Distribution Agreement

Signature:

______________________________ _______________
Crystal Deloris Grant Date

The Epigenetics of Aging: Exploring Biomarkers and the Interplay Within the Aging Epigenome

By

Crystal Deloris Grant
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

Karen Conneely
Advisor

Peng Jin
Advisor

Michael Epstein
Committee Member

Stephanie Sherman
Committee Member

Paula Vertino
Committee Member

Hao Wu
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

The Epigenetics of Aging: Exploring Biomarkers and the Interplay Within the Aging Epigenome

By

Crystal Deloris Grant
B.A., Cornell University, 2013

Advisors: Karen Conneely, Ph.D. and Peng Jin, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

2020

Abstract

The Epigenetics of Aging: Exploring Biomarkers and the Interplay Within the Aging Epigenome

By Crystal Deloris Grant

The process of aging is poorly understood yet age remains the main predictor of physiological decline and disease development in humans. Aging is marked by widespread, reproducible changes to the epigenome. The most studied epigenetic modification is DNA methylation (DNAm), which shows robust, genome-wide changes with age. These DNAm changes have been used to construct highly accurate blood-based models of chronological age. Using these models, it was found that individuals with a predicted DNAm age higher than their actual chronological age are at increased risk of all-cause mortality. This measure, termed the participants' epigenetic age acceleration, may then serve as a proxy for some measures of health. First, I present a study of how this age acceleration term contributes to longitudinal models of phenotypes associated with Type II Diabetes (T2D)—an age-related disease. I found that this epigenetic age acceleration term remained stable over the 16 years the participants were sampled, and that this term does associate with risk factors for T2D. Our results suggest that DNAm has the potential to act as a mediator between aging and diabetes-related phenotypes, or alternatively, that it may serve as a biomarker of these phenotypes. Next, I present work that aimed to uncover how variability in DNAm with age may be useful in modeling risk for developing adverse age-related phenotypes. I identified age-related variably methylated cytosines, then used these sites to construct a score indicating the amount of epigenetic drift an individual was undergoing. Though this score did not appear to contribute significantly to longitudinal models of aging phenotypes or mortality risk, other biomarkers that incorporate information about DNAm variability maybe be informative. Lastly, looking to other levels of the epigenome as part of a pilot study, I characterized changes in chromatin accessibility and three histone modifications with age. This led to the identification of regions of age-related change as well as the observation of which histone modifications can be informative in future aging studies. My dissertation work sheds light on which types of epigenetic changes can be used to inform biomarkers of biological aging, informing future studies of the epigenetics of aging.

The Epigenetics of Aging: Exploring Biomarkers and the Interplay Within the Aging Epigenome

By

Crystal Deloris Grant
B.A., Cornell University, 2013

Advisors: Karen Conneely, Ph.D. and Peng Jin, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology
2020

Acknowledgments

I would like to thank my family for all of their patience, love, support, and encouragement.

I would like to thank my friends for the joy and sense of belonging they brought me. Through knowing them, I found myself. For that, I am and will remain eternally grateful. They made grad school, dare I say, fun.

I would like to thank my committee members and co-advisor for their help and guidance in contributing to the work in this dissertation (also for some very fun and very blurry memories at the GMB retreats).

I would like to thank the faculty who made the IMSD program possible, especially Gillian, Pat, and Amanda. The program helped me find my tribe at Emory.

I would also like to thank the faculty in the M2M program, especially Nael, for their mentorship and their leadership in developing a PhD track for people like me who don't quite fit in in their GDBBS program.

I would like to thank my lab mates, Nick and Liz, for lots of laughs, good conversation, and for their always helpful advice. It's been wonderful distracting them and my other 341 officemates from getting their work done these last few years.

Lastly, I would like to thank my principal advisor, Karen, for being exactly the mentor that I needed. I cannot imagine having worked with anyone else (I certainly wouldn't have enjoyed my PhD nearly as much as I have). I think often of the day in April when I nervously asked if I could join her lab, and how shocked I was that she agreed. Each time I think of it, I consider how truly lucky I am that she saw something in me and, much like a stray cat that your lease says you're not allowed to have, took me in.

Table of Contents

Tables and Figures

Chapter II

Chapter III

Chapter IV

Supplementary Tables and Figures

Chapter II

Chapter III

Chapter IV

Chapter I: Introduction

Epidemiology of Aging

There is a single phenomenon that is the most profound risk factor for nearly all non-communicable diseases afflicting humans—that phenomenon is aging. Despite the concept of aging and the changes that accompany it being familiar to all who undergo it, at the biological level, aging is not fully understood by humans.[1] Aging is a complex process characterized by the time-dependent, generalized decline in function of an organism, decreasing its fitness. This increasing dysregulation increases an organism's susceptibility to stress, disease, and injury, ultimately leading to mortality. Age is the main risk factor for disease development in humans;[2,3] in many countries, age-related diseases like cardiovascular diseases, diabetes, cancer, and neurodegenerative disorders are among the dominant health problems faced by the population.[4] Considering the negative influence of aging on an organism's fitness and ability to reproduce, it is paradoxical that such a process would evolve to exist in nearly all living organisms. While many theories have been proposed to explain the paradox of aging,[5] the process remains poorly understood.[1]

Worldwide, the population aged 65 years and older is growing rapidly, with a 150% expansion projected over the next few decades.[6] The over 65 population worldwide is expected to jump from approximately 130 million in 1950 to an estimated 1.6 billion by 2050.[7] This is due not to a slowing of the aging process, as this rate appears stable across populations and periods of time,[8] but instead can be attributed to an increase of the number of individuals surviving to later stages in life. Despite these recent global gains in life expectancy, age-related disease burden and the incidence of chronic disabilities remain high.[9] Considering that in many economically-developed countries, much of the costs of both health and social care occur in the final decades of an individual's life,[10] these shifting population demographics could cause severe economic and societal ramifications. Without intervention, the proportion of individuals afflicted with chronic age-related diseases, including those with co-occurrences of multiple chronic conditions, termed comorbidity, will continue to increase.[11] Thus, as the human lifespan continues to increase, interventions to ensure that this time is spent in good health are vital.

