cardiac muscle hypertrophy
GO:0003057	0.0011200298	2676.0000	1	3	regulation of the force of heart contraction by chemical signal
GO:0045986	0.0011200298	2676.0000	1	3	negative regulation of smooth muscle contraction
GO:0044058	0.0011200298	2676.0000	1	3	regulation of digestive system process
GO:0051929	0.0011200298	2676.0000	1	3	positive regulation of calcium ion transport via voltage-gated calcium channel activity
GO:0048636	0.0011200298	2676.0000	1	3	positive regulation of muscle organ development
GO:0055021	0.0014932336	1783.6667	1	4	regulation of cardiac muscle tissue growth
GO:0003084	0.0014932336	1783.6667	1	4	positive regulation of systemic arterial blood pressure
GO:0055013	0.0033582066	668.2500	1	9	cardiac muscle cell development
GO:0045823	0.0033582066	668.2500	1	9	positive regulation of heart contraction
GO:0003014	0.0044763536	485.7273	1	12	renal system process
GO:0010578	0.0041037077	534.4000	1	11	regulation of adenylate cyclase activity involved in G-protein signaling pathway
GO:0007189	0.0041037077	534.4000	1	11	activation of adenylate cyclase activity by G-protein signaling pathway
GO:0003044	0.0041037077	534.4000	1	11	regulation of systemic arterial blood pressure mediated by a chemical signal
GO:0046620	0.0037309920	593.8889	1	10	regulation of organ growth
GO:0043500	0.0037309920	593.8889	1	10	muscle adaptation
GO:0032412	0.0089426659	231.7826	1	24	regulation of ion transmembrane transporter activity

Table A.33: E7 - Active TSS flankin - lm

Intersection

No highly significant age prediction probes are associated with this state.

A.2.1.8 E8 - Active enhancer

Random Forest

rf-No highly significant age prediction probes are associated with this state.

Linear Regression

XXX

Intersection

No highly significant age prediction probes are associated with this state.

A.2.1.9 E9 - Weak enhancer

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0006200	0.0047179084	Inf	1	24	ATP catabolic process
GO:0006636	0.0043247494	Inf	1	22	unsaturated fatty acid biosynthetic process
GO:0060326	0.0090426578	Inf	1	46	cell chemotaxis
GO:0042908	0.0001965795	Inf	1	1	xenobiotic transport
GO:0015911	0.0007863181	Inf	1	4	plasma membrane long-chain fatty acid transport
GO:0046618	0.0001965795	Inf	1	1	drug export
GO:0043526	0.0017692156	Inf	1	9	neuroprotection
GO:0033700	0.0007863181	Inf	1	4	phospholipid efflux
GO:0043449	0.0021623747	Inf	1	11	cellular alkene metabolic process
GO:0019370	0.0019657952	Inf	1	10	leukotriene biosynthetic process
GO:0015908	0.0035384313	Inf	1	18	fatty acid transport
GO:0006692	0.0031452723	Inf	1	16	prostanoid metabolic process

Table A.34: E9 - Weak enhancer - rf

rf-Metabolic processes - metabolism slows with age.

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0043414	0.0070768626	Inf	1	36	macromolecule methylation
GO:0016458	0.0064871240	Inf	1	33	gene silencing
GO:0043045	0.0001965795	Inf	1	1	DNA methylation involved in embryo development
GO:0044030	0.0001965795	Inf	1	1	regulation of DNA methylation
GO:0044027	0.0001965795	Inf	1	1	hypermethylation of CpG island
GO:0090116	0.0003931590	Inf	1	2	C-5 methylation of cytosine
GO:0043046	0.0003931590	Inf	1	2	DNA methylation involved in gamete generation
GO:0006349	0.0005897385	Inf	1	3	regulation of gene expression by genetic imprinting
GO:0006346	0.0005897385	Inf	1	3	methylation-dependent chromatin silencing
GO:0045814	0.0023589542	Inf	1	12	negative regulation of gene expression, epigenetic
GO:0006305	0.0021623747	Inf	1	11	DNA alkylation

Table A.35: E9 - Weak enhancer - lm

Intersection

No highly significant age prediction probes are associated with this state.

A.2.1.10 E10 - Weak transcribed/ active proximal

Random Forest

rf-No lumi probes were found.

Linear Regression

XXX

Intersection

XXX

No lumi probes were found.

A.2.1.11 E11 - Strong transcription

Random Forest

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0008089	0.0008548835	Inf	1	4	anterograde axon cargo transport
GO:0010970	0.0040606967	Inf	1	19	microtubule-based transport

Table A.36: E11 - Strong transcription - rf

Intersection

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0008089	0.0008548835	Inf	1	4	anterograde axon cargo transport
GO:0010970	0.0040606967	Inf	1	19	microtubule-based transport

Table A.37: E11 - Strong transcription - intersection

Linear Regression

GOBPID	Pvalue	OddsRatio	Count	Size	Term
GO:0008089	0.0008548835	Inf	1	4	anterograde axon cargo transport
GO:0010970	0.0040606967	Inf	1	19	microtubule-based transport

Table A.38: E11 - Strong transcription - lm

A.2.2 CC - Cellular Component - Random Forest

A.2.2.1 E1 - PC Repressed

"No results met the specified criteria" (R's error message when running hyperGTest).

A.2.2.2 E2 - Low signal

No highly significant age prediction probes are associated with this state.

A.2.2.3 E3 - Heterochromatin

No lumi probes were found (converting the probe id to lumi id resulted in NA).

A.2.2.4 E4 - CNV/Rep

"No results met the specified criteria" (R's error message when running hyperGTest).

A.2.2.5 E5 - Strong promoter

GOCCID	Pvalue	OddsRatio	Count	Size	Term
GO:0005882	0.003397633	Inf	1	31	intermediate filament

Table A.39: E5-Strong promoter

A.2.2.6 E6 - Poised promoter

"No results met the specified criteria" (R's error message when running hyperGTest).

A.2.2.7 E7 - Active TSS flankin

GOCCID	Pvalue	OddsRatio	Count	Size	Term
GO:0031527	0.0008366801	Inf	1	5	filopodium membrane

Table A.40: E7 - Active TSS flankin

A.2.2.8 E8 - Active enhancer

No highly significant age prediction probes are associated with this state.

A.2.2.9 E9 - Weak enhancer

"No results met the specified criteria" (R's error message when running hyperGTest).

A.2.2.10 E10 - Weak transcribed/ active proximal

No lumi probes were found.

A.2.2.11 E11 - Strong transcription

"No results met the specified criteria" (R's error message when running hyperGTest).

Appendix B

Figures

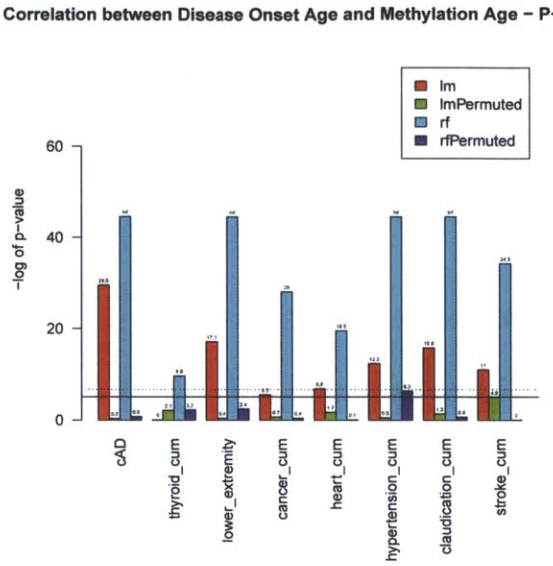


Figure B.0.1: Correlation between the age of disease onset and methylation age compared to real age - p values

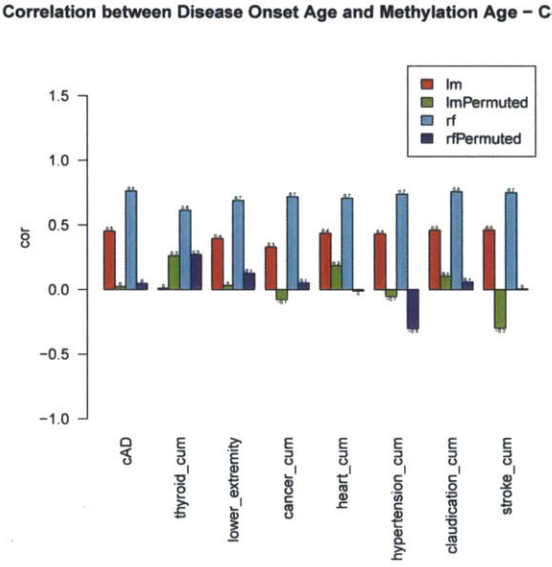
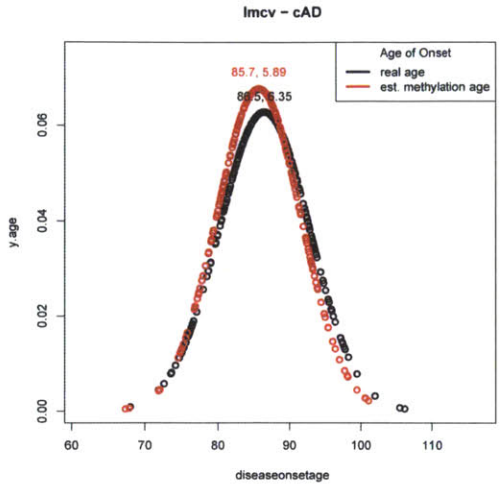


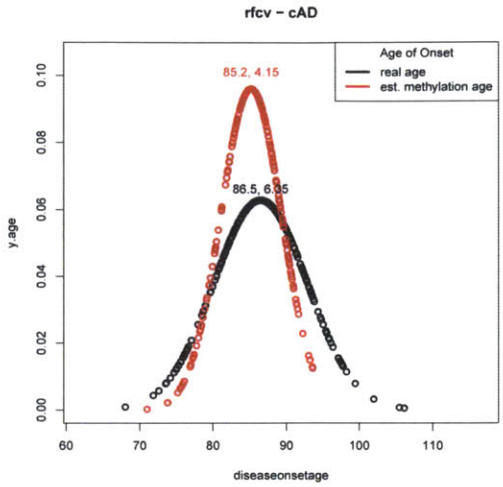
Figure B.0.2: Correlation between the age of disease onset and methylation age compared to real age - correlation values

Figure B.0.2 shows the correlations between the age of disease onset and negative age acceleration (cor(real age-methylation age, disease onset)). Figure B.0.1 shows the significance of the correlation values. The black line is the significance threshold

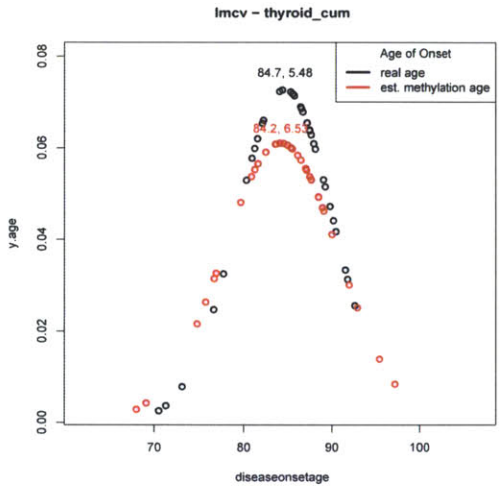
(p value = .05) and the dotted line shows a stricter significance threshold (p value = .01).



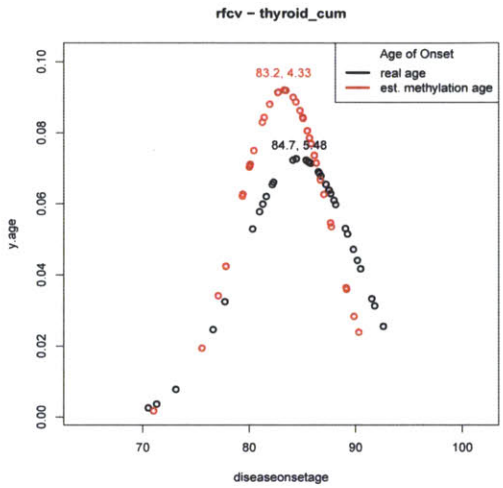
(a) lm- AD



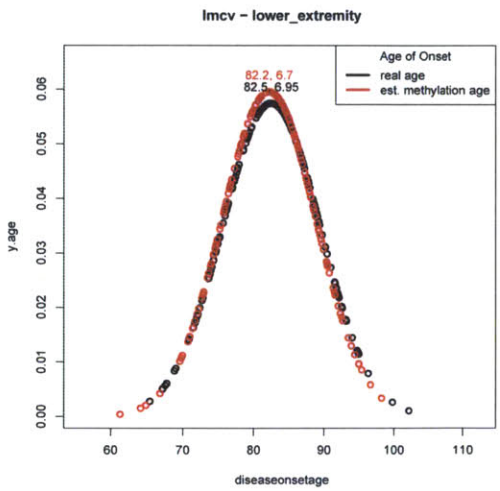
(b) rf- AD



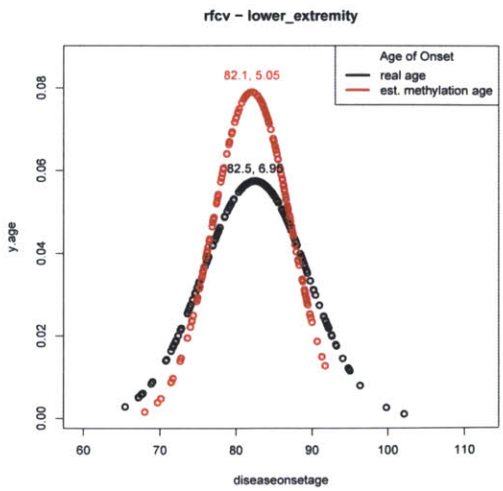
(c) lm- thyroid



(d) rf-thyroid

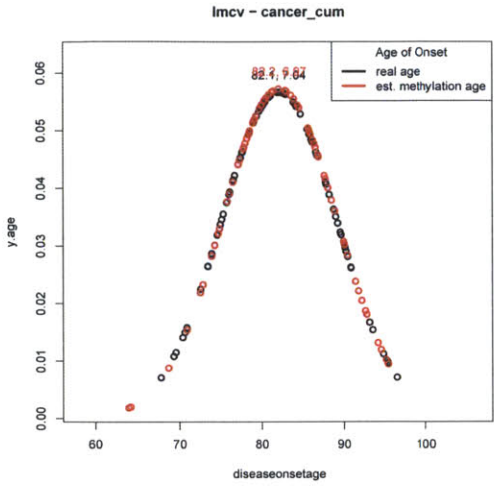


(e) lm- lower extremity pain

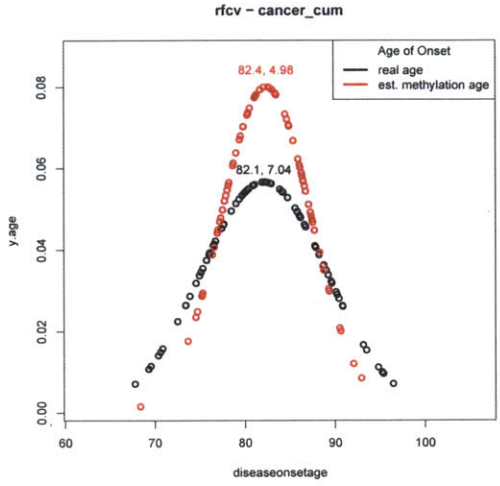


(f) rf- lower extremity pain

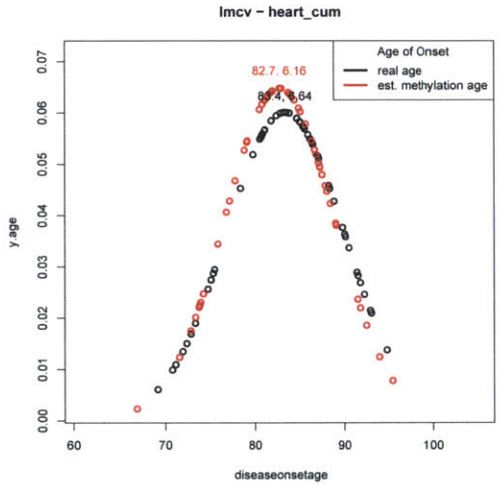
Figure B.0.3: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms



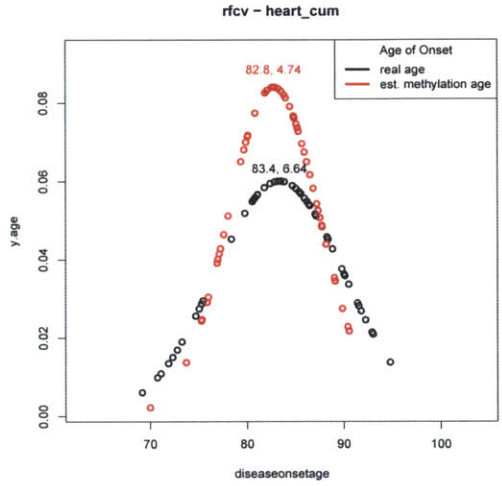
(g) lm- cancer



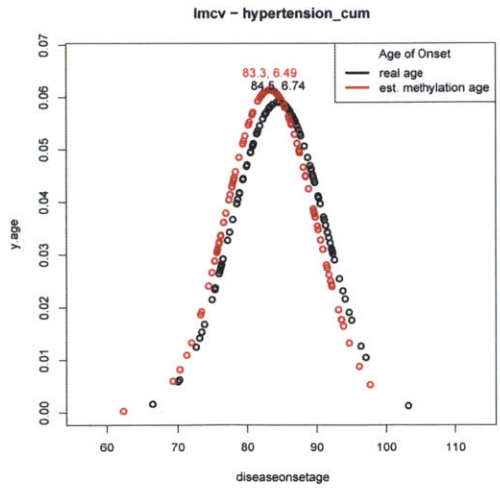
(h) rf- cancer



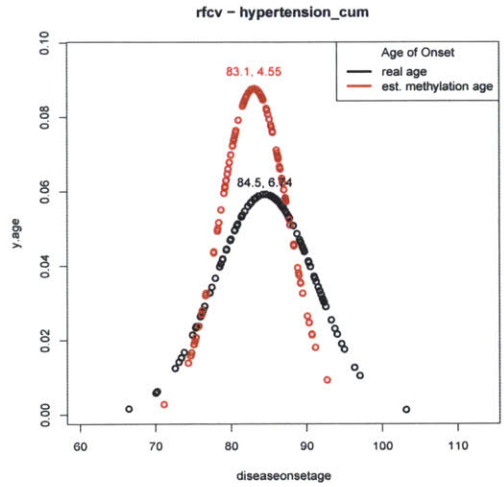
(i) lm- heart



(j) rf- heart

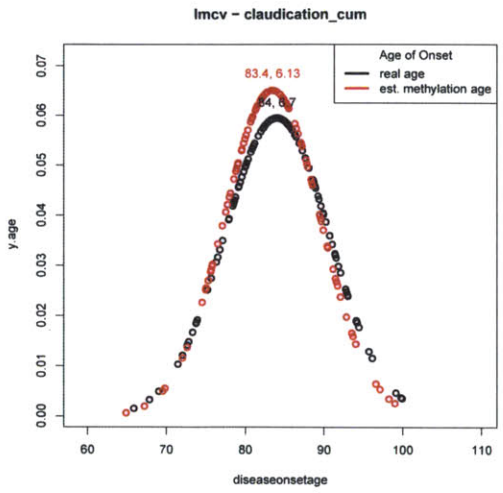


(k) lm- hypertension

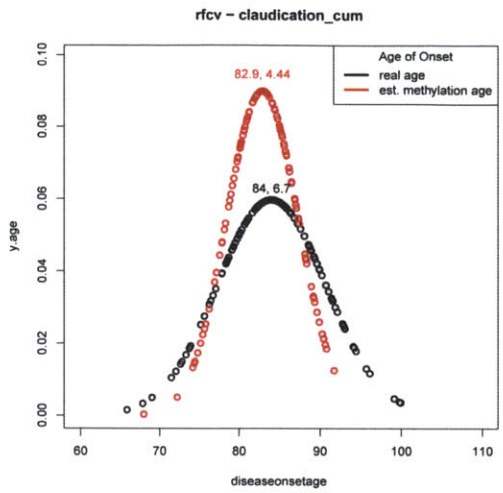


(l) rf- hypertension

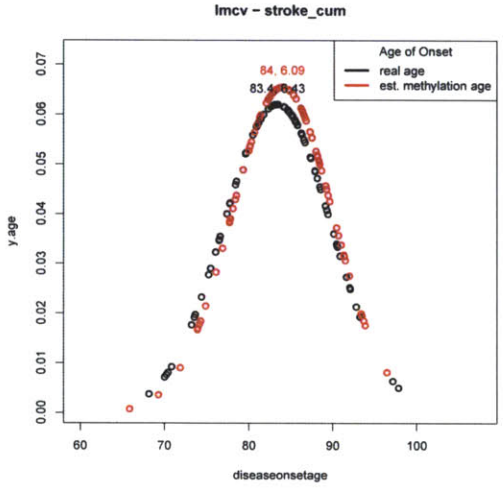
Figure B.0.3: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms



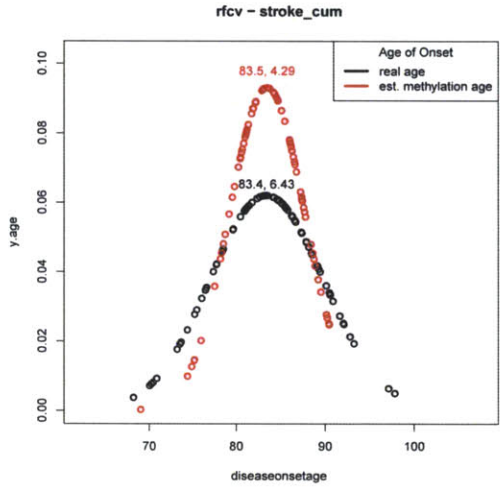
(m) lm- claudication



(n) rf- claudication

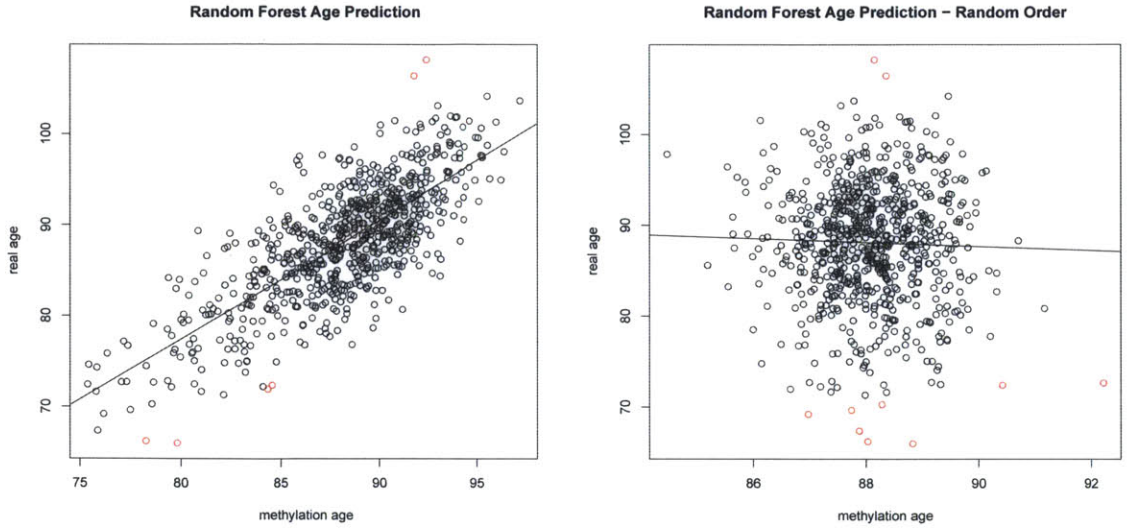


(o) lm- stroke



(p) rf- stroke

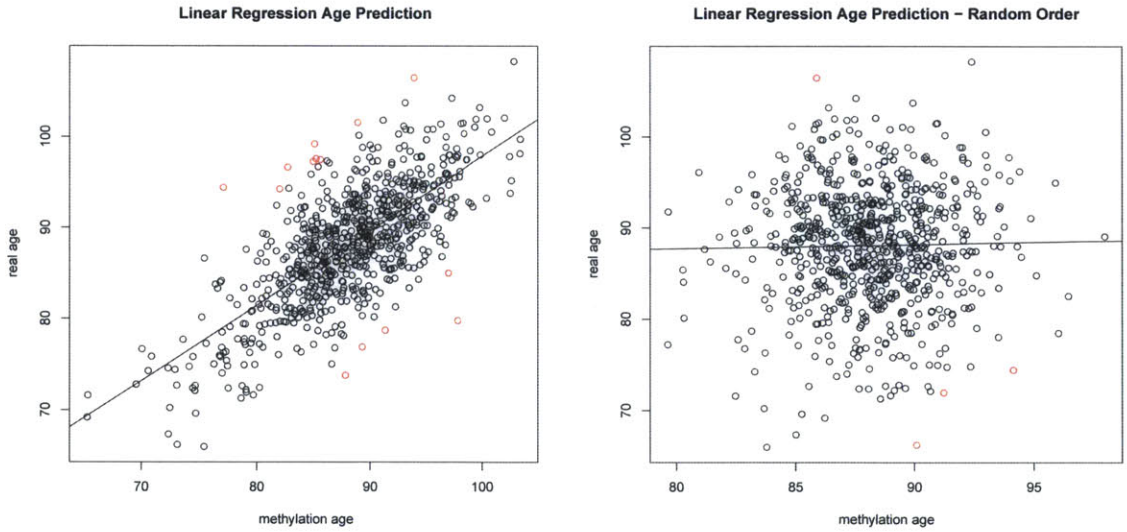
Figure B.0.3: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms



(a) Age Prediction based on Methylation using Random Forest

(b) Age Prediction based on Methylation using Random Forest - Permuted Ages

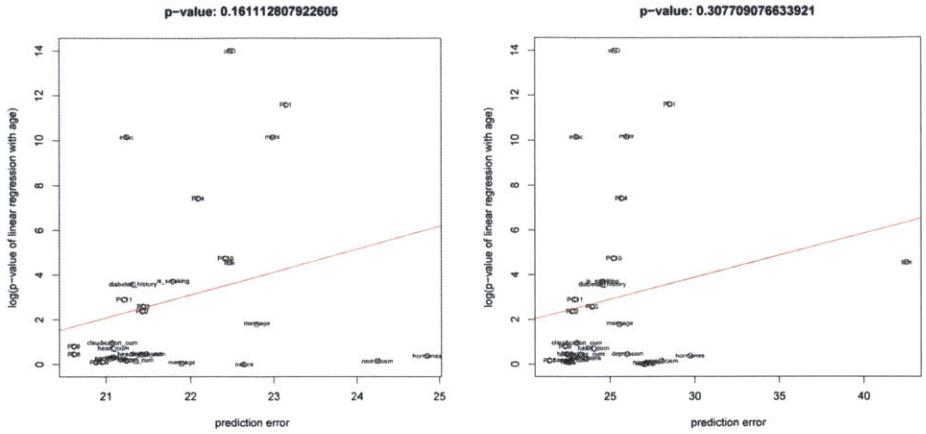
Figure B.0.4: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms



(a) Age Prediction based on Methylation using Random Forest

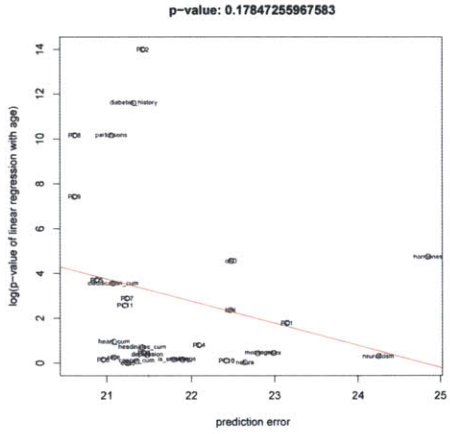
(b) Age Prediction based on Methylation using Random Forest - Permuted Ages

Figure B.0.5: Disease Onset Mean and Standard Deviation Differences Across Diseases and Prediction Algorithms

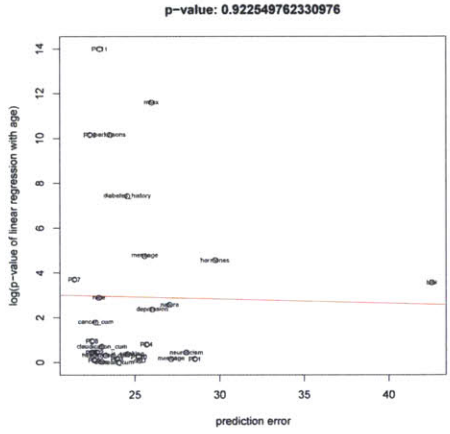


(a) Current results - rf

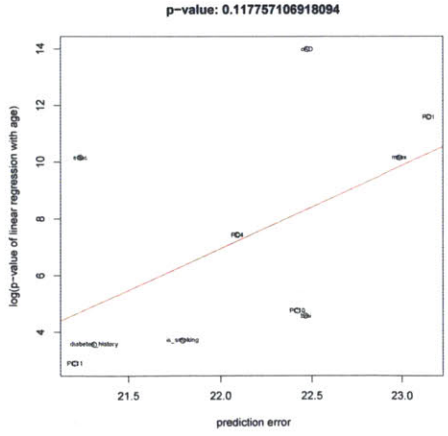
(b) Current results - lm



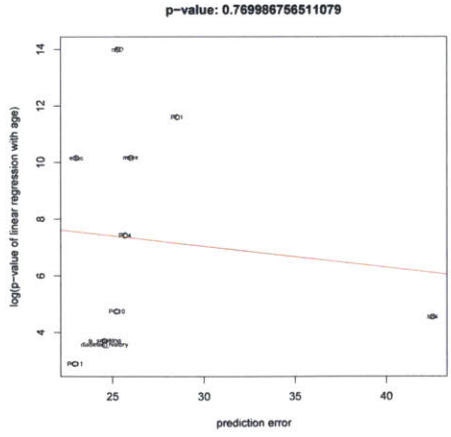
(c) Current results - rf random



(d) Current results - lm random



(e) Current results for covariates that are significantly associated with age - rf



(f) Current results for covariates that are significantly associated with age - lm

Figure B.0.6: Current Results for the Effects of Adjusting for Covariates on Age Prediction Error

Bibliography

[1] Genome-wide methylation profiles reveal quantitative views of human aging rates. URL http://www.ncbi.nlm.nih.gov.libproxy.mit.edu/geo/query/acc.cgi?acc=GSE40279. Genome wide DNA methylation profiling of individuals across a large age range. The Illumina Infinium 450k Human DNA methylation Beadchip was used to obtain DNA methylation profiles across approximately 450k CpGs from human whole blood.

[2] Herv Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010. ISSN 1939-0068. doi: 10.1002/wics.101. URL http://dx.doi.org/10.1002/wics.101.

[3] Jose Almirall. Lecture 4 - overview. URL http://www2.fiu.edu/~almirall/Lecture4.pdf.

[4] Jordana Bell, Athma Pai, Joseph Pickrell, Daniel Gaffney, Roger Pique-Regi, Jacob Degner, Yoav Gilad, and Jonathan Pritchard. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biology*, 12(1):R10, 2011. ISSN 1465-6906. doi: 10.1186/gb-2011-12-1-r10. URL http://genomebiology.com/2011/12/1/R10.

[5] Jordana T. Bell, Pei-Chien Tsai, Tsun-Po Yang, Ruth Pidsley, James Nisbet, Daniel Glass, Massimo Mangino, Guangju Zhai, Feng Zhang, Ana Valdes, So-Youn Shin, Emma L. Dempster, Robin M. Murray, Elin Grundberg, Asa K. Hedman, Alexandra Nica, Kerrin S. Small, Emmanouil T. Dermitzakis, Mark I. McCarthy, Jonathan Mill, Tim D. Spector, Panos Deloukas, and The MuTHER Consortium. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet*, 8(4):e1002629, 04 2012. doi: 10.1371/journal.pgen.1002629. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1002629.

[6] Johannes Bohacek and Isabelle M Mansuy. Epigenetic inheritance of disease and disease risk. *Neuropsychopharmacology*, 38(1):220–236, 01 2013. URL http://dx.doi.org/10.1038/npp.2012.110. taken from http://www.nature.com/npp/journal/v38/n1/fig_tab/npp2012110b1.html.

[7] Leo Breiman. Random forests. Technical report, University of California, january 2001. URL http://oz.berkeley.edu/~breiman/randomforest2001.pdf.

[8] Daniela Cihakova. *Graves' Disease.* Johns Hopkins University School of Medicine & Johns Hopkins Health System. URL http://autoimmune.pathology.jhmi.edu/diseases.cfm?systemID=3&DiseaseID=21. Autoimmune Disease Research Center.

[9] Sergio Cocozza, Giovanni Scala, Gennaro Miele, Imma Castaldo, and Antonella Monticelli. A distinct group of cpg islands shows differential dna methylation between replicas of the same cell line in vitro. *BMC Genomics*, 14(1): 692, 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-692. URL http://www.biomedcentral.com/1471-2164/14/692.

[10] John Cunha. *High Blood Pressure (cont.).* MedicineNet, 11 2012. URL http://www.onhealth.com/high_blood_pressure/page7.htm#what_causes_high_blood_pressure. Medically Reviewed by a Doctor on 4/11/2012.

[11] John Daintith. statistical methods, 2004. URL http://www.encyclopedia.com/doc/1O11-statisticalmethods.html.

[12] Matthew Eaton and Manolis Kellis. Personal genomics, disease epigenomics, systems approaches to disease. MIT 6.047/6.878/HST.507 - Computational Biology: Genomes, Networks, Evolution.

[13] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotech*, 28(8):817–825, August 2010. doi: 10.1038/nbt.1662. URL http://dx.doi.org/10.1038/nbt.1662.

[14] Agustin F. Fernandez, Yassen Assenov, Jose Ignacio Martin-Subero, Balazs Balint, Reiner Siebert, Hiroaki Taniguchi, Hiroyuki Yamamoto, Manuel Hidalgo, Aik-Choon Tan, Oliver Galm, Isidre Ferrer, Montse Sanchez-Cespedes, Alberto Villanueva, Javier Carmona, Jose V. Sanchez-Mut, Maria Berdasco, Victor Moreno, Gabriel Capella, David Monk, Esteban Ballestar, Santiago Ropero, Ramon Martinez, Marta Sanchez-Carbayo, Felipe Prosper, Xabier Agirre, Mario F. Fraga, Osvaldo Graa, Luis Perez-Jurado, Jaume Mora, Susana Puig, Jaime Prat, Lina Badimon, Annibale A. Puca, Stephen J. Meltzer, Thomas Lengauer, John Bridgewater, Christoph Bock, and Manel Esteller. A dna methylation fingerprint of 1628 human samples. *Genome Research*, 22(2):407–419, 2012. doi: 10.1101/gr.119867.110. URL http://genome.cshlp.org/content/22/2/407.abstract.

[15] Kaylee Finn. How to calculate outliers. URL http://www.ehow.com/how_5201412_calculate-outliers.html.

[16] Burt Gerstman. *StatPrimer.* San Jose State University, version 6.4 edition. URL http://www.sjsu.edu/faculty/gerstman/StatPrimer/anova-a.pdf. Chapter 12: Analysis of Variance.

[17] ScottF. Gilbert. Ageing and cancer as diseases of epigenesis. *Journal of Biosciences*, 34(4):601–604, 2009. ISSN 0250-5991. doi: 10.1007/s12038-009-0077-4. URL http://dx.doi.org/10.1007/s12038-009-0077-4.

[18] Aaron D. Goldberg, C. David Allis, and Emily Bernstein. Epigenetics: A landscape takes shape. *Cell*, 128(4):635–638, 02 2007. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867407001869.

[19] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359 – 367, 2013. ISSN 1097-2765. doi: http://dx.doi.org/10.1016/j.molcel.2012.10.016. URL http://www.sciencedirect.com/science/article/pii/S1097276512008933.

[20] Holger Heyn, Sebastian Moran, Irene Hernando-Herraez, Sergi Sayols, Antonio Gomez, Juan Sandoval, Dave Monk, Kenichiro Hata, Tomas Marques-Bonet, Liewei Wang, and Manel Esteller. Dna methylation contributes to natural human variation. *Genome Research*, 23(9):1363–1372, 2013. doi: 10.1101/gr.154187.112. URL http://genome.cshlp.org/content/23/9/1363.abstract.

[21] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013. ISSN 1465-6906. doi: 10.1186/gb-2013-14-10-r115. URL http://genomebiology.com/2013/14/10/R115.

[22] Robert S. Illingworth, Ulrike Gruenewald-Schneider, Shaun Webb, Alastair R. W. Kerr, Keith D. James, Daniel J. Turner, Colin Smith, David J. Harrison, Robert Andrews, and Adrian P. Bird. Orphan cpg islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet*, 6(9):e1001134, 09 2010. doi: 10.1371/journal.pgen.1001134. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1001134.

[23] Alireza Kashani, ve Lepicard, Odile Poirel, Catherine Videau, Jean Philippe David, Catherine Fallet-Bianco, Axelle Simon, Andr Delacourte, Bruno Giros, Jacques Epelbaum, Catalina Betancur, and Salah El Mestikawy. Loss of {VGLUT1} and {VGLUT2} in the prefrontal cortex is correlated with cognitive decline in alzheimer disease. *Neurobiology of Aging*, 29(11):1619 – 1630, 2008. ISSN 0197-4580. doi: http://dx.doi.org/10.1016/j.neurobiolaging.2007.04.010. URL http://www.sciencedirect.com/science/article/pii/S0197458007001741.

[24] Daniel Katzman, Jessica Moreno, Jason Noelanders, and Mark Winston-Galant. Comparisons of two means. 2007. URL https://controls.engin.umich.edu/wiki/index.php/Comparisons_of_two_means.

[25] Shinji Maegawa, George Hinkal, Hyun Soo Kim, Lanlan Shen, Li Zhang, Jiexin Zhang, Nianxiang Zhang, Shoudan Liang, Lawrence A. Donehower, and Jean-Pierre J. Issa. Widespread and tissue specific age-related dna methylation changes in mice. *Genome Research*, 20(3):332–340, 2010. doi: 10.1101/gr.096826.109. URL http://genome.cshlp.org/content/20/3/332.abstract.

[26] MedlinePlus. *Graves disease*. MedlinePlus. URL http://www.nlm.nih.gov/medlineplus/ency/article/000358.htm. Updated by: Brent Wisse, MD, Associate Professor of Medicine, Division of Metabolism, Endocrinology & Nutrition, University of Washington School of Medicine. Also reviewed by A.D.A.M. Health Solutions, Ebix, Inc., Editorial Team: David Zieve, MD, MHA, Bethanne Black, Stephanie Slon, and Nissi Wang.

[27] Albert Montillo. Random forests. 2009. URL http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf.

[28] Kris Montis and Timothy Peil. First quartile and third quartile. URL http://www.unc.edu/~rls/s151-09/class4.pdf.

[29] nature.com. chromatin, . URL http://www.nature.com/scitable/definition/chromatin-182.

[30] nature.com. methylation, . URL http://www.nature.com/scitable/definition/methylation-95. 2008 by Sinauer Associates, Inc. All rights reserved. Sadava, D. Life: the science of biology. 8th Edition.

[31] NIH. What are single nucleotide polymorphisms (snps)? URL http://ghr.nlm.nih.gov/handbook/genomicresearch/snp.

[32] R Nowak. Lecture 2: Introduction to classication and regression. 2007. URL http://nowak.ece.wisc.edu/SLT07/lecture2.pdf.

[33] The University of Akron. Principal component analysis. URL http://ull.chemistry.uakron.edu/chemometrics/11-A_PCA.pdf.

[34] The University of North Carolina. Iqr rule for outliers. URL http://www.unc.edu/~rls/s151-09/class4.pdf.

[35] A Dictionary of Nursing. covariate. URL http://www.encyclopedia.com/doc/1062-covariate.html.

[36] Theresa Phillips. The role of methylation in gene expression, 2008. URL http://www.nature.com/scitable/topicpage/the-role-of-methylation-in-gene-expression-1070. Nature Education 1(1):116.

[37] Princeton. linear regression. URL http://wordnetweb.princeton.edu/perl/webwn?s=linear+regression&sub=Search+WordNet.

[38] Jane Qiu. Epigenetics: Unfinished symphony. *Nature*, 441(7090):143 – 145, 2006. ISSN 0028-0836. doi: doi:10.1038/441143a. URL http://dx.doi.org/10.1038/441143a.

[39] Vardhman K. Rakyan, Thomas A. Down, David J. Balding, and Stephan Beck. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*, 12(8):529 – 541, 2011. ISSN 1471-0056. doi: 10.1038/nrg3000. URL http://dx.doi.org/10.1038/nrg3000.

[40] Markus Ringner. What is principal component analysis? *Nat Biotech*, 26(3): 303–304, March 2008. doi: 10.1038/nbt0308-303. URL http://dx.doi.org/10.1038/nbt0308-303.

[41] Jeff Schneider. Cross validation. 1997. URL http://www.cs.cmu.edu/~schneide/tut5/node42.html.

[42] Rachel Sealfon and Melissa Gymrek. Recitation 6: Random forests and affinity propagation. 6.047/6.878 Fall 2012, October 2012. URL http://stellar.mit.edu/S/course/6/fa12/6.047/courseMaterial/topics/topic4/lectureNotes/recitation6/recitation6.pdf.

[43] Alan Sykes. An introduction to regression analysis. Technical report, University of Chicago, December 1992. URL http://www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf.

[44] TheFreeDictionary. covariate. URL http://medical-dictionary.thefreedictionary.com/covariate.

[45] WebMD. *Hypothyroidism.* WebMD, July 2010. URL http://www.webmd.com/a-to-z-guides/hypothyroidism-topic-overview. WebMD Medical Reference from Healthwise.

[46] Gene Ontology website. An introduction to the gene ontology. URL http://www.geneontology.org/GO.doc.shtml.

[47] Xuan Zhou, Zhanchao Li, Zong Dai, and Xiaoyong Zou. Prediction of methylation cpgs and their methylation degrees in human dna sequences. *Computers in biology and medicine*, 42(4):408–413, 04 2012. URL http://linkinghub.elsevier.com/retrieve/pii/S0010482511002484?showall=true.