Georgia State University
ScholarWorks @ Georgia State University

Public Health Theses | School of Public Health

Spring 5-13-2016

Modeling Non-Linear Relationships Between DNA Methylation And Age: The Application of Regularization Methods To Predict Human Age And The Implication Of DNA Methylation In Immunosenescence

Nicholas Johnson

Follow this and additional works at: http://scholarworks.gsu.edu/iph_theses

Recommended Citation
Johnson, Nicholas, "Modeling Non-Linear Relationships Between DNA Methylation And Age: The Application of Regularization Methods To Predict Human Age And The Implication Of DNA Methylation In Immunosenescence." Thesis, Georgia State University, 2016.
http://scholarworks.gsu.edu/iph_theses/473

This Thesis is brought to you for free and open access by the School of Public Health at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Public Health Theses by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

ABSTRACT

Modeling Non-Linear Relationships Between DNA Methylation And Age: The Application of Regularization Methods To Predict Human Age And The Implication Of DNA Methylation In Immunosenescence

By

Nicholas David Johnson

May 4, 2016

Background: Gene expression is regulated via highly coordinated epigenetic changes, the most studied of which is DNA methylation (DNAm). Many studies have shown that DNAm is linearly associated with age, and some have even used DNAm data to build predictive models of human age, which are immensely important considering that DNAm can predict health outcomes, such as all-cause mortality, better than chronological age. Nevertheless, few studies have investigated non-linear relationships between DNAm and age, which could potentially improve these predictive models. While such investigations are relevant to predicting health outcomes, non-linear relationships between DNAm and age can also add to our understanding of biological responses to late-life events, such as diseases that afflict the elderly.

Objectives: We aim to (1) examine non-linear relationships between DNAm and age at specific loci on the genome and (2) build upon regularization methods by comparing prediction errors between models with both non-transformed and square-root transformed predictors to models that include only non-transformed predictors. We used both the sparse partial least squares (SPLS) regression model and the lasso regression model to make our comparisons.

Results: We found two age-differentially methylated sites implicated in the regulation of a gene known as KLF14, which could be involved in an immunosenescent phenotype. Inclusion of the square-root transformed variables had little effect on the prediction error of the SPLS model. On the other hand, the prediction error increased substantially in the lasso regression model, particularly when few predictors (<30) were included in the model and when many predictors (>70) were included.

Conclusion: The growing amount and complexity of biological data coupled with advances in computational technology are indispensable to our understanding of biological pathways and perplexing biological phenomena. Moreover, high-dimensional biological data have enormous implications for clinical practice. Our findings implicate a possible biological pathway involved in immunosenescence. While we were unable to improve the predictive models of human age, future research should investigate other possible non-linear relationships between DNAm and human age, considering that such statistical methods can improve predictions of health outcomes.

Modeling Non-Linear Relationships Between DNA Methylation And Age: The Application of Regularization Methods To Predict Human Age And The Implication Of DNA Methylation In Immunosenescence.

by

Nicholas David Johnson

B.A., LAWRENCE UNIVERSITY

A Thesis Submitted to the Graduate Faculty
of Georgia State University in Partial Fulfillment
of the
Requirements for the Degree

MASTER OF PUBLIC HEALTH

ATLANTA, GEORGIA
30303

APPROVAL PAGE

Modeling Non-Linear Relationships Between DNA Methylation And Age: The Application of Regularization Methods To Predict Human Age And The Implication Of DNA Methylation In Immunosenescence.

by

Nicholas David Johnson

Approved:

Ruiyan Luo
Committee Chair

Karen Conneely
Committee Member

April 25, 2016
Date

Acknowledgements

I would like to thank Dr. Ruiyan Luo and Dr. Karen Conneely for their patience, suggestions, edits, and overall guidance as I worked on my research and writing for this thesis. Moreover, the graduate students in Dr. Karen Conneely's lab, namely, Chloe Robins, Crystal Grant, and Liz Kennedy were immensely kind, helpful and patient as I troubleshooted R programming issues and picked their brains to gain a deeper understanding of recent research in genetics and epigenetics. Additionally, I would like to thank Dr. Katherine Masyn for her phenomenal instruction in applied linear regression and her willingness to meet and share her statistical advice. My brother Zack Johnson was also an excellent brainstormer and biologist to bounce ideas off of pertaining to my research. Lastly, I would like to thank Dr. Matt Hayat for being a great instructor and phenomenal mentor, who gave excellent research and career advice as I advanced through the Master of Public Health program.

Author's Statement Page

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Georgia State University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote from, to copy from, or to publish this thesis may be granted by the author or, in his/her absence, by the professor under whose direction it was written, or in his/her absence, by the Associate Dean, School of Public Health. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without written permission of the author.

Nicholas David Johnson
Signature of Author

TABLE OF CONTENTS

Chapter 1. The Implication Of DNA Methylation In Immunosenescence

1.1 Tables & Figures

Gene series #	Study	Abbreviation	Tissue type	Publication
N/A	Grady Trauma Project	GTP	Peripheral blood	Barfield et al. 2014
GSE60132	TOPS Family Study	TOPS	Peripheral blood	Ali et al. 2015
GSE56581	MESA Epigenomics and Transcriptomics Study (human T cells)	MESA-T	Purified T cells	Reynolds et al. 2014
GSE56046	MESA Epigenomics and Transcriptomics Study (human monocytes)	MESA-M	Purified monocytes	Reynolds et al. 2015
GSE59065	Estonian Genome Center Investigation of Age-related epigenetics and immune system function in PBL, CD4+ and CD8+ T cells	EGC-PBL EGC-CD4 EGC-CD8	Peripheral blood leukocytes (PBL), CD4+ and CD8+ T cells	Tserel et al. 2014
GSE74193	Schizophrenia-related DNAm and gene expression	CTX	Dorsolateral prefrontal cortex	Jaffe et al. 2016

Table 1. Reference information regarding each of the four datasets analyzed.

Dataset	*N*	β_{age}	SE_{age}	T_{age}	P_{age}
GTP	336	9.6E-4	6.4E-5	14	2.0E-16
TOPS	192	1.2E-3	1.0E-4	12	2.0E-16
MESA-T	214	1.1E-3	1.4E-4	7.6	1.0E-12
MESA-M	1202	1.6E-3	9.7E-5	16	2.0E-16
EGC-PBL	97	2.9E-3	2.2E-3	1.2	0.20
EGC-CD4	99	6.8E-4	1.7E-3	0.40	0.69
EGC-CD8	100	-6.9E-4	2.6E-3	-0.26	0.79
CTX	346	1.4E-3	2.8E-4	4.9	1.4E-6

Table 2. Statistics corresponding to the age term of the regression model fitted separately to each of the six data sets for the CpG site, cg07955995. The first two columns include abbreviations of the dataset, according to Table 1 and the sample size. Columns 3 and 4 include the slope coefficient estimate of age and corresponding standard error, and columns 5 and 6 include the T-statistic for age and corresponding p-value. The regression model included covariates and excluded the age-quadratic term.

Dataset	*N*	β_{age}	SE_{age}	T_{age}	P_{age}
GTP	336	4.9E-4	5.0E-5	9.8	2.0E-16
TOPS	192	6.3E-4	6.5E-5	9.7	2.0E-16
MESA-T	214	7.4E-4	1.1E-4	6.7	1.8E-10
MESA-M	1202	9.3E-4	6.4E-5	15	2.0E-16
EGC-PBL	97	4.7E-5	1.3E-3	0.035	0.97
EGC-CD4	99	8.0E-4	2.0E-3	0.40	0.69
EGC-CD8	100	-2.3E-3	1.8E-3	-1.3	0.21
CTX	346	9.0E-4	2.8E-4	3.2	0.0013

Table 3. Statistics corresponding to the age term corresponding to the regression model fitted separately to each of the four data sets for the CpG site, cg22285878. The type of information included in each column is the same as described in Table 2. The regression model included covariates and excluded the age-quadratic term.

Dataset	*N*	β_{age^2}	SE_{age^2}	T_{age^2}	P_{age^2}	Var_{young}	Var_{old}	P_{Var}
GTP	336	3.6E-5	4.1E-6	8.8	2.0E-16	8.3e-5	2.7e-3	4.0e-17
TOPS	192	3.4E-5	3.9 E-6	8.6	4.3E-15	6.5e-5	3.0e-3	1.5e-48
MESA-T	214	1.3E-5	1.6E-5	0.84	0.40	1.9e-4	4.2e-4	4.1e-4
MESA-M	1202	2.2E-5	1.0E-5	2.2	0.025	4.6e-4	1.7e-3	5.3e-6
EGC-PBL	97	1.6 E-5	2.1 E-5	-0.74	0.46	1.8e-4	1.4e-3	1.7e-11
EGC-CD4	99	2.9 E-6	1.6 E-5	0.18	0.86	2.9e-4	6.6e-4	2.3e-3
EGC-CD8	100	2.0E-5	2.5 E-5	0.79	0.43	3.9e-4	1.9e-3	5.9e-8
CTX	346	2.1 E-6	3.5 E-6	0.61	0.54	4.3e-4	1.2e-3	8.7e-5

Table 4. Statistics corresponding to the age-quadratic term of the regression model fitted separately to each of the four data sets for the CpG site, cg07955995. The regression model included the age term and covariates. The first two columns include abbreviations of the dataset, according to Table 1 and the sample size. Columns 3 and 4 include the slope coefficient estimate of age² and corresponding standard error, and columns 5 and 6 include the T-statistic for age and corresponding p-value. Columns 7 and 8 include the variances of the young age group and variance of the old age group for each of the 8 datasets. Column 9 is the p-value corresponding to the F-statistic calculated as the ratio of variances of the young and old age groups.

Dataset	*N*	β_{age^2}	SE_{age^2}	T_{age^2}	P_{age^2}	Var_{young}	Var_{old}	P_{Var}
GTP	336	2.1E-5	6.8E-3	6.2	1.4E-9	6.6e-5	2.0e-3	3.0e-7
TOPS	192	1.5E-5	2.7E-6	5.7	5.0E-8	6.6e-5	6.0e-4	9.9e-8
MESA-T	214	4.8E-6	1.2E-5	0.39	0.70	1.4e-4	2.1e-4	9.9e-3
MESA-M	1202	1.8E-5	6.5E-6	2.7	6.6E-3	1.8e-4	7.3e-4	2.3e-62
EGC-PBL	97	5.5E-6	1.3E-5	0.43	0.67	1.5e-4	4.3e-4	1.7e-4
EGC-CD4	99	-5.4E-7	1.9E-5	-0.028	0.97	1.9e-4	1.2e-3	1.2e-9
EGC-CD8	100	3.0E-5	1.7E-5	1.73	0.087	3.6e-4	8.0e-4	3.3e-3
CTX	346	3.0E-6	3.5E-6	0.87	0.38	4.5e-4	1.4e-3	1.6e-3

Table 5. Statistics corresponding to the age-quadratic term corresponding to the regression model fitted separately to each of the four data sets for the CpG site, cg22285878. The type of information included in each column is the same as described in Table 4. The regression model included the age term and covariates. Columns denote the same information

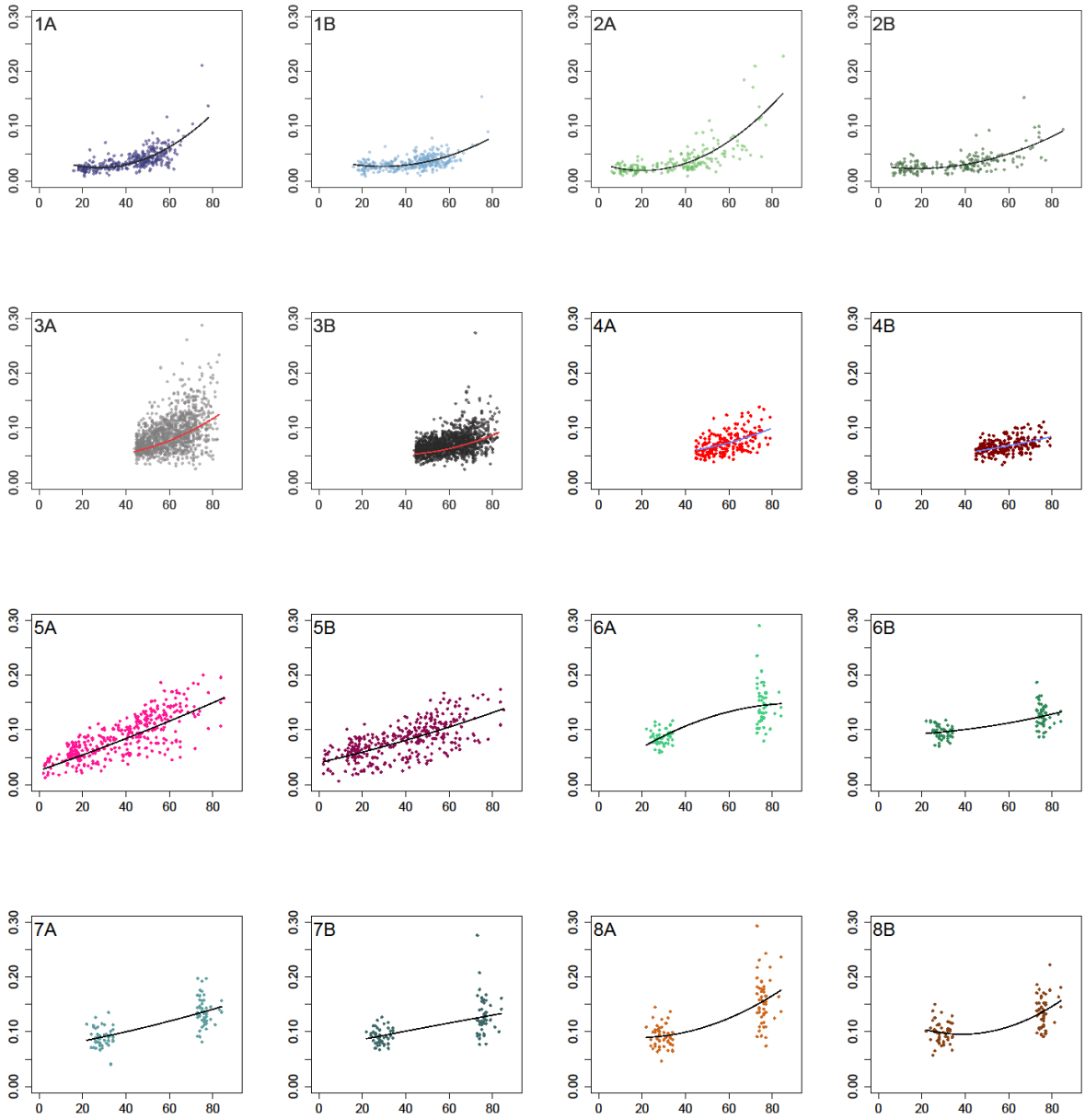


Figure 1. Fitted regression lines with age-quadratic term for each of the 8 datasets, holding covariates constant at their mean values. The alphanumeric characters in the top left corner of each graph indicates the dataset and CpG site. The letter A refers to cg07955995 and the letter B refers to cg22285878. The number 1 corresponds to GTP, 2 to TOPS, 3 to MESA-M, 4 to MESA-T, 5 to CTX, 6 to EGC-PBL, 7 to EGC-CD4, and 8 to EGC-CD8. Colors of regression lines were chosen to clearly contrast the datapoints.

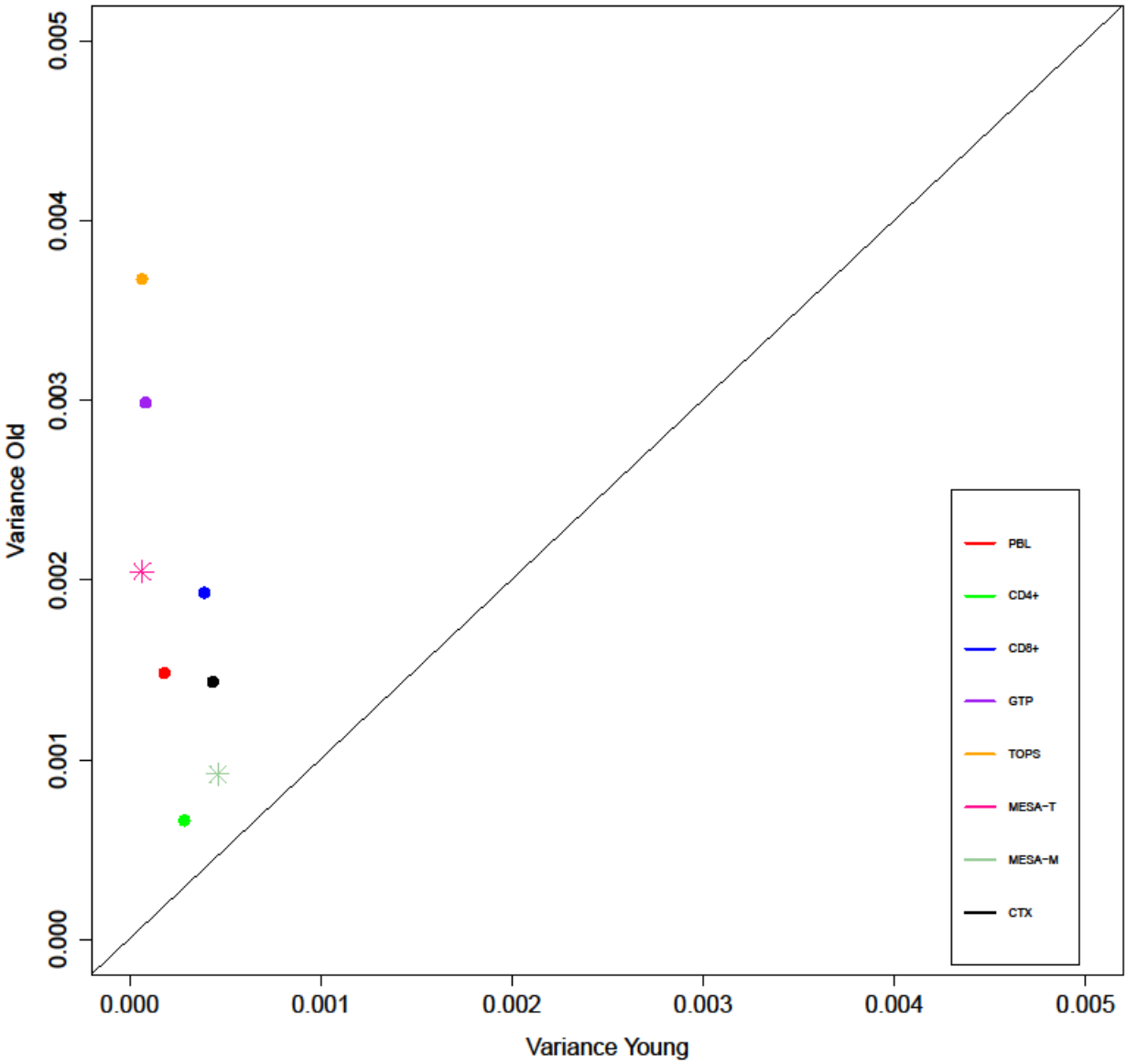


Figure 2. Plot of the variance of methylation at cg07955995 for the old age group (73 years or older) against the young age group for each dataset. MESA-M and MESA-T appear with different symbols because subjects in these datasets were too old to include an age group less than 34 years of age. For these two datasets the median ages (in years) were used (58 for MESA-M and 60 for MESA-T) to create young groups (≤median) and old groups (>median).

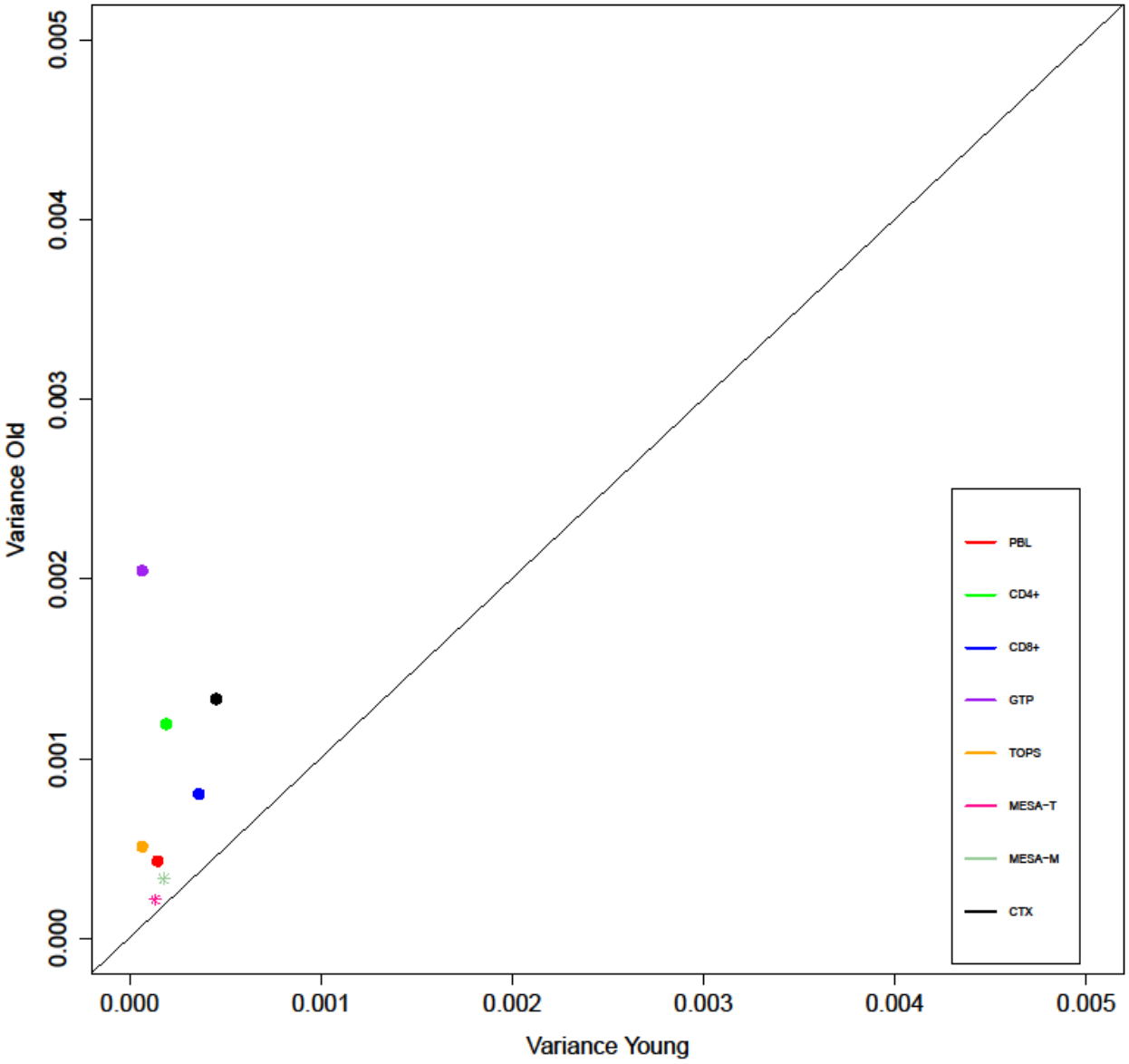
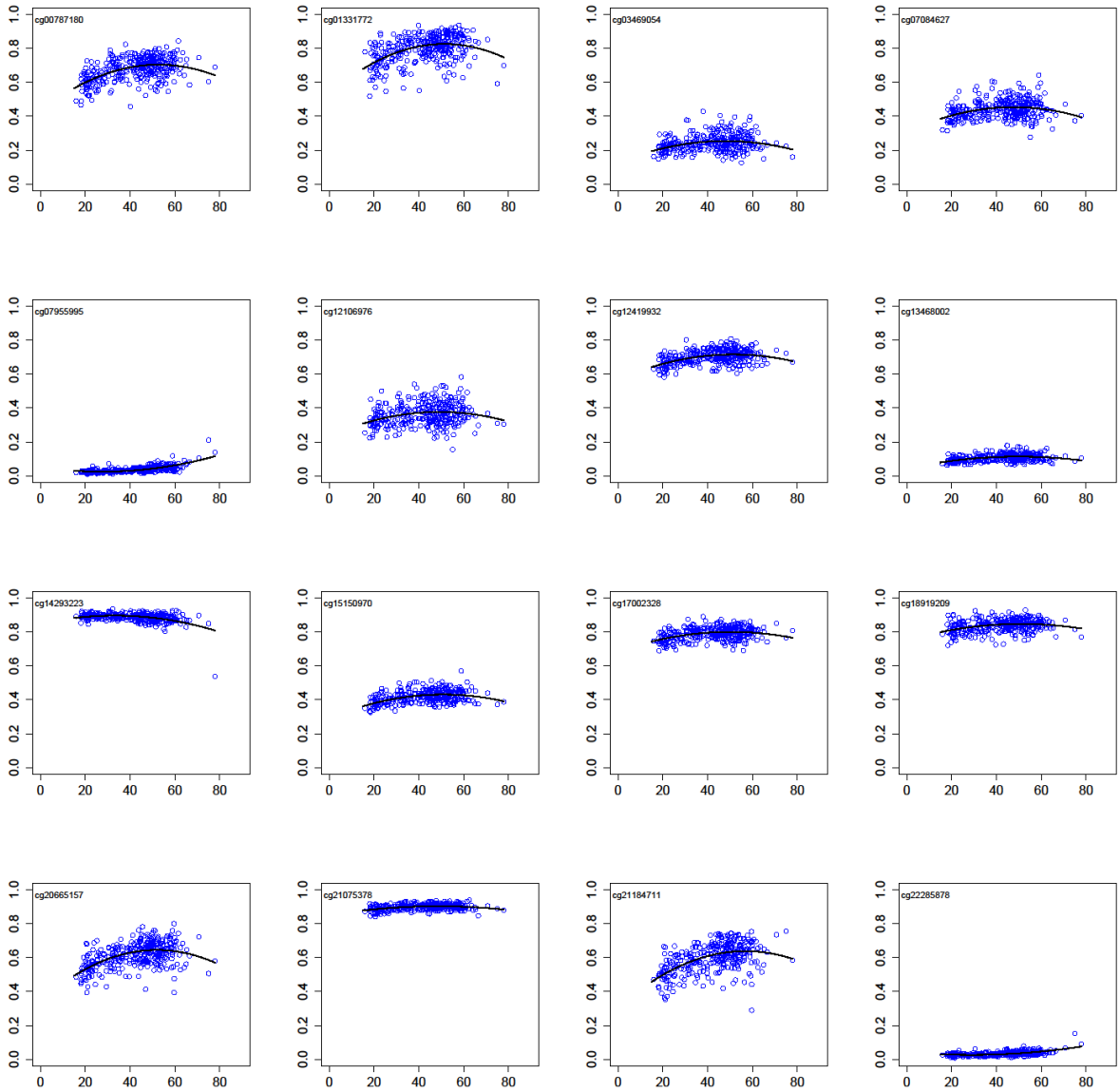


Figure 3. Plot of the variance of methylation at cg22285878 for the old age group (73 years or older) against the young age group (34 years of age or younger) for each dataset. MESA-M and MESA-T appear with different symbols because subjects in these data sets were too old to include an age group less than 34 years of age. For these two datasets the median ages (in years) were used (58 for MESA-M and 60 for MESA-T) to create young groups (≤median) and old groups (>median).

Supplemental Tables & Figures:



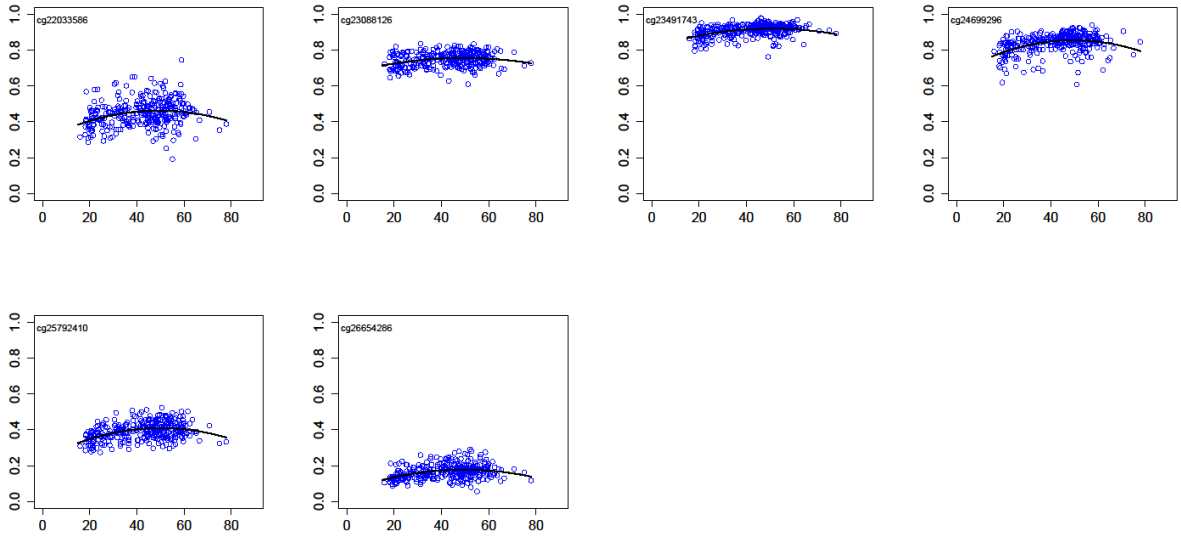


Figure S1. Beta values plotted against age for the 22 CpG sites with a Holm-significant quadratic term (p<5.038799e-05) in the primary analysis. Regression lines correspond to a model with an age term and age² term, holding covariates constant at their mean values.

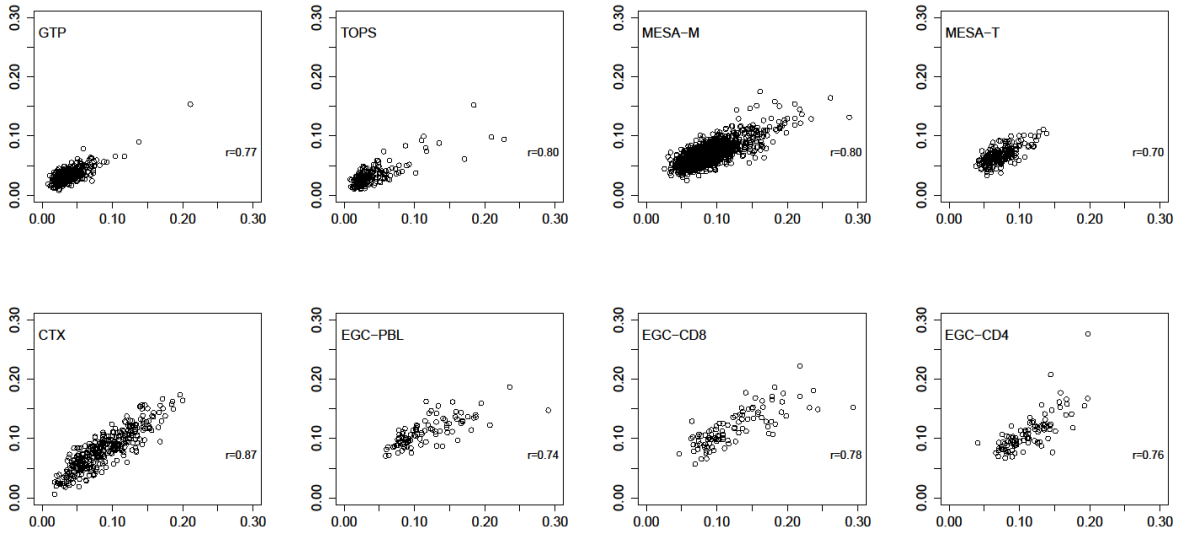


Figure S2. Beta values of cg22285878 plotted against beta values of cg07955995 for each of the eight datasets (top left corner) with Pearson's r correlation (bottom right corner).

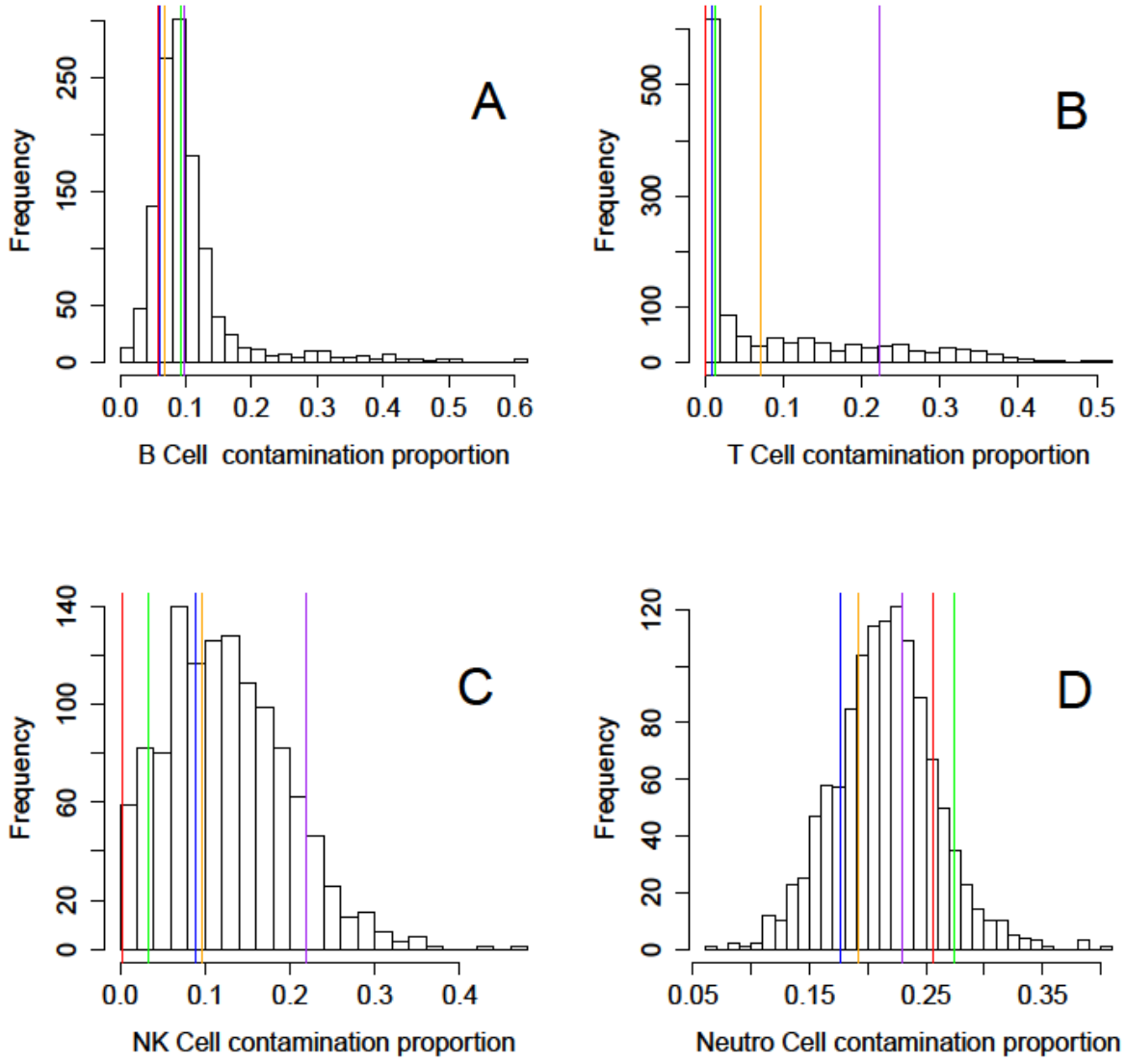


Figure S3. Histograms of cell type contamination proportions across MESA-M individuals. Colored lines indicate the contamination proportions among the five individuals with the highest β-values at cg07955995.

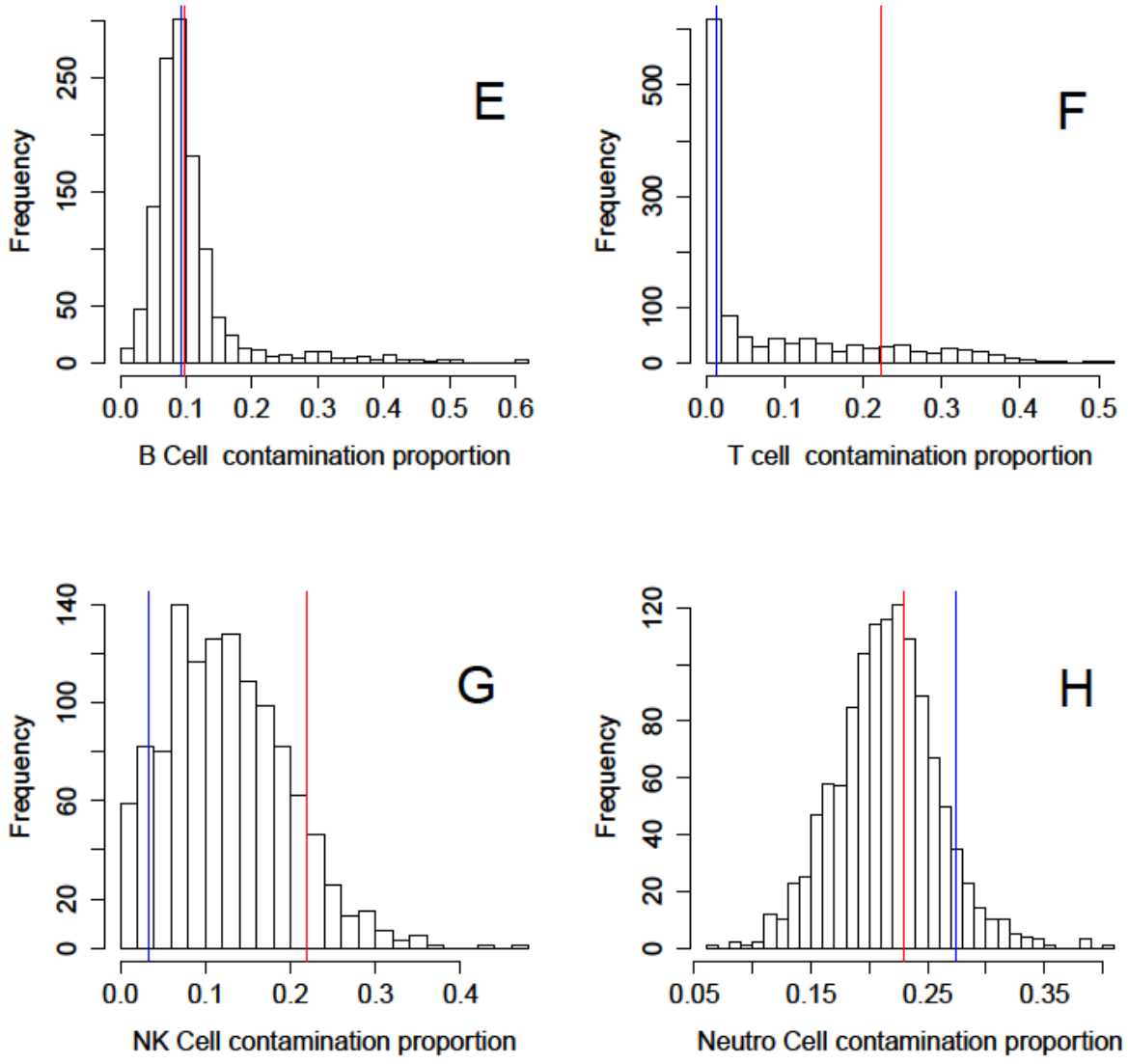


Figure S4. Histograms of cell type contamination proportions across MESA-M individuals. Colored lines indicate the contamination proportions among the five individuals with the highest β-values at cg22285878.

	β_{age}	SE_{age}	T_{age}	P_{age}	β_{age^2}	SE_{age^2}	T_{age^2}	P_{age^2}
cg00787180	9.7E-03	1.4E-03	7.15	5.8E-12	-9.2E-05	1.6E-05	-5.66	3.3E-08
cg01331772	1.1E-02	1.8E-03	6.22	1.5E-09	-1.1E-04	2.2E-05	-5.04	7.7E-07
cg03469054	4.8E-03	8.2E-04	5.84	1.2E-08	-4.8E-05	9.8E-06	-4.92	1.4E-06
cg07084627	5.4E-03	8.7E-04	6.28	1.1E-09	-5.6E-05	1.0E-05	-5.38	1.4E-07
cg07955995	-2.0E-03	3.4E-04	-5.83	1.3E-08	3.6E-05	4.1E-06	8.76	1.1E-16
cg12106976	4.8E-03	8.1E-04	5.91	8.5E-09	-4.8E-05	9.7E-06	-4.90	1.5E-06
cg12419932	5.4E-03	8.2E-04	6.55	2.2E-10	-5.2E-05	9.9E-06	-5.30	2.2E-07
cg13468002	2.7E-03	4.4E-04	6.14	2.5E-09	-2.6E-05	5.3E-06	-5.03	8.1E-07
cg14293223	2.6E-03	6.0E-04	4.40	1.5E-05	-4.1E-05	7.2E-06	-5.65	3.6E-08
cg15150970	4.7E-03	6.9E-04	6.88	3.1E-11	-4.7E-05	8.3E-06	-5.64	3.6E-08
cg17002328	4.4E-03	6.9E-04	6.32	8.3E-10	-4.4E-05	8.3E-06	-5.31	2.0E-07
cg18919209	3.3E-03	5.3E-04	6.27	1.2E-09	-3.2E-05	6.3E-06	-5.04	7.7E-07
cg20665157	1.2E-02	1.4E-03	8.51	6.5E-16	-1.2E-04	1.7E-05	-6.76	6.4E-11
cg21075378	2.2E-03	3.4E-04	6.34	7.8E-10	-2.2E-05	4.1E-06	-5.38	1.4E-07
cg21184711	1.3E-02	1.6E-03	7.82	7.3E-14	-1.1E-04	1.9E-05	-5.71	2.5E-08
cg22033586	5.4E-03	8.6E-04	6.30	9.5E-10	-5.4E-05	1.0E-05	-5.16	4.3E-07
cg22285878	-1.2E-03	2.8E-04	-4.38	1.6E-05	2.1E-05	3.3E-06	6.23	1.4E-09
cg23088126	3.3E-03	5.5E-04	5.87	1.1E-08	-3.3E-05	6.7E-06	-4.99	9.8E-07
cg23491743	4.7E-03	7.7E-04	6.11	2.9E-09	-4.6E-05	9.3E-06	-4.94	1.2E-06
cg24699296	7.5E-03	1.2E-03	6.25	1.3E-09	-7.5E-05	1.4E-05	-5.19	3.7E-07
cg25792410	6.4E-03	1.1E-03	6.09	3.2E-09	-6.3E-05	1.3E-05	-4.99	9.6E-07
cg26654286	3.9E-03	6.4E-04	6.09	3.1E-09	-3.9E-05	7.6E-06	-5.14	4.7E-07

Table S1. Regression statistics corresponding to the 22 CpG sites with a Holm-significant quadratic term (p<5.038799e-05) from the primary analysis. Column 1 indicates the CpG probe site according to the Illumina Infinium 450k Human Methylation Array. Columns 2-5 correspond to the slope coefficient estimate, standard error, T-statistic, and P-value corresponding to the age term in the primary analysis. Columns 6-9 correspond to the slope coefficient estimate, standard error, T-statistic, and P-value corresponding to the age^2 term in the primary analysis.

1.2 Introduction:

Gene expression is regulated via highly coordinated epigenetic changes, the most studied of which is DNA methylation (DNAm). DNAm is the binding of a methyl group (-CH_3) to DNA, which, in mammals, occurs most commonly at a cytosine nucleotide that resides 5' to a guanine nucleotide, referred to as a CpG site (Mendizabal 2014). CpG sites are clustered in CpG-rich regions known as CpG islands (CGIs), regions directly adjacent to CGIs (CpG shores), and CpG-poor regions (CpG shelves). The position of the CpG site within the transcription unit (typically, comprising a promoter, the RNA coding sequence, and terminator) has a substantial effect on how DNAm impacts expression of the downstream gene, e.g. upregulation or downregulation (Jaenisch and Bird 2003). Generally, CpG-poor regions are hypermethylated whereas CpG-rich regions are hypomethylated (Day 2013). Evidence suggests that DNAm in the gene body stimulates transcription whereas DNAm of transcription start sites results in gene silencing (Jones 2012).

While patterns of DNAm across the genome vary according to tissue type and environmental exposure, many studies have shown that age explains a substantial portion of the variation in human DNAm (Bell 2012; Christensen 2009; Issa et al. 1994, 1996; Ahuja et al. 1998; Nakagawa et al. 2001; Fraga et al. 2005; So et al. 2006; Fraga and Esteller 2007; Bjornsson et al. 2008). Furthermore, monozygotic (MZ) twin studies indicate that age-related DNAm is not wholly explained by genetic factors: DNAm varies more widely between older MZ twins than younger, and these differences are less marked among MZ twins who share more time with each other and have similar lifestyles (Fraga et al., 2005).

DNAm has been observed to decrease genome-wide with age (Bollati et al. 2009), although the relationship between DNAm and age is much more nuanced. For example, CGIs are associated with an increasing rate of DNAm with age whereas non-CGIs are associated with a decreasing rate of DNAm with age (Christensen 2009). Furthermore, post-natal DNA is hypomethylated and undergoes a rapid increase in DNAm in early life before stabilizing in adulthood, followed by a gradual decrease later in life (Jones 2015). But not all CpG sites that undergo age-related DNAm follow this trend: while CGI promoters are generally hypomethylated with age, one study found age-related hypermethylated CGI promoters associated with gene silencing (Shen 2007), and another study found a class of CGIs with stably methylated shores associated with high gene expression (Edgar 2014).

While general trends in DNAm are important, deviations from these trends will provide essential insights into the complex relationship between DNAm and senescence. Heretofore, studies investigating age-related DNAm have shown a few stages in life where rapid DNAm changes occur followed by stabilization. For example, differences in the rates of DNAm have been observed between pediatric and adult subjects (Alish et al. 2012). Less studied is the reverse trend: Stable DNAm levels followed by rapid methylation/demethylation. Here, we present two analyses to better understand non-linear trends between DNAm and human aging. We analyze genome-wide DNAm data in a set of peripheral blood samples with the goal of identifying loci for which DNAm increases or decreases at an increasing rate with age. We follow up this analysis in seven

datasets comprising distinct tissue types including dorsolateral prefrontal cortex and peripheral blood, along with purified blood cell types isolated from peripheral blood including monocytes, CD4+ T cells, and CD8+ T cells.

1.3 Methods:

Samples:
A total of eight datasets were analyzed (Table 1). For each of the datasets, DNA was collected from tissue samples and was bisulphite-treated for cytosine to thymine conversion and hybridized to the Illumina Infinium 450k Human Methylation Beadchip (Ali et al. 2015; Reynolds et al. 2014; Reynolds et al. 2015; Barfield et al. 2015; Gillespie et al. 2009; Jaffe et al. 2016;).

Grady Trauma Project (GTP):
The primary analysis was performed using a subset of 336 individuals ranging in age from 16 to 78 from data collected as part of the Grady Trauma Project, a study investigating the effects of genetic and environmental factors on individuals' response to stressful life events. Participants were recruited from waiting rooms at Grady Memorial Hospital in Atlanta, GA between 2005 and 2008. Individuals who provided informed consent provided either salivary or blood samples. The Institutional Review Boards of Emory University School of Medicine and Grady Memorial Hospital approved all procedures of the Grady Trauma Project. (Gillespie et al. 2009; Barfield et al. 2014).

Follow-up analyses were performed on the publicly available datasets listed below to replicate findings of the primary analysis.

TOPS Family Study (TOPS)

The TOPS Family Study included methylation data collected from peripheral blood of 192 individuals. Individual ages ranged from 6 to 85 and each individual belonged to 1 of 7 extended families of Northern European descent. In order to be included in the study, each nuclear family was required to have two obese siblings and at least one parent or sibling who was never obese. The NCBI Gene Expression Omnibus accession number corresponding to the TOPS Family Study is GSE60132. (Ali et al. 2015)

Multi-Ethnic Study of Atherosclerosis (MESA):

The Multi-Ethnic Study of Atherosclerosis (MESA) collected methylation data for samples of $CD4^+$ T cells (MESA-CD4; 214 subjects) and monocytes (MESA-M; 1,202 subjects) isolated from peripheral blood. The age range was 45-79 for MESA-CD4 subjects and 44-83 for MESA-M subjects. The MESA study was conducted to collect population-based information on the prevalence and progression of subclinical cardiovascular disease. Subjects were recruited from six sites: Baltimore City and Baltimore County, Maryland; Chicago, Illinois; Forsyth County, North Carolina; Los Angeles County, California; New York, New York; and St. Paul, Minnesota. The NCBI Gene Expression Omnibus accession number corresponding to MESA-CD4 and MESA-M are GSE56581 and GSE56046, respectively. (Reynolds et al. 2014; Reynolds et al. 2015)

Estonian Genome Center Investigation of Age-related Epigenetics and Immune System function in PBL, CD4$^+$ and CD8$^+$ T cells (EGC-PBL, EGC-CD4, EGC-CD8):

Peripheral blood leukocytes (EGC-PBL) were collected, in addition to CD4$^+$ T cells (EGC-CD4) and CD8$^+$ T cells (EGC-CD8), which were isolated from peripheral blood from healthy donors of the Estonian Genome Center of the University of Tartu. A total of 101 subjects were in the study, and were divided into a young age group and an old age group whose ranges were 22-34 and 73-84, respectively. After quality control, a total of 296 samples were collected from the 101 subjects: 99 CD4$^+$ T cell samples, 100 CD8$^+$ T cell samples, and 97 PBL samples. The NCBI Gene Expression Omnibus accession number corresponding to the Estonian Genome Center Investigation of Age-related epigenetics and immune system function in PBL, CD4+ and CD8+ T cells is GSE59065. (Tserel et al. 2014)

Schizophrenia-related DNAm and gene expression (CTX):

Samples of dorsolateral prefrontal cortex were surgically removed from post-mortem brain samples donated to the NIMH Brain Tissue Collection at the National Institute of Health in Bethesda, Maryland. Our investigation focused on 346 samples from the control group with an age range of 2-85. The NCBI Gene Expression Omnibus accession number corresponding to the Schizophrenia-related DNAm and gene expression data is GSE74193. (Jaffe et al. 2016)

Data Analysis

The eight data sets were analyzed using the suite of R functions CpGassoc (Barfield et al. 2012). Within the CpGassoc package, the cpg.qc function was used to perform quality control on all the datasets. Quality control included the removal of probes with missing data for >5% of samples, removal of samples with >5% missing probe sites, and the removal of samples with a detection p-value > 0.001. Methylated and unmethylated signals were then quantile normalized. A beta-value (β) was then computed from the methylated signal (M) and unmethylated signal (U) as follows: $\beta = \frac{M}{U+M}$.

The GTP data set was analyzed by using the R function cpg.assoc to perform a separate linear regression for each CpG site where β was regressed on age, age^2, sex, and cell type proportions, which were imputed using the R package minfi. The quadratic term for age was included to allow identification of CpG sites demonstrating an increasing rate of change with age. CpG sites were considered significant if the p-value corresponding to the t-statistic on the age^2 term was smaller than the Holm-Bonferroni corrected level of significance (α=1.034342e-06).

For CpG sites demonstrating an increasing rate of change with age in GTP, follow-up analyses were performed on the remaining data sets. Only CpG sites with a significant quadratic term were included in the subsequent analyses, which were performed using the R function lm(). The response variable β was regressed on the predictors age and age^2 similar to the analysis performed on the first data set. Our TOPS, MESA-T, and MESA-M analyses all included a 'racegendersite' variable as a covariate in the regression models. While our TOPS regression model also included imputed cell type proportions as covariates, the MESA-T and MESA-M

models included contaminated cell type proportions, namely, proportions of B-cells, monocytes, natural killer cells, and neutrophils.

1.4 Results:

A separate linear regression was performed on each CpG site for data obtained from the GTP. Of the 483,399 CpG sites analyzed, 9,923 were significantly associated with age (P<1.03E-7). Subsequently, the 9,923 aDM CpG sites were fitted with an age term and age^2 term as independent variables. For the subsequent analysis 22 CpG sites had a Holm-significant age^2 term (P<5.04E-05). Of these 22 CpG sites (Figure 5 & Table 6), three general trends were observed:
(1) Two of the 22 CpG sites that were modeled exhibited a low level of DNAm and near-zero slope between β and age at the minimum age of the GTP age range (age_{min}=15.9 years) that positively accelerated as age increased.
(2) One CpG site exhibited a near-zero slope and high DNAm at age_{min} whose slope decreased with age. However, when the oldest individual was removed from the model, both the age term and the age^2 term were no longer significant.
(3) The remaining 19 CpG sites that were modeled exhibited a positive relationship between DNAm and age at age_{min} whose slope decreased with age until reaching zero before becoming increasingly negative among older individuals.

The two CpG sites (cg07955995 and cg22285878) that exhibited trend (1) are 14 base pairs apart from each other. The GTP analysis and subsequent analyses revealed that models with a significant quadratic relationship between β and age for cg07955995 also had a significant quadratic relationship between β and age for cg22285878, which is likely attributable to the proximity of these two CpG sites on the genome. In other words, it is plausible that both CpG sites are detecting the same signal (Table 4).

There are three SNPs within 10,000 base pairs of these two CpG sites. Because the three oldest subjects in GTP had the largest proportion of cells methylated, we sought to determine whether a SNP was driving the relationship found. When compared to the remaining 333 subjects, none of the three oldest subjects had a unique SNP genotype.

To replicate this result in independent datasets, a separate linear regression was fitted for each of the two CpG sites identified (cg07955995 and cg22285878) in seven additional datasets (Table 1). Among the additional analyses, an increasing rate of DNAm with age was most notably observed in the TOPS, similar to observations in the GTP analyses (Figure 1). The age-quadratic terms were significant for cg07955995 in TOPS and MESA-M ($p=4.33x10^{-15}$ and p=0.0247, respectively) but not in MESA-T, EGC-PBL, EGC-CD4, EGC-CD8, and CTX (p=0.402, p=0.460, p=0.857, p=0.430, p=0.544, respectively) (Table 4; Figure 1). The other CpG site (cg22285878) yielded similar results. The age-quadratic term was significant in TOPS ($p=4.95x10^{-8}$) and MESA-M (p=0.006643) but not MESA-T, EGC-PBL, EGC-CD4, EGC-CD8, and CTX (p=0.698, p=0.668, p=0.978, p=0.087, p=0.384, respectively) (Table 5).

These results provide evidence that age-related DNAm at these two CpG sites may vary by tissue type and blood cell type. In particular, MESA-M, a sample of isolated monocytes, exhibited evidence of an age-quadratic relationship with DNAm whereas MESA-T, a sample of isolated T cells, exhibited no evidence of an age-quadratic relationship with DNAm. To investigate whether cell type contamination was driving this relationship, cell type contamination proportions were compared between those individuals with the highest β values to the remaining individuals in MESA-M (Figure S3 & S4). We found no substantial difference between the contamination proportions between the two groups.

In addition to investigating quadrilinear relationships between DNAm and age, we also investigated whether the variance in DNAm differed between young and old individuals by cell and tissue type. Because the EGC data was already split into a young age group (≤34 years) and an old age group (≥73 years), we split data from GTP, TOPS, and CTX in a similar fashion. Because MESA-M and MESA-T have age ranges of 45-79 and 44-83, we could not split these data sets in the same age groups as the other data sets. For MESA-M and MESA-T, we instead used the median values (58 and 60, respectively) to split the datasets into young groups (≤median) and old groups (>median) (Figure 2 & 3).

There was a significantly greater variance in the old age group compared to the young age group for all data sets for both cg07955995 and cg22285878 (Tables 4 & 5). While Figure 2 and 3 plot the deviation of these variances from the y=x line for all data sets, EGC-CD4, EGC-CD8, and EGC-PBL are the most meaningfully comparable considering the data is derived from the same sample of subjects. The other datasets, on the other hand, were recruited from a variety of different people across the United States. When comparing EGC-CD4, EGC-CD8 for cg07955995, EGC-CD8 exhibited the highest variance in the old age group while CD4 exhibited the lowest variance. For cg22285878, CD4 exhibited the highest variance in the old age group whereas PBL exhibited the lowest variance among the old age group.

1.5 Discussion:

Many studies have investigated the relationship between DNAm and human age. While significantly different rates of DNAm have been observed between pediatric and adult populations (Alisch et al. 2012), we are unaware of studies investigating CpG sites whereby DNAm increases exponentially or decreases exponentially with age within the same cohort. Such investigations are important to unravel the complex relationship between DNAm and senescence. In addition to investigating a quadrilinear relationship between DNAm and age, we further explored age-related DNAm at these two CpG sites by comparing the variances between young and old individuals across cell types. In order to understand the importance of these results, we first review evolutionary models of human aging and discuss their relevance to quadrilinear age-related DNAm. Next, we discuss the tissue and cell type specific patterns we observe and a possible relationship with immune system function. Lastly, we consider how these effects could be implicated in gene regulation, particularly, the regulation of a gene known as KLF14.

Evolutionary Models of Human Aging:

Methylome-wide association studies provide a novel approach to testing evolutionary models of human aging. The focus of this analysis is the rate of change of DNAm over the lifespan, which provides insight into the evolution of aging. Major models of the evolution of aging include mutation accumulation, disposable soma, optimal lifespan, and the developmental model of aging. While these models are not mutually exclusive, the specific association between age and DNAm varies from one model to another. Here, we briefly outline each model, along with the type of age-related DNAm it is consistent with.

The mutation accumulation evolutionary model of senescence posits that mutations deleterious to late life survival accumulate more readily than those that affect early life survival because mutations that threaten survival prior to reproduction –before mutations have been passed on to the next generation- are less likely to pass on to the next generation than mutations that threaten post-reproductive survival (Medawar 1952). The disposable soma model of human aging proposes that there exists a tradeoff between reproduction and longevity (Kirkwood 1977). For example, a biological process that allocates more resources to early life reproduction will have less resources to spare for bodily maintenance resulting in the deterioration over time, that is, the more resources allocated to reproduction, the less resources there will be to stave off deterioration of the organism and eventual mortality. Thirdly, the developmental model of aging views aging as the prolongation into late life of biological processes beneficial in early life. Lastly, the optimal lifespan model of aging posits that there comes a certain point in the late life after which it is more likely that an organism's genes will be passed on to future generations if resources are invested into its offspring's survival and reproduction as opposed to its own survival and reproduction (Williams 1957).

The mutation accumulation model, where mutations whose consequences are deleterious later in life, is consistent with age differentially methylated sites (aDM) that manifest a change in the rate at which the proportion of cells are methylated later in life. Likewise, the optimal lifespan model is consistent with changes in the rate of DNAm, which could correspond to a biological mechanism driving a late life event, such as mortality. Conversely, the disposable soma model and developmental model of aging are likely to exhibit a similar or decreasing rate of proportion of cells methylated over the course of the lifespan. The findings of this study, whereby two age-differentially methylated CpG sites exhibited a quadrilinear increase in DNAm with age, are potentially consistent with the optimal lifespan and mutation accumulation models of aging. While this study cannot distinguish between the two models of aging, further investigation into the evolutionary origin of DNAm at these two CpG sites could provide further insight. Furthermore, evidence for one model does not negate another model of senescence, as they are not mutually exclusive, that is, multiple evolutionary models could contribute to senescence intraspecifically.

Tissue and Cell Type Specific DNAm:

While an increased rate of DNAm with age is an interesting finding, it remains unclear what biological phenomenon might be responsible for such age-related DNAm. Many studies have found not only tissue specific DNAm patterns but also cell type specific DNAm patterns (Xie et al. 2013; Lokk et al. 2012). Furthermore, such tissue- and cell type-specific DNAm patterns could be involved in cell lineage differentiation (Reinius et al. 2012). Here, we investigated age-related DNAm patterns among datasets of different tissue and cell types in order to explore the possibility that a particular cell type could be driving the change in rate of DNAm with age. While our investigation is far from exhaustive, we consider it a preliminary exploration into the possibility that our findings are specific to a particular cell type or set of cell types.

The analyses on GTP and TOPS data (both peripheral blood) exhibited a significant quadratic relationship between β and age. Subsequent to these analyses, we investigated datasets with isolated blood cell types to explore the possibility that a particular cell type was driving the effect. First MESA-T and MESA-M were analyzed. From these analyses we found that MESA-M had a significant quadrilinear relationship between β and age but MESA-T exhibited no such relationship, suggesting that monocytes may be responsible or partially responsible for the original finding. Nevertheless, the age range of MESA-T (45-79 years) lacked young individuals, which would render it difficult to detect such an effect, if there was one. Subsequent analyses were performed on EGC-PBL, EGC-CD4, and EGC-CD8 to further investigate cell-type specific effects. While a quadrilinear relationship between age and DNAm was not identified in EGC-PBL, EGC-CD4, and EGC-CD8, we cannot conclude that no such relationship exists considering the data did not include individuals between the ages 23 and 72, which could hinder the ability to detect such a quadratic effect. In order to further understand the quadratic relationship between β and age among different cell types at the two CpG sites of interest, future studies should focus on cell-type specific data with large samples and a wide age range. However, our subsequent comparison of the variation between young and old individuals suggests that there is an increase in variation with age for both of these CpG sites in all cell types and tissues examined. This increase is likely responsible for the significant quadrilinear terms observed in GTP, TOPS, and MESA-M, and can itself be explained by a small number of older subjects demonstrating increased DNAm at these sites, which are unmethylated in most subjects.

Putative Biological Function of cg07955995 and cg22285878:

cg07955995 and cg22285878 reside in the promoter region of the gene KLF14. KLF14 belongs to a family of 17 proteins known as Kruppel-like factor (KLF) proteins, which are involved in immune cell differentiation. KLF proteins exert their effect via the binding to gene promoters and enhancers (Sarmento et al. 2015).

A recent study found that KLF14 represses FOXP3 via the epigenetic regulation at the Treg-specific demethylated region (TSDR) of FOXP3 *in vitro* and *in vivo* using a mouse model (Sarmento et al. 2015). Previous work has shown that FOXP3 induces and is required for the conversion of $CD14^{+}CD25^{-}$ naïve T cells to $CD4^{+}CD25^{+}$ regulatory T cells via the joint stimulation of TCR and TGF-

β (Chen et al 2003; Fontenot et al. 2003; Hori et al. 2003). A study of CD25-deficient mice found that the over-expression of proinflammatory cytokines, including IL-2 and IFN-γ, in response to bacterial superantigen stimulation, particularly staphylococcal enterotoxin B, was curbed by injection of $CD4^+CD25^+$ suggesting that $CD4^+CD25^+$ plays a role in regulating antigen-induced inflammatory responses (Pontoux et al. 2002). In consideration of the aforementioned studies and our findings, we suggest a biological pathway whereby late-life DNAm at cg07955995 and cg22285878 could induce immunosenescence: (1) Age-related DNAm in the promoter region of KLF14 downregulates expression of KLF14; (2) In the absence of normal levels of KLF14, which otherwise inhibits FOXP3, FOXP3 expression is upregulated; (3) Higher levels of FOXP3 induces a greater number of naïve $CD14^+25^-$ T cells to convert to $CD14^+25^+$ T cells (sometimes referred to as Treg cells); and (4) $CD14^+25^+$ T cells regulate the inflammatory response to superantigen.

While our results can only be taken as a preliminary investigation, we suggest DNAm at cg22285878 could regulate KLF14 thereby contributing to an immunosenescent phenotype in older adults. Immunosenescence is the general deterioration of the immune system over the course of the lifespan. While aging affects all aspects of the immune system, T cells are the most severely affected (Linton et al. 2004). While many factors may contribute to an immunosenescent phenotype, one such factor, namely, infection with cytomegalovirus (CMV) is particularly prevalent among the elderly and has been implicated in CD8+ T cell proliferation (Tserel 2015). Moreover, a substantial portion of peripheral blood CD4+ T cell and CD8+ T responses has been observed in CMV-seropositive individuals comprising approximately 10% of T cell memory compartments (Sywester et al. 2005). Our investigation compared variances of DNAm between old age groups and young age groups among EGC-PBL, EGC-CD4, and EGC-CD8. Among these three cell types, we found that $CD4^+$ T cells had the greatest variance in DNAm in the old age group for cg22285878. The higher variance among old adults in $CD4^+$ T cells relative to other cell types could reflect an immunosenescent phenotype in some elderly people, but not others. Moreover, we found that a few old people had particularly high DNAm at cg07955995 and cg22285878, which could indicate a possible immunosenescent phenotype in these individuals. More work must be done to understand the role of age-related DNAm at cg22285878 and its possible relationship with immunosenescence, such as CMV infection.

Conclusion:

Our study is the first to implicate two age differential methylated CpG sites, which reside in the promoter region of KLF14, in immune system function, most notably the immunosenescent phenotype seen among the elderly. In order to further understand this putative biological pathway, future research should pay particular attention to DNAm data in immunosenescent individuals, including those who are CMV-seropositive. While our investigation suggests differences in the variances of DNAm at cg22285878 among the old age group, we were unable to discover a significant quadrilinear relationship between DNAm and age in $CD4^+$ T cells. Nevertheless, our comparison of $CD4^+$ T cells, $CD8^+$ T cells, and PBL used data that lacked individuals between 23 and 72 years of age, which greatly limits the ability to detect such an age-related effect. Thus, in order to further understand this relationship, future research should collect DNAm data from isolated blood cell types among individuals with a wide range of ages.

A number of limitations of this study are worth noting. Firstly, all eight datasets are cross-sectional, so it is not possible to infer that an increase in age yields an exponential increase in DNAm, merely that there is an association. Moreover, it is possible that environmental factors, such as a pollutant accumulates in the body over time, resulting in an age-related relationship at these two CpG sites. Ideally, longitudinal data could better assess the relationship in question.

Overall, our study is the first to utilize methylome-wide association studies to investigate a quadratic relationship between DNAm and age. We found two CpG sites that exhibit stable DNAm early in life followed by a rapid increase in DNAm in late life. We then determined these CpG sites reside in the promoter region of KLF14, which has recently been shown to be involved in immune system function via the suppression of FOXP3. These findings highlight the importance of DNAm in furthering our understanding of aging, immunology, and biological pathways, more generally.

Chapter 2. The Application of Regularization Methods To Predict Human Age

2.1 Tables & Figures

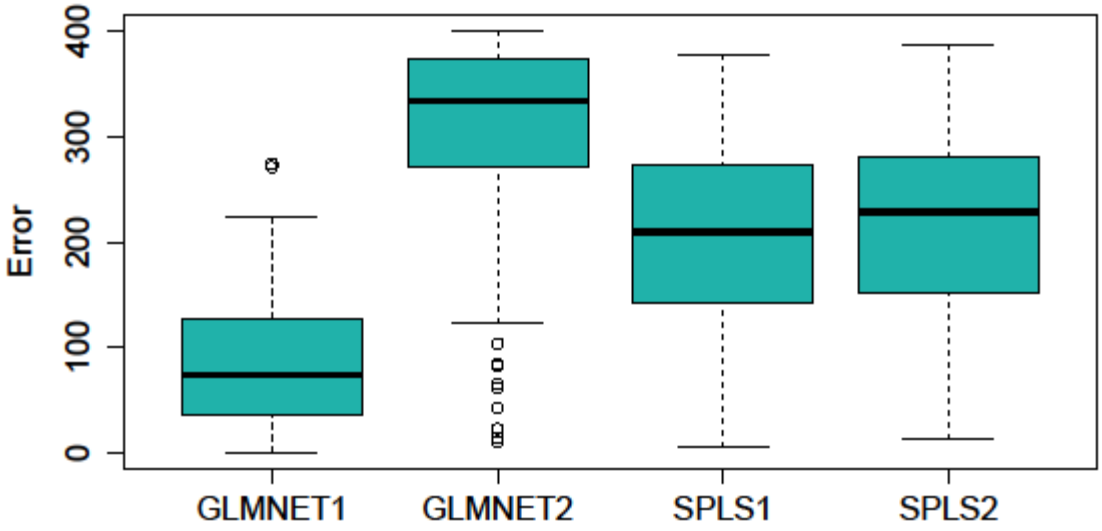


Figure 2. Boxplots of the distributions of test errors corresponding to the 100 iterations for each of the four analyses. GLMNET1 corresponds to the lasso model of the non-transformed data, GLMNET2 corresponds to the lasso model of the combined data, SPLS1 corresponds to the sparse partial least squares model of the non-transformed data, and SPLS2 corresponds to the sparse partial least squares model of the transformed data.

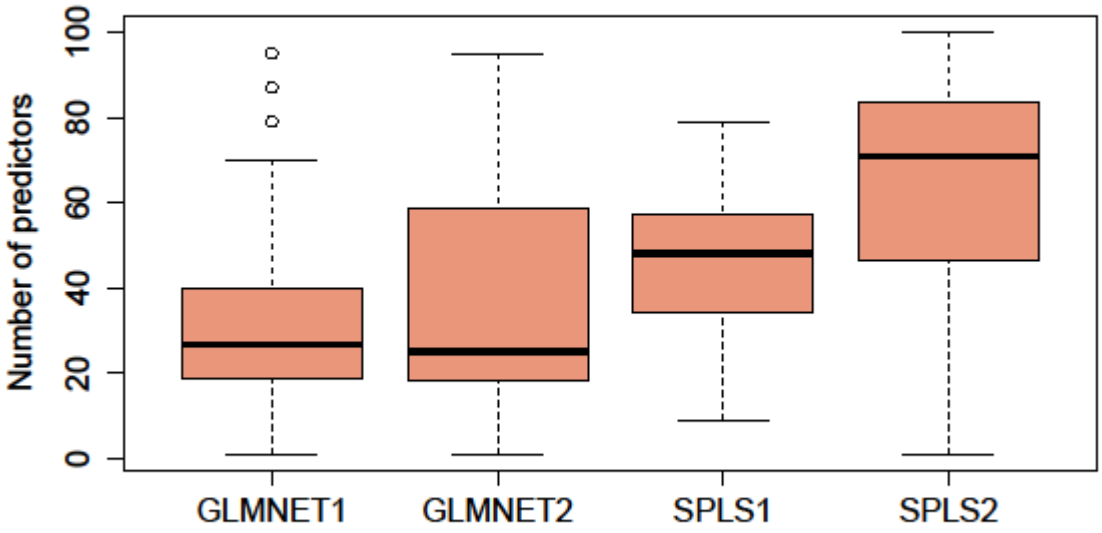


Figure 3. Boxplots of the distributions of the number of predictors corresponding to the 100 iterations for each of the four analyses. GLMNET1, GLMNET2, SPLS1, SPLS2 designate the same analyses specified in Figure 1.

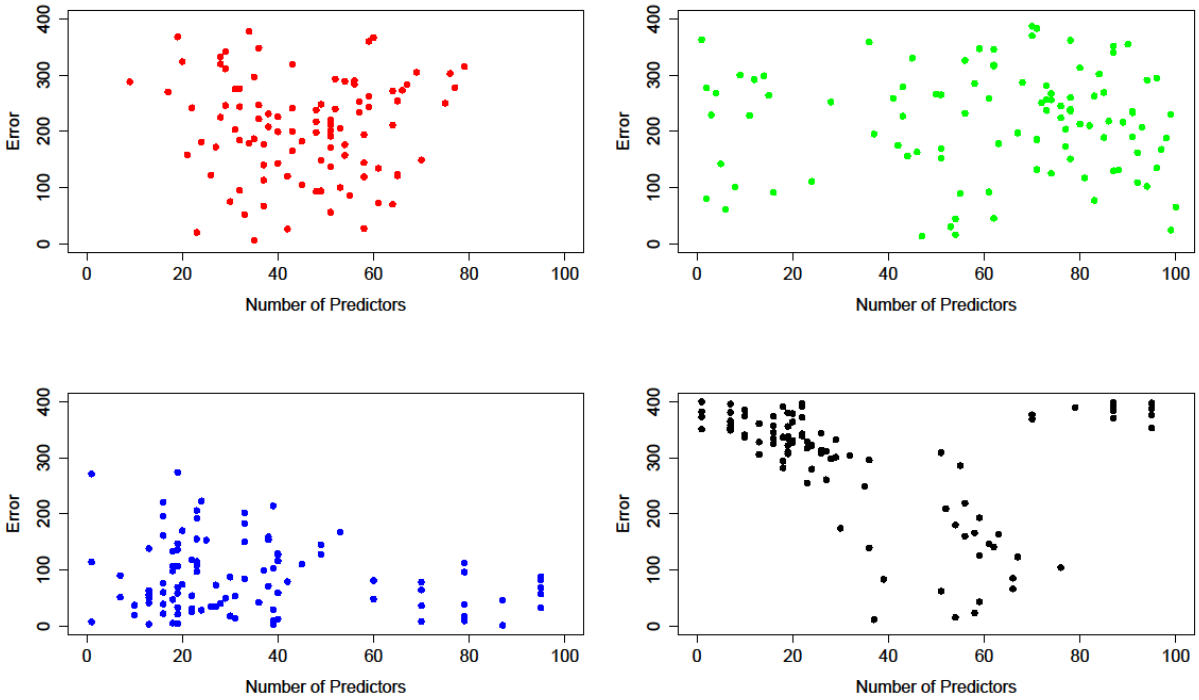


Figure 3. Error vs. number of predictors for each of the four analyses. The first plot (red dots) corresponds to the SPLS model of the non-transformed data. The second plot (green dots) corresponds to the SPLS model of the combined data. The third plot (blue dots) corresponds to the lasso model of the non-transformed data. The fourth plot (black dots) corresponds to the lasso model of the combined data.

	Error Mean (SD)	**Number of predictors** Mean (SD)
GLMNET1	86.4 (63.7)	45.6 (15.0)
GLMNET2	296.2 (104.1)	62.1 (28.0)
SPLS1	204.2 (87.7)	34.82 (24.3)
SPLS2	215.1 (93.0)	37.0 (27.5)

Table 1. Mean and standard deviation (SD) values corresponding to the error and number of predictors for each of the four analyses.

2.2 Introduction:

Over the past few decades, technological advancements have permitted the collection of high-dimensional biological data, which have resulted in the realization that biological systems are much more complex than previously imagined (Quackenbush 2007). In addition to the collection of high-dimensional genomics and transcriptomics data, biology has also witnessed the introduction of high-dimensional epigenetic data, most notably, DNA methylation (DNAm). While previous microarrays included 1,505 CpG sites and 27,000 CpG sites, respectively, currently available is the Infinium HumanMethylation450 BeadChip, which permits the detection of differential DNAm at 485,764 CpG sites (Sandoval 2011).

A wide range of studies have observed that a substantial portion of differentially methylated CpG sites associate with age across many tissues and cell types (Alisch 2012; Barfield 2013; Christensen 2009; Florath 2013; Fraga 2005; Gentilini 2013; Heyn 2011; Horvath 2012; Jaffe et al. 2016; Jones et al 2015; Linton 2004; Martin 2005; Martino 2011; McClay 2013; Poulsen 2007; Shroeder 2011). Furthermore, many researchers have utilized statistical tools to find a linear combination of CpG sites in order to predict human age and all-cause mortality. While some of these studies have focused on aging in particular tissues and cell types, others have been adapted to predict age across multiple tissues (Bocklandt 2011; Hannum 2013; Horvath 2013; Weidner 2014).

The tools utilized to create a predictive model of aging using DNAm data build on statistical techniques developed several decades ago to address problems corresponding to high dimensional data in general (Wold 1966). More specifically, statisticians have been charged with the task of developing a regression method that finds a linear combination of predictors while simultaneously addressing common high dimensional data issues, including a large number of predictors, high collinearity among these predictors, and a small sample size (Chung 2010).

While the utilization of biological data to predict chronological age is interesting in its own right, its importance resides in its ability to detect senescence, that is, the physiological deterioration and eventual mortality of an organism over time in general and as a specific consequence of disease (Hannum 2012). Many studies have observed a relationship between environmental exposures and senescence. For example, smoking tobacco has been observed to be a risk factor for all-cause mortality (Blair 1989; Prescott 2002). There remains substantial work in order to bridge the gap between environmental exposures and biomarkers of physiological decline, particularly DNAm. Nevertheless, among the studies utilizing methylation data to create predictive models of aging, one study observed accelerated aging in breast cancer tissue, forecasting the promise of such statistical tools as a potential diagnostic tool (Horvath 2013).

While these data-reducing statistical tools have been previously used to predict age via a linear relationship between age and DNAm, studies have observed different rates of DNAm change among subjects of different age groups (Alisch et al. 2012). These observations raise the question as to whether a data transformation that permits modeling a non-linear relationship between age and DNAm could improve the prediction of age.

Our study aims to compare predictive models of human aging that only permit the modeling of a linear relationship with age to models that permit the modeling of both a linear and non-linear relationship with age. In order to accomplish this, we build prediction models whereby a linear combination of age differentially methylated (aDM) CpG sites serve as predictors of age using two R packages (glmnet and SPLS), which perform lasso regression models and sparse partial least squares regression models, respectively (Chung 2010; Friedman 2010). We apply these models to methylation data obtained from peripheral blood samples (Gillespie et al. 2009; Barfield et al. 2014). We then perform a square root transformation on the aDM CpG sites and add these transformed predictors to the original dataset, thereby permitting the modeling of both a linear and non-linear relationship between age and aDM CpG sites. Subsequently, we compare the age predictions corresponding to the non-transformed predictors with the age predictions corresponding to data that includes both square root transformed predictors and non-transformed predictors.

2.3 Methods:

Sample:

The analysis was performed on data obtained from the Grady Trauma Project, which included 336 subjects (see Methods 1 for more details). While the original data set included all CpG sites on the Illumina Infinium 450k Human Methylation Array (approximately 485,000 CpG sites), our analysis was conducted on the 16,747 CpG sites that are also on the Illumina Infinium 27k Human Methylation Array (approximately 27,000 CpG sites). Including only those CpG sites that exist on both arrays permits our tool to be used on data obtained from substantially more studies than would be the case had we included CpG sites included on one array.

Normalization:

Beta-values were computed using methylated and unmethylated signals (see Methods 1). The probe sites corresponding to the CpG sites on the Illumina arrays come in two different designs, which yields beta-values of substantially different distributions. In order to correct for this bias, we used the same beta-mixture quantile normalization method used in Horvath (2013), which was adapted from another study (Teschendorff 2013).

Data Transformation:

A square root transformation was performed on the 16,747 CpG sites that served as predictors. These transformed values were then added as a new set of predictors to the non-transformed data. Thus, the new data set contained non-transformed and square root transformed predictors bringing the total number of predictors to 33,494 predictors.

Analysis:

Two data sets were used for the analysis: the non-transformed data set with 16,747 CpG sites and the combined data set with 33,494 CpG sites, which included 16,747 non-transformed CpG sites as predictors, along with the square root transformation of these same 16,747 CpG sites. Chronological age (in years) was used as the dependent variable.

The analysis comprised one hundred iterations of the following steps (1-3): (1) The data set was randomly subsetted into a test dataset with 60 subjects and a training data set with 276 subjects. (2) A sparse partial least squares regression was performed on the training data set to predict age, using cross-validation to select the tuning paramaeter, and the fitted model was applied to the test data set to get the test error, using the R package SPLS. The test error is defined as below. (3) Step 2 was repeated using a lasso regression model. Test Errors were computed for each method as follows:

$$Error = \frac{\sum(observed\ age - predicted\ age)^2}{sample\ size}$$

The prediction error for each method was calculated as the average of test errors in the 100 iterations.

2.4 Results:

The lasso model of combined data had the highest average error (296.2) and the greatest average number of predictors (62.1) among the four analyses (Table 1; Figure 1). The lasso model of non-transformed data, on the other hand, had the lowest average error (86.4). The sparse partial least squares regression model corresponding to the non-transformed data and the sparse partial least squares regression model corresponding to the combined data yielded similar distributions of errors. Figure 1 and Figure 2 show boxplots of the distribution of errors and the distribution of number of predictors, respectively, for each of the four analyses.

The error vs. number of predictors for each iteration was plotted for each of the four analyses (Figure 3). The two SPLS plots show no distinct pattern. The two GLMNET plots, on the other hand, each show a distinct pattern. Pertaining to GLMNET1, as the number of predictors increases, the error appears to decrease. GLMNET2 shows a similar pattern, although the error is much larger when the number of predictors are less than forty. Moreover, as the number of predictors exceeds 70, the errors increase substantially.

2.5 Discussion:

The original aim of this study was to investigate whether a square root transformation of the methylation data could improve the prediction of age. Neither the lasso model nor the sparse partial least squares model showed any improvement in the prediction of age (Table 1). In fact, the lasso model had a substantially higher average error for the combined data compared to the non-transformed data.

Theoretically, square root transformed predictors should allow the lasso regression model to fit a non-linear relationship between DNAm and age, and possibly improve the prediction of age. Thus, it is unclear why the prediction of age would be worse for the lasso model of the combined data. That being said, the scatterplots indicate that the error is markedly higher for few predictors (<40) and many predictors (>70) for the combined data compared to the non-transformed data.

While it is unclear why a square root transformation would substantially increase the prediction error, it is not surprising that the prediction error would remain unimproved as is the case for the spare partial least squares regression model. Several studies have observed thousands of age-associated CpG sites on the Illumina Infinium 27k Human Methylation Array and tens of thousands of age-associated CpG sites on the Illumina Infinium 450k Human Methylation Array in multiple human tissues based on a linear relationship between DNAm and age (Hernandez 2011; Zongli 2010). Moreover, predictive models using a linear combination of these age-associated CpG sites have yielded relatively accurate predictions of age based on a linear relationship between DNAm and age (Horvath 2013). Although some age-associated CpG sites are better modeled as a non-linear relationship with age (see Results 1), it is possible that the linearly-related age-associated CpG sites are sufficient to build a relatively accurate prediction model of human age.

While the square root transformation did not improve the prediction of age, we should not exclude the possibility that other transformations that would permit a non-linear relationship could improve the prediction of age. Moreover, while the square root transformation did not improve the prediction of age using methylation data, it is possible that such a transformation could improve the prediction for other types of data, particularly those whose relationship between the dependent variable and the predictor variables exhibit a non-linear association. Further research that extends these models to a variety of transformations and multiple data types is necessary to better elucidate how these models would fare under the aforementioned circumstances.

References

Ahuja, N., Li, Q., Mohan, A.L., Baylin, S.B., Issa, J.P., 1998. Aging and DNA methylation in colorectal mucosa and cancer. Cancer research 58, 5489-5494.

Ali, O., Cerjak, D., Kent, J.W., Jr., James, R., Blangero, J., Carless, M.A., Zhang, Y., 2015. An epigenetic map of age-associated autosomal loci in northern European families at high risk for the metabolic syndrome. Clinical epigenetics 7, 12.

Alisch, R.S., Barwick, B.G., Chopra, P., Myrick, L.K., Satten, G.A., Conneely, K.N., Warren, S.T., 2012. Age-associated DNA methylation in pediatric populations. Genome Res 22, 623-632.

Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S.Y., Dempster, E.L., Murray, R.M., Grundberg, E., Hedman, A.K., Nica, A., Small, K.S., Dermitzakis, E.T., McCarthy, M.I., Mill, J., Spector, T.D., Deloukas, P., 2012. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS genetics 8, e1002629.

Bjornsson, H.T., Sigurdsson, M.I., Fallin, M.D., Irizarry, R.A., Aspelund, T., Cui, H., Yu, W., Rongione, M.A., Ekstrom, T.J., Harris, T.B., Launer, L.J., Eiriksdottir, G., Leppert, M.F., Sapienza, C., Gudnason, V., Feinberg, A.P., 2008. Intra-individual change over time in DNA methylation with familial clustering. Jama 299, 2877-2883.

Blair, S.N., Kohl, H.W., Iii, Paffenbarger, R.S., Jr, Clark, D.G., Cooper, K.H., Gibbons, L.W., 1989. Physical fitness and all-cause mortality: A prospective study of healthy men and women. Jama 262, 2395-2401.

Bocklandt, S., Lin, W., Sehl, M.E., Sanchez, F.J., Sinsheimer, J.S., Horvath, S., Vilain, E., 2011. Epigenetic predictor of age. PLoS One 6, e14821.

Bollati, V., Schwartz, J., Wright, R., Litonjua, A., Tarantini, L., Suh, H., Sparrow, D., Vokonas, P., Baccarelli, A., 2009. Decline in genomic DNA methylation through aging in a cohort of elderly subjects. Mechanisms of ageing and development 130, 234-239.

Chen, W., Jin, W., Hardegen, N., Lei, K.J., Li, L., Marinos, N., McGrady, G., Wahl, S.M., 2003. Conversion of peripheral CD4+CD25- naive T cells to CD4+CD25+ regulatory T cells by TGF-beta induction of transcription factor Foxp3. The Journal of experimental medicine 198, 1875-1886.

Christensen, B.C., Houseman, E.A., Marsit, C.J., Zheng, S., Wrensch, M.R., Wiemels, J.L., Nelson, H.H., Karagas, M.R., Padbury, J.F., Bueno, R., Sugarbaker, D.J., Yeh, R.F., Wiencke, J.K., Kelsey, K.T., 2009. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS genetics 5, e1000602.

Chung, D., Keles, S., 2010. Sparse partial least squares classification for high dimensional data. Stat Appl Genet Mol Biol 9, Article17.

Ding, J., Reynolds, L.M., Zeller, T., Muller, C., Lohman, K., Nicklas, B.J., Kritchevsky, S.B., Huang, Z., de la Fuente, A., Soranzo, N., Settlage, R.E., Chuang, C.C., Howard, T., Xu, N., Goodarzi, M.O., Chen, Y.D., Rotter, J.I., Siscovick, D.S., Parks, J.S., Murphy, S., Jacobs, D.R., Jr., Post, W., Tracy, R.P., Wild, P.S., Blankenberg, S., Hoeschele, I., Herrington, D., McCall, C.E., Liu, Y., 2015. Alterations of a Cellular Cholesterol Metabolism Network Are a Molecular Feature of Obesity-Related Type 2 Diabetes and Cardiovascular Disease. Diabetes 64, 3464-3474.

Edgar, R., Tan, P.P., Portales-Casamar, E., Pavlidis, P., 2014. Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. Epigenetics & chromatin 7, 28.

Florath, I., Butterbach, K., Muller, H., Bewerunge-Hudler, M., Brenner, H., 2014. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Hum Mol Genet 23, 1186-1201.

Fontenot, J.D., Gavin, M.A., Rudensky, A.Y., 2003. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. Nature immunology 4, 330-336.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T.D., Wu, Y.Z., Plass, C., Esteller, M., 2005. Epigenetic differences arise during the lifetime of monozygotic twins. Proceedings of the National Academy of Sciences of the United States of America 102, 10604-10609.

Fraga, M.F., Esteller, M., 2007. Epigenetics and aging: the targets and the marks. Trends in genetics : TIG 23, 413-418.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software 33, 1-22.

Gentilini, D., Mari, D., Castaldi, D., Remondini, D., Ogliari, G., Ostan, R., Bucci, L., Sirchia, S.M., Tabano, S., Cavagnini, F., Monti, D., Franceschi, C., Di Blasio, A.M., Vitale, G., 2013. Role of epigenetics in human aging and longevity: genome-wide DNA methylation profile in centenarians and centenarians' offspring. Age (Dordrecht, Netherlands) 35, 1961-1973.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., Zhang, K., 2013. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell 49, 359-367.

Hernandez, D.G., Nalls, M.A., Gibbs, J.R., Arepalli, S., van der Brug, M., Chong, S., Moore, M., Longo, D.L., Cookson, M.R., Traynor, B.J., Singleton, A.B., 2011. Distinct DNA methylation changes highly correlated with chronological age in the human brain. Human Molecular Genetics.

Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J., Puca, A.A., Sayols, S., Pujana, M.A., Serra-Musach, J., Iglesias-Platas, I., Formiga, F., Fernandez, A.F., Fraga, M.F., Heath, S.C., Valencia, A., Gut, I.G., Wang, J., Esteller, M., 2012. Distinct DNA methylomes of newborns and centenarians. Proceedings of the National Academy of Sciences of the United States of America 109, 10522-10527.

Hori, S., Nomura, T., Sakaguchi, S., 2003. Control of regulatory T cell development by the transcription factor Foxp3. Science 299, 1057-1061.

Horvath, S., 2013. DNA methylation age of human tissues and cell types. Genome biology 14, R115.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R.S., Boks, M.P., van Eijk, K., van den Berg, L.H., Ophoff, R.A., 2012. Aging effects on DNA methylation modules in human brain and blood tissue. Genome biology 13, R97.

Issa, J.P., Ottaviano, Y.L., Celano, P., Hamilton, S.R., Davidson, N.E., Baylin, S.B., 1994. Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. Nature genetics 7, 536-540.

Jaenisch, R., Bird, A., 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature genetics 33 Suppl, 245-254.

Jones, M.J., Goodman, S.J., Kobor, M.S., 2015. DNA methylation and healthy human aging. Aging cell.

Kirkwood, T.B., Holliday, R., 1979. The evolution of ageing and longevity. Proc R Soc Lond B Biol Sci 205, 531-546.

Linton, P.J., Dorshkind, K., 2004. Age-related changes in lymphocyte development and function. Nature immunology 5, 133-139.

Lokk, K., Modhukur, V., Rajashekar, B., Martens, K., Magi, R., Kolde, R., Koltsina, M., Nilsson, T.K., Vilo, J., Salumets, A., Tonisson, N., 2014. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome biology 15, r54.

Martin, G.M., 2005. Epigenetic drift in aging identical twins. Proceedings of the National Academy of Sciences of the United States of America 102, 10413-10414.

Martino, D.J., Tulic, M.K., Gordon, L., Hodder, M., Richman, T.R., Metcalfe, J., Prescott, S.L., Saffery, R., 2011. Evidence for age-related and individual-specific changes in DNA methylation profile of mononuclear cells during early immune development in humans. Epigenetics 6, 1085-1094.

Medawar, P.B., 1952. An Unsolved Problem of Biology. Western Printing Services LTD Bristol, University College London.

Mendizabal, I., Keller, T.E., Zeng, J., Yi, S.V., 2014. Epigenetics and evolution. Integrative and comparative biology 54, 31-42.

Nakagawa, H., Nuovo, G.J., Zervos, E.E., Martin, E.W., Jr., Salovaara, R., Aaltonen, L.A., de la Chapelle, A., 2001. Age-related hypermethylation of the 5' region of MLH1 in normal colonic mucosa is associated with microsatellite-unstable colorectal cancer development. Cancer research 61, 6991-6995.

Pontoux, C., Banz, A., Papiernik, M., 2002. Natural CD4 CD25(+) regulatory T cells control the burst of superantigen-induced cytokine production: the role of IL-10. International immunology 14, 233-239.

Poulsen, P., Esteller, M., Vaag, A., Fraga, M.F., 2007. The epigenetic basis of twin discordance in age-related diseases. Pediatr Res 61, 38R-42R.

Prescott, E., Scharling, H., Osler, M., Schnohr, P., 2002. Importance of light smoking and inhalation habits on risk of myocardial infarction and all cause mortality. A 22 year follow up of 12 149 men and women in The Copenhagen City Heart Study. Journal of epidemiology and community health 56, 702-706.

Reinius, L.E., Acevedo, N., Joerink, M., Pershagen, G., Dahlen, S.E., Greco, D., Soderhall, C., Scheynius, A., Kere, J., 2012. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One 7, e41361.

Reynolds, L.M., Taylor, J.R., Ding, J., Lohman, K., Johnson, C., Siscovick, D., Burke, G., Post, W., Shea, S., Jacobs, D.R., Jr., Stunnenberg, H., Kritchevsky, S.B., Hoeschele, I., McCall, C.E., Herrington, D.M., Tracy, R.P., Liu, Y., 2014. Age-related variations in the methylome

associated with gene expression in human monocytes and T cells. Nature communications 5, 5366.

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M., Esteller, M., 2011. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics 6, 692-702.

Sarmento, O.F., Svingen, P.A., Xiong, Y., Xavier, R.J., McGovern, D., Smyrk, T.C., Papadakis, K.A., Urrutia, R.A., Faubion, W.A., 2015. A Novel Role for Kruppel-like Factor 14 (KLF14) in T-Regulatory Cell Differentiation. CMGH Cellular and Molecular Gastroenterology and Hepatology 1, 188-202.e184.

Schroeder, J.W., Conneely, K.N., Cubells, J.C., Kilaru, V., Newport, D.J., Knight, B.T., Stowe, Z.N., Brennan, P.A., Krushkal, J., Tylavsky, F.A., Taylor, R.N., Adkins, R.M., Smith, A.K., 2011. Neonatal DNA methylation patterns associate with gestational age. Epigenetics 6, 1498-1504.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R.A., Issa, J.P., 2007. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS genetics 3, 2023-2036.

So, K., Tamura, G., Honda, T., Homma, N., Waki, T., Togawa, N., Nishizuka, S., Motoyama, T., 2006. Multiple tumor suppressor genes are increasingly methylated with age in non-neoplastic gastric epithelia. Cancer science 97, 1155-1158.

Sylwester, A.W., Mitchell, B.L., Edgar, J.B., Taormina, C., Pelte, C., Ruchti, F., Sleath, P.R., Grabstein, K.H., Hosken, N.A., Kern, F., Nelson, J.A., Picker, L.J., 2005. Broadly targeted human cytomegalovirus-specific CD4+ and CD8+ T cells dominate the memory compartments of exposed subjects. The Journal of experimental medicine 202, 673-685.

Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., Beck, S., 2013. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics (Oxford, England) 29, 189-196.

Tserel, L., Kolde, R., Limbach, M., Tretyakov, K., Kasela, S., Kisand, K., Saare, M., Vilo, J., Metspalu, A., Milani, L., Peterson, P., 2015. Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. Sci Rep 5, 13107.

Weidner, C.I., Lin, Q., Koch, C.M., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D.O., Jöckel, K.-H., Erbel, R., Mühleisen, T.W., Zenke, M., Brümmendorf, T.H., Wagner, W., 2014. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome biology 15, 1-12.

Williams, G.C., 1957. Pleiotropy, Natural Selection, and the Evolution of Senescence. Evolution 11, 398-411.

Wold, H.O.A., 1968. Nonlinear estimation by iterative least square procedures.

Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., Yang, H., Wang, T., Lee, A.Y., Swanson, S.A., Zhang, J., Zhu, Y., Kim, A., Nery, J.R., Urich, M.A., Kuan, S., Yen, C.A., Klugman, S., Yu, P., Suknuntha, K., Propson, N.E., Chen, H., Edsall, L.E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.Y., Chi, N.C., Antosiewicz-Bourget, J.E., Slukvin, I., Stewart, R., Zhang, M.Q., Wang, W.,

Thomson, J.A., Ecker, J.R., Ren, B., 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 153, 1134-1148.

Xu, Z., Taylor, J.A., 2014. Genome-wide age-related DNA methylation changes in blood and other tissues relate to histone modification, expression and cancer. Carcinogenesis 35, 356-364.