The number of years spent in good health is termed the healthspan.[9] Among the aging population, healthspan remains highly variable, with some maintaining good health throughout their lives while others fall ill.[3] Thus, it is not necessarily the case that all who age are destined for sickness. Differential susceptibility to age-related diseases can be attributed to biological differences between individuals, which work to modify disease risk.[12] Because aging is highly reflective of an individual's biology—incorporating everything from an individual's genetics, life style choices and behaviors, diet, environment, and stochastic factors, it is a highly individual process. Part of what complicates the study of aging is that aging is highly multi-factorial, both in the factors that affect it and those that are influenced by it. It is not the case that one system goes awry in the process of aging; instead, all biological systems reflect dysregulation in distinct yet interrelated ways.[13]

The adverse health outcomes associated with aging begin at the molecular level. The accumulation of damage and increased dysregulation impairing individuals' ability to repair damage leads to compromised cell and tissue function, manifesting as the characteristics commonly observed in aging.[9,14] The following are considered the nine hallmarks of aging: genomic instability, telomere attrition, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, altered intercellular communication, and epigenetic alterations.[13] The breadth of these hallmarks reflect both the complexity of aging—in that it is the case of multiple mechanism going awry in parallel instead of a single point of failure, as well as the interrelatedness of contributors to the phenomenon of aging. Additionally, while some ground has been broken in studying aging by characterizing interventions and genotypes that lead to longer lifespans in model organisms,[15,16] the translation to clinical and behavioral interventions that can be made by humans has been minimal. This underlies both the importance of leveraging network theories of aging in which molecular, cellular, and systemic level changes are integrated,[17] and the utility of human epidemiological studies in better understanding aging at a fundamental level and using this understanding to improve health.

Biomarkers of Aging

Biomarkers are medical signs used to objectively indicate biological processes underlying a patient's health in the absence of symptoms of disease.[18] Examples of biomarkers include simple, non-invasive measures such as a patient's pulse or blood pressure, through to more complex laboratory tests requiring clinical samples of blood or other tissues. The clinical utility of a biomarker is dependent on its high levels of reproducibility and accuracy in predicting the incidence or outcome of disease in individuals across different demographics and populations.[19] Considering the link between aging and disease risk in humans, biomarkers specifically predictive of the aging process would be highly valuable. These biomarkers would inform both what phenotypes are consistent with the normal, healthy aging process as well as contribute to personalized medicine, a method of clinical practice that uses new technologies to provide individualized medical decisions regarding the prediction, prevention, and treatment of disease.[20] In short, an accurate biomarker of the aging process could transform the field of personalized medicine by aiding in early diagnosis and identifying at-risk patients to whom medical interventions would be most efficacious—ultimately lowering healthcare costs and improving health outcomes.[21]

A Biomarker of Aging (BoA) is a measurable indicator of the biological age of an organism, in the absence of disease.[22] Biological aging refers to the change over time to biological processes in an organism; this differs from chronological aging, which is defined solely by the passage of time.[2,22] Thus, aging can be viewed as a process occurring at different rates for different individuals, irrespective of the actual passage of time. The American Federation for Aging Research, in proposing their criteria for a BoA, suggest that it: be able to predict the rate of aging, monitor the aging process in the absence of disease, be functional in model organisms, and can be performed repeatedly on a subject without harm.[23] Developing such a marker has been the focus of many recent studies with the goal of optimizing healthspan and longeveity.[24,25]

Many candidate BoA have been developed that assay aging at different levels. One study, using photographs of individual's faces, asked subjects to rate them on their perceived age. This perceived age based on a photograph was found to associate with the biological age measure in the cohort of young individuals of the same chronological age.[26] Moreover, this perceived age based on photographs was even found to predict mortality in monozygotic (MZ) twins aged ≥70 years, finding that the greater the difference in perceived age, the more likely that the older-looking twin died first.[27] Many other candidate biomarkers relying on physical indicators of health have been found to predict morbidity and mortality well, including walking speed, grip strength, forced expiratory volume, and more.[17] Biomarkers that rely on molecular indicators of health, including circulating C-reactive protein, creatinine, and fasting glucose,[7] have also been developed and are useful in indicating health status, though no single measurement has yet been found to accurately capture biological age.[28]

By definition, biomarkers of biological age should be better indicators of overall health than chronological age.[22,26] Despite age being the main risk factor in disease development, individuals of the same age who share nearly identical genomes at birth can be discordant for disease development. These cases, involving MZ twins, suggest that in addition to accounting for an individual's genome in indicating their risk of an age-related disease, a BoA must also account for other factors influencing an individual's health. Thus, a way to improve existing biomarkers is by incorporating an individual's genetic information in addition to their distinct interactions with their environment by including epigenetic data.

Epigenetics

Introduction

The term epigenetics, coined in 1942 by Conrad Waddington,[29] refers to biological events and phenotypes not wholly explained by genetic principles. Epigenetics represents the bridge between genotype and phenotype, encompassing the changes in gene products and cellular phenotypes in the absence of changes to the DNA sequence itself.[30] Epigenetic alterations have the potential to be reversible, making them a promising tool in personalized medicine involved targeted epigenetic therapies.[31] Additionally, they are dynamic, serving as a response to both intra-and extra-cellular stimuli.[32] It is this plasticity that allows for cellular diversity in organisms; though each cell possesses essentially the same genetic sequence, through epigenetic alterations, distinct cellular identities are established from the same genetic material.[33] The coordinated epigenetic changes in somatic cells are vital for proper organismal development, and the erasure of these changes in the germline vital for reproduction.[34] It is the establishment and stable passage of these alterations during cellular division that allow different tissues to maintain their cellular identities. Epigenetic changes can be influenced by genetics, mediated by an organism's interaction with its environment, and can occur stochastically as a result of drift.[35] Thus, it is the epigenome's mutability that empowers it to mediate necessary changes from the, largely immutable, genetic code and pose it as a promising target in understanding human health and aging.

Part of what allows the epigenome its mutability is that it is mediated by the covalent and noncovalent modifications to individual base pairs of the DNA and histone proteins; taken together, these interactions compose the substructure of chromatin.[30] Each level of the epigenome encodes different information and has a reciprocal relationship with the other levels, in which it can influence them and be influenced by them.[36] Chromatin, which can be broadly classified as euchromatin (transcriptionally active, less compact) or heterochromatin (transcriptionally inactive, more compact),[37] can undergo remodeling mediated by chromatin remodeler enzymes which act in a cell-type specific and developmental-stage specific manner.[38] The substructure of chromatin can exert regulatory functions by modifying the binding sites to transcription factors (TFs), as well as the spatial accessibility of DNA to transcription machinery, thus chromatin structure can directly affect cellular processes such as transcription, DNA repair, and replication.[39] Chromatin's substructure can be assayed by several high throughput techniques in which accessible regions are cleaved followed by sequencing to reveal where the chromatin is most accessible.[40] This can be performed by nucleases and targeted to the level of the individual nucleosome, as in MNase-Seq,[41] or more broadly, as in DNase-Seq.[42] In the newest, and most efficient technique, a transposase inserts a sequence containing adapters into accessible regions which are then amplified and sequenced in the Assay for Transposase Accessible Chromatin (ATAC-Seq).[43]

Chromatin structure is then influenced by the presence or absence of histone octamers comprising nucleosomes as well as the structural and functional variants of these histone proteins. A nucleosome consists of 147 base pairs of DNA wrapped around a histone octamer, where the octamer is made of two of each of the core histone proteins H2A, H2B, H3 and H4.[44] The lysine residues on these different histone variants can then be modified in many ways, including: phosphorylation,[45] sumoylation,[46] ubiquitylation,[47] acetylation,[48] and methylation.[49] Histone modifications are established by enzymes that can transfer the specific group, for example histone acetyltransferases (HATs) for acetyl groups, histone methyltransferase (HMTs) for methyl groups, to specific amino acids on the histone proteins.[50] These modifications to different histones can trigger binding to the chromatin of different proteins, resulting in biochemically-induced structural changes to chromatin architecture as well as functional changes to gene expression.[48] For example, the addition of an acetyl group to a histone tail often attenuates the slight positive charge on the histone protein, resulting in it being less tightly wound to the negatively charged DNA—which is often associated with increased gene transcription.[51] While the methylation of histones can be associated with either an increase or decrease in gene expression depending on its location on the histone octamer.[33]

In addition to specific histone modifications directly influencing function, the composition of the modifications can also be used to identify the functional state of different regions of the genome. More specifically, histone H3 lysine 4 trimethylation (H3K4me3) is associated with promoter regions; H3 lysine 27 acetylation (H3K27ac) is associated with active enhancer regions; H3 lysine 27 trimethylation (H3K27me3) is associated with Polycomb repression; H3 lysine 9 trimethylation (H3K9me3) is associated with heterochromatin regions; and H4 lysine 20 trimethylation (H4K20me3) is associated with constitutive heterochromatin.[33] These individual modifications can be assayed and their genomic locations identified using enrichment assays with antibodies made to specific modifications followed by sequencing, or chromatin immunoprecipitation with sequencing (ChIP-Seq).[52] Histones provide the architecture through which DNA is compacted and this DNA is similarly subject to modification.

DNA Methylation

In 1975, two papers suggested that methylation of a cytosine (C) occurring next to a guanine (G), forming a CpG dinucleotide, could serve as an epigenetic mark in vertebrates.[53,54]

Since then, because of its stability and relative ease to assay genome-wide, DNA methylation (DNAm) has become the most studied modification to chromatin. DNAm generally refers to a CpG dinucleotide that has had a methyl group covalently attached, forming 5-methylcytosine (5-mC),[55] though methylation can occur to cytosines in other contexts.[56,57] The *de novo* addition of the methyl group is catalyzed by DNA methyltransferases (DNMTs), specifically, DNMT3A and DNMT3B, while the maintenance of DNAm patterns after cellular replication is carried out by DNMT1.[34] There are around 28 million CpG sites in the human genome with the majority, approximately 70–80%, being methylated.[56,58]

DNAm can be assayed across the genome using several methods which involve a bisulfite conversion step.[59] In this step, unmethylated cytosines are deaminated to uracil, while the methyl groups on 5-mC cytosines are protected from this step, remaining cytosines. The DNA can then be sequenced at base-pair resolution, termed bisulfite sequencing (BS-Seq), to allow for the mapping of which CpGs are methylated in the genome. While this method is useful in determining methylation patterns in the genome, BS-Seq is unable to distinguish between different modifications made to 5-mC. 5-mC can be oxidized to 5-hydroymethylcytosine (5-hmC) by the ten-11 translocation (TET) enzyme family proteins.[60] TET can further oxidize 5-hmC to 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC). Research has suggested both the these modified version of 5-mC are part of an active DNA demethylation process,[61] and that 5-hmC can act as its own stable mark showing enrichment at gene bodies and cell-type specific changes in its location indicating a functional role.[62,63] In addition to BS-Seq, array-based methods offer the benefit of parallelization in that the methylation profiles of thousands of CpGs can be queried simultaneously.[64] Illumina microarrays have been developed that assay over 27k, 450k, and 850k CpGs.

While most of the genome is relatively depleted of CpG, regions of the genome with a high density of CpGs exist and are referred to as CpG islands (CGIs). About 60% of CGIs overlap gene promoters.[65] 5-mC is involved in the epigenetic regulation of gene expression in that promoter methylation is often associated with a transcriptionally repressive state. Methylation within genes associates with a transcriptionally active state.[66] While intergenic methylation is suggested to affect gene expression through enhancer regulation.[67] The observation that patterns of methylation undergo frequent change during early embryogenesis, and that different tissues feature distinct methylation profiles, suggests that this modification is vital for normal cell function and development. Moreover, DNAm plays a role in many other cellular processes including: the silencing of repetitive and centromeric sequences, X chromosome inactivation, the formation of heterochromatin, and mammalian imprinting.[30] Thus, DNA methylation represents a critical component of the epigenome, and, unsurprisingly, can give rise to disease when it goes awry.

DNAm patterns are often dysregulated with disease in humans. For example, cancer can be marked by reduced levels of global DNAm with the hypermethylation of some regions including certain promoters and tumor-suppressor genes. Additionally, changes in DNAm have been implicated in several other diseases, including cardiovascular disease (CVD), neurological disorders, metabolic disorders, and autoimmune diseases.[68] DNAm can also be used to predict disease incidence; it has been found to change with obesity and these changes can predict the incidence of Type II Diabetes (T2D).[69] In order to tease out the direction of causation between DNAm changes and obesity, the statistical approach Mendelian Randomization (MR) can be employed.[70] Through the application of MR, the alterations to DNAm have been found to be a consequence of adiposity rather than a cause of it.[71] Consequential DNAm changes with disease could arise from disease-associated variants negatively affecting the DNAm patterns across the genome—referred to as the methylome,[72] or could be a symptom of an individual's internal environment, or some combination.

The methylome appears to change through lifestyle factors like smoking,[73] exercise,[74] diet,[75] as well as environmental factors like air quality,[74] stress,[76] early life events,[77] and many more. It was discovered, through studies of MZ twin pairs, that methylomes of young twins start out similar but diverge over time due to environmental factors or spontaneous stochastic errors in the DNAm maintanence.[78] This divergence appears enhanced if the twins were not living in the same environment.[79] The differences that arise over time among nearly genetically identical individuals reinforces the influence of epigenetic factors on phenotypic variation over the lifetime. It is thought that some of the changes observed in DNAm with time are due to imperfect copying of DNAm from parent to daughter DNA strands, leading to an accumulation of epigenetic errors seen both in disease development and aging.[68]

DNA Methylation and Aging

Among the nine hallmarks of aging are epigenetic alterations.[13] Numerous studies detail that, much like other biological levels, the epigenome displays a progressive loss of configuration with age. While some parts of this loss of configuration appear stochastic, others appear to occur in a directional manner and in specific regions of the genome, suggesting an underlying biological mechanism in the aging process.[80] Aging is correlated with a global decrease in DNA methylation, though many promoter-associated CpG islands (regions rich in CpG sites) have been observed to show hypermethylation with aging.[81] The global decrease in DNAm seen in aging is thought to be due to the decline in levels of DNMT1.[82,83] This suggests that distinct mechanisms may be at work, with hypermethylation resulting from programmed changes while hypomethylation is more a result of environmental and stochastic processes.[84] Interestingly, many of the changes seen in normal aging are also seen in cancer.

In fact, the finding that DNAm changes with age in healthy tissues was made serendipitously in 1994 by *Issa et al.* who, upon screening healthy cells, noted a CGI expected to be unmethylated was somewhat methylated while tumor cells were completely methylated.[85] They noted that this increase in methylation appeared dependent on the age of the participant whose cells were used. That DNAm is a hallmark of cancer cell has been well documented since its discovery in 1983,[86] but the finding that many of the alterations seen in cancer are also seen in healthy aging have led some to propose that the two processes share a common pathway.[68] Thus, a better understanding of the changes in the methylome during healthy aging may help inform the causes of cancer.

Several epigenome-wide association studies (EWASs) have characterized the robust changes to the methylome with age across multiple tissues,[87-91] and find that these changes appear to be highly tissue specific.[92] However, assaying DNAm in blood to gain insight into age-related changes can be complicated by the finding that proportions of white blood cells vary with age.[93] The proportion of CD8+ T cells decreases with age due to age-related thymic involution.[94] Because individual blood cells feature distinct methylation profiles,[95] not correcting for cell type proportions has the potential to confound observed age-differential DNAm patterns if not accounted for. In order to address this issue, Houseman *et al.*[96] developed a method to infer leukocyte identities and proportions in whole blood samples based on DNAm signatures. Employing this or more recently developed tools, including reference-free methods for assessing cell type mixtures in samples,[97,98] EWASs of age have been able to address the issue of possible confounding in their studies in tissues containing a mix of cell types, like blood.

Blood is a promising tissue in the development of a BoA because of the ease with which it can be assayed. A blood-based BoA complies with the American Federation for Aging

Research's requirement that an ideal BoA is able to be performed repeatedly on a subject without harm. Blood-based biomarkers have recently been developed that use observed epigenetic age-related changes to measure biological aging. These epigenetic biomarkers are highly predictive of chronological age, even being referred to as 'aging clocks,' and are promising candidates to also predict biological age.[90,99-101]

Linear DNA Methylation Changes with Age

Development of DNAm Aging Clocks

Methylation shows robust, genome-wide changes with age.[87,102,103] Capitalizing on these changes, researchers have created models that use DNAm information at a subset of age-related differentially methylated cytosines (aDMCs) throughout the genome, where DNAm has a significant linear relationship with age.[99,101,104] These models, termed DNAm aging clocks, output an estimate of chronological age based on DNAm patterns (DNAm age), and these DNAm ages have been found to correlate very closely with individuals' chronological ages. These clocks are generally built by using a linear regression algorithm trained against the chronological ages of sampled participants DNAm array data.[105] While tens of thousands of CpGs across the genome appear to be aDMCs,[106] relying on supervised machine learning methods such as a penalized regression (for example, least absolute shrinkage and selection operator, lasso, or elastic net) to reduce redundancy,[7] some clocks are able to narrow the number of CpGs used as input. These clocks, due to the different study populations and techniques used in their development, show variability not just in the number and identity of CpGs they contain but also in their ability to model chronological aging.

The first DNAm aging clock was developed in 2011 from saliva samples. It found that DNAm information from just two genes (NPTX2 and Tom1L1) predicted the age of an individual within 5.2 years.[90] A study published the following year found that DNAm in blood samples at the CpG islands of three genes, ELOVL2, FHL2, and PENK, strongly correlated with age—with the correlation for ELOVL2 in particular being 92%.[100] In the same vein, a study by Weidner *et al.*, using blood samples, found that data at just 3 CpGs, located in the genes ITGA2B, ASPA and PDE4, could predict chronological age within 5 years.[107] While these studies focused in on several genes, two additional clocks published in 2013 more broadly utilize DNAm arrays in predicting chronological age without tying the CpGs used as input to specific genetic pathways.

The first of these clocks, developed by Hannum *et al.*,[99] assayed DNAm using the 450k array on whole blood samples from 656 individuals. Using an elastic net regression model with an input of both DNAm data and clinical parameters including gender and body mass index (BMI), 71 CpGs across the genome were identified; these 71 sites accurately predicted chronological age within 3.9 years. In addition to this finding in blood, Hannum *et al.* found that their sites were somewhat predictive across other tissues tested (including breast, kidney, lung, and skin), with a 72% correlation between the age predicted by the model (DNAm age) and actual chronological age of the participant providing the sample. The genes linked to the 71 CpGs in the model occur within or near genes with known functions in aging, including: Alzheimer's disease, cancer, tissue degradation, DNA damage, and oxidative stress. Because cell type proportions were not included as a covariate in the Hannum clock, its predictive ability in blood is, in part, driven by age-related alterations in blood cell composition.

The second clock was constructed to be tissue-agnostic; Horvath,[101] using data from Hannum *et al.* and additional samples, developed an aging clock that functions across different tissues. Limiting the CpGs to those present on the 27k array, this clock also used an elastic net regression model on 8,000 samples from over 30 different tissues. This model selected 353 CpGs for Horvath's clock, which is able to accurately model chronological age within 3.6 years across tissues. There were, however, varying degrees of accuracy in the prediction depending on the tissue; the correlation between predicted and chronological DNAm age are quite low across: breast tissue (cor = 0.87), uterine endometrium (cor = 0.55), skeletal muscle tissue (cor = 0.70), and heart tissue (cor = 0.77). The fact that it works well across many tissues, in addition to the publicly availability of the pipeline, encouraged the widespread use of this early version of the Horvath clock in many studies.[108] Interestingly, though both the Hannum and Horvath clock appear highly accurate, they share only 6 CpGs in common.[109]

Other interesting finding from the Horvath clock includes the finding that in stem cells (both induced pluripotent and embryonic), DNAm age appears to increase with passage number—reflecting their limited proliferation and differentiation potential after several rounds of cell divisions. Additionally, these stem cell samples have a DNAm age near zero, again suggesting that this measure is capturing their biological age.[101] Horvath also introduced the concept of age acceleration (Δ_{age}), which refers to the positive difference between the predicted DNAm age and the actual chronological age of the participant. This term appears to be highly heritable when tested in MZ twin pairs, with that heritability decreasing over time, again supporting the findings of previous research supporting that non-genetic factors like one's environment and stochastic factors become more relevant over time. Interestingly, cancer tissues showed significant age acceleration relative to normal tissues, suggesting that this Δ_{age} measure could be useful as a biomarker. [101]

In an attempt to test the utility of the Hannum and Horvath clocks as potential BoAs, in 2015, Marioni *et al.*[110] assessed whether an individual's degree of accelerated aging, or Δ_{age}, predicted their risk of mortality. The Weidner clock[107] was also examined but after it was found to correlate poorly with chronological age, it was not used in further analysis. Using DNAm data from four longitudinal cohorts, a meta-analysis was performed to estimate the association between Δ_{age} and mortality. Participants with a Δ_{age} of 5 or greater (so whose predicted DNAm age was 5 years higher than their chronological age) had a 21% higher mortality risk than those without this degree of age acceleration. This increased risk of mortality persisted even after adjustments for hypertension, diabetes, CVD, and APOE e4 status—all well-characterized risk factors for early mortality. This finding, that a measure of accelerated aging can predict all-cause mortality better than chronological age alone, suggests that methylation is a meaningful indicator of biological aging.

Modeling Age-Related Phenotypes

Since this initial finding that linked the deviation between an individual's chronological and predicted DNAm age, or Δ_{age}, to their all-cause mortality risk, many studies have sought to also link this measure to risk of individual diseases as well as to note what phenotypic factors can influence it. Studies have linked accelerated epigenetic aging to: higher BMI;[111,112] early menopause;[113] risk of cancer incidence;[114] increased frailty;[115] risk of Down Syndrome;[116] stress;[117] obesity;[118] Alzheimer's disease,[119] and more. Conversely, being extremely long-lived seemed linked to relatively lower Δ_{age} in participants and their offspring.[120] While the Horvath and Hannum clocks have consistently proved their utility in modeling some age-related diseases and time to death, even in a large scale meta-analysis,[104] they appear to lack predictive ability in informing some disease risk and outcomes.[121]

It was hypothesized that the reason these clocks fell short in their predictive ability was that they were trained only on chronological age and did not take into account environmental exposures known to influence disease risk. These initial clocks, though accurate in modeling chronological age and predicting mortality risk, have been updated to incorporate more phenotypic and clinical data—dramatically improving their ability to model both lifespan and healthspan.[104,122,123] Interestingly, by leveraging the use of longitudinal data, future prediction can be made based on epigenetic age acceleration, including the future onset of lung cancer,[124] and mortality from both cancer and cardiovascular events.[125]

While many of the studies using DNAm aging clocks have been performed on cross-sectional data, comparing different individuals of different ages, longitudinal data, in which the same individuals are profiled over a period of time, offers the benefit of studying the trajectory of aging within individuals, and thus removing the possibility that observed associations are due to confounders in the individual process of aging. In fact, using longitudinal data to observe Δ_{age} over time has revealed its notable stability,[126] suggesting even that the measure becomes fixed at some point before adulthood. This finding was echoed by another study that observed a relationship between Δ_{age} and BMI, but that this was only observable in middle age, perhaps due to confounding factors (development and survival bias, respectively) in the extremely young and older cohorts.[118] Thus, these DNAm clocks represent the best candidate epigenetic BoA to date. Using these DNAm aging clocks in conjunction with longitudinal data may provide insight into individuals' aging processes and risk of adverse health outcomes.

Changes in DNA Methylation Variability with Age

Studies of the functional role of aDMCs in humans imply that these sites, while they may be tracking changes consistent with all-cause mortality risk, are not necessarily indicative of functionally relevant regions of the genome in their effects on gene expression changes with age.[106,127] Thus, while these CpGs are informative of chronological age, they may provide only limited insight into the transcriptional changes in an organism as it ages—limiting its use as a biomarker. This could be due to the methods used to identify the aDMCs in models like Horvath's,[101] in which informative aDMCs are identified by linear regression techniques and the residual between chronological and predicted age (Δ_{age}) used to model aging-related phenotypes and mortality risk. A recent study found this residual to be dependent on the population size used when identifying aDMCs, with a larger population size leading to a smaller residual; this suggestion, that the measure can vary based on the size of the population in which it is calculated, further challenges its use as a reliable BoA.[104,128]

Studies have reported that there is a change in the variability of DNAm with age; this suggests that DNAm profiles may be varying at different rates and in different directions across different individuals.[106,129] Considering the variability in DNAm age among people of the same chronological age,[130] between MZ twins in healthy aging[129] as well as the observation that heritability of DNAm age among MZ twins decreases over time,[101] modeling methylomic variability may be essential to capturing phenotypic variability. Perhaps a more informative model of aging would feature CpGs that show more variability over time instead of those reflecting linear age-related methylation differences. Several studies[106,131,132] have characterized such CpGs, termed age-related variably methylated cytosines (aVMCs), using diverse methods of identification (methods reviewed in [133]). These studies have sought to identify in both cross sectional and longitudinal data, individual CpGs particularly susceptible to methylomic drift.

These studies have found that these sites representing drift in increased variability appear to occur especially at age-associated CpGs,[134] and that these CpGs are often near genes involved in the aging process.[132] Genes near these sites were enriched in pathways involved in aging and development and, more specifically, aVMCs are enriched in transcription factors binding sites.[131] Additionally, aVMCs identified in blood appear to also reflect variability in other tissues (including colon, lung, and skin)[106]—suggesting that assaying blood could give insight into drift occurring throughout the body. In a study identifying regions of CpGs reflecting high variability in methylation across different tissues, irrespective of age, termed variably methylated regions (VMRs), VMRs specific to cell type and shared between them were identified—where different tissues sharing VMRs also shared common developmental origins.[135] This study also found that VMR networks were highly responsive to environment and are enriched in enhancers. This finding further suggests the functional link between changes at these sites, reflecting epigenetic drift, and resultant transcriptional drift and suggests that modeling variability at the molecular level could aid in modeling health.

Negative health outcomes, like cancer, and aging are both marked by an increase in stochastic DNAm drift.[136] This drift is not a directional like the linear DNAm changes used in building the aging clocks—with specific loci undergoing either hyper- or hypomethylation in a reliable manner. Additionally, this drift is not uniform across the genome, nor across individuals of the same age.[67] These observations, in addition to the finding that the degree of drift appears linked to the rate of proliferation of the tissue, implies that drift may be more a result of dysregulation in DNAm maintenance systems than an innate, programmed aging phenomenon. The rate and severity of this drift, while stochastic, does appear to follow some patterns in terms of the regions of the genome that it most affects as well as its severity increasing with some health factors like the degree of chronic inflammation in the body.[137] While this methylomic drift can have negative effects on the integrity of cells overall, it can be especially detrimental to stem cells, in which differential methylation can lead to differential gene expression of cells of the same tissue, termed transcriptional drift, as well as overall impaired stem cell function—one of the hallmarks of aging.[13]

Modeling variability may be key to understanding and quantifying the degree to which the integrity of the epigenome is disrupted in the process of aging, and in disease development. One study found that methylomic drift is more severe in MZ twins discordant for an autoimmune disease, with the diseased twins' methylomes reflecting more variability than those of the healthy twins.[138] Single-cell epigenetic profiling is another method that can be used to assay variability between samples taken from young and older participants. A study that aimed to characterize within-individual variability among histone modifications using single-cell ChIP-Seq noted that variability in histone modifications increased with age and that this increase in variability was driven by non-genetic factors.[139] This finding supports the interconnectedness of different levels of the epigenome in that both linear and variability changes with age are observed at the sequence level in DNAm and at the chromatin-level in histone modifications.

Summarizing the high dimensional variability occurring across thousands of sites into a single score indicating epigenetic drift may also provide insight into the gene expression patterns directly influenced by allowing this measure to be treated as a predictor of biological age in future analysis. DNAm, however, is limited in its ability to accurately model gene expression as it is often outperformed by data from histone modifications comprising the chromatin's local microenvironment.[140,141] This suggests that modeling the effect of drift on the epigenome and the resulting phenotypic dysregulation would be better served by integrating other epigenetic data more relevant to gene expression.

Chromatin, Histone Modification, and Expression Changes With Age

Interplay of Epigenetic Marks

Epigenetic modifications vary according to their level of influence but often contribute overlapping information.[142] For example, regions in the genome that feature differential methylation with age, specifically age-related hypermethylation at CGIs, also often feature repressive histone marks.[143] Conversely, active transcriptional start sites (TSSs) are often in nucleosome-depleted regions (NDRs) marked both by trimethylation of histone H3 at lysine 4 (H3K4me3) and the histone variant H2A.Z, which has been observed to repel DNMTs.[65,144]

This indicates that DNA methylation patterns are interrelated with patterns in histone modifications which can influence expression. DNAm patterns can also affect the binding of transcription factors (TFs), some of which can show a preference for either methylated or unmethylated DNA—another method through which it directly influences gene expression.[145] Taken together, these interactions are vital to proper chromatin structure and function.[142] It is hypothesized that these redundancies and layers of complexity and control employed by the epigenome are in place to guard against aberrant gene expression.[146]

These covalent histone modifications have the ability to induce changes in chromatin structure as well as recruit proteins important in chromatin regulation and gene expression.[37] In addition to their local contribution to transcription, histone modifications have important functional consequences in establishing global chromatin environments. These modifications are the basis of chromatin domains that contribute to whether DNA is transcriptionally accessible or inaccessible. Measures of chromatin accessibility seem to be interrelated with DNA methylation patterns as well. DNA methylation, which is associated with heterochromatin formation, appears to become depleted with age, lending to the theory that heterochromatin is lost in aging.[66,147] Because of their importance in chromatin function and their interrelatedness to DNA methylation, both histone modifications and chromatin accessibility are likely to also reflect age-related changes; they also have the potential to modify disease risk through their regulatory influence on gene expression and thus may also be informative in a model of aging.

Chromatin Structure

Aging chromatin is marked by disruptions in the normal maintenance of chromatin structure and its overall instability. With age, the protective ends of chromosomes composed of repeats of a short DNA sequence, or telomeres, shorten—one of the hallmarks of aging. This telomere attrition can lead to replicative senescence in which a cell will no longer divide, leading to lost information.[148] Additionally, the loss of DNAm globally with age across mammalian cells appears to occur especially at repetitive DNA sequences, where DNAm had previously exerted a silencing influence. Due to passive hypomethylation and a decrease in the number of histone proteins with age, the repressed heterochromatin microenvironment home to these repetitive sequences is not as well maintained in a condensed, repressive state and some regions are no longer silenced during aging.[80] This heterochromatin loss with aging can be observed both through the digestion of aged chromatin with MNase, in which the spacing between nucleosome becomes more irregular, as well as through the observation of less dense 30 nm fibers with age.[149] This loss of structure could cause an increase in genomic perturbances, including DNA breaks, translocations, and insertions.

This loss of the repressive DNAm on a larger scale is also thought to worsen observed genomic instability with age by allowing for the reactivation of typically silenced regions like inactive X-chromosomes and transposable elements (TE). Much of the repetitive, repressed DNA in the genome consists of retrotransposons, a class of transposable elements capable of moving around the genome;[150] this movement, can interrupt normal gene function and is even capable of causing cancer.[151] The increased mobilization of TEs is observed in cancer but has also been observed to occur to some degree even in healthy aging. These translocations, however, can be counteracted in mice by interventions known to slow the process of aging in model systems,[152] suggesting that interventions for healthy aging could mitigate some of the effects driven by chromatin aging.

Histone Modifications

Histone modifications have also been found to change with age, though these marks are more dynamic than DNAm, reflecting fluctuations in gene expression, so these observations are not as robust or well-characterized.[51] An additional mechanism driving aging chromatin's less compact structure is that the synthesis of core histone proteins decreases with age in human cells in vitro.[153] In fact, one study found in vitro fibroblasts from a 92-year-old had a 50% reduction in the synthesis of histones compared to those of a 9-year-old.[154] The factors driving this global histone loss, are poorly understood, though some studies suggest it may be linked to the shortening of telomeres.[153] Studies in mammalian model organisms and human cells have observed specific changes in histone modifications, including an increase in H3K9me3,[155] an increase in H4K20me3,[156] and an increase in H3K27me3 with age.[157] Overall, observed trends in histone modification changes in model systems indicate that there is an increase in the appearance of activating modifications and a decrease in repressive modifications with age, this is consistent with the observation in chromatin on the larger scale that the genome is becoming less compact with age.[158]

DNAm patterns and histone modifications are interrelated, and the directionality of their influence can change over the course of development.[142] Underlying the coordination between DNAm and histone modifications, CGIs featuring hypermethylation with age also featured H3K27 methylation—a repressive histone modification,[159] conversely, regions of the genome featuring histone modifications or variants associated with more open chromatin can work to shield CpGs sites within them from DNMTs.[160] Because of the interrelation between histone modifications and gene expression, it is likely that the age-related changes in histone modifications, in part, drive the widespread age-associated changes in gene expression.[127]

Objectives

The following chapters introduce three studies: 1) a study of the utility of the Horvath-derived Δ_{age} term in modeling an age-related disease longitudinally, 2) a study identifying aVMCs, using them to construct a score, and testing how the score contributes to modeling age-related phenotypes, and 3) a study that characterizes regions of the epigenome showing age-related differences in chromatin accessibility and histone modifications.

The first study (Chapter II), acknowledges that most research into the utility of the Δ_{age} term as a biomarker has been limited to cross-sectional (CS) studies which found the term had little predictive ability in disease incidence. The finding that the term could predict risk of all-cause mortality indicated that, theoretically, it should give some insight into overall health. It was hypothesized that leveraging the power of longitudinal data might remove any confounding factors leading to the lack of correlation between the term and health measures in CS data. Thus, this research is important because it was one of the first studies to characterize Δ_{age} longitudinally in its relationship to risk factors for type II diabetes.

The second study (Chapter III), aimed to create a novel model of methylomic variability that could be used as a biomarker indicating epigenetic drift. Recently, it has been suggested that understanding variability in aging would be important in characterizing both healthy aging and risk of disease incidence. Several studies have determined sites throughout the genome increasing in variability, and linked these sites to pathways important in aging. To test the relationship between variability at these sites and aging phenotypes, a score of variability must be developed. This research is important because it develops such a score and tests its link to aging-related phenotypes and mortality risk in a longitudinal cohort.

The third study (Chapter IV), characterizes age-related changes in the epigenome among two groups at extreme ends of the aging spectrum. While the levels of the epigenome interact and often complement one another in terms of the functionality of their modifications, DNA methylation is most often the sole subject of study because of its relative stability and the ease with which it can be assayed. This research is important because it aims to characterize aging at the level of broad histone modifications and chromatin accessibility in humans with the goal of uncovering what information the other levels of the epigenome can to contribute to the current understanding of molecular changes with age.

Finally, Chapter V is a discussion of the findings from the three studies as well predictions for future directions in the development of epigenetic biomarkers of aging. Suggestions of cutting-edge epigenetic profiling technologies and methodologies are discussed. Additionally, the importance of leveraging longitudinal studies in diverse human populations is addressed to ensure aging biomarkers can be widely used to improve care.

Chapter II. A Longitudinal Study of DNA Methylation as a Potential Mediator of Age-Related Diabetes Risk

Crystal D. Grant, Nadereh Jafari, Lifang Hou, Yun Li, James D. Stewart, Guosheng Zhang, Archana Lamichhane, JoAnn E. Manson, Andrea A. Baccarelli, Eric A. Whitsel, Karen N. Conneely

Adapted from an original publication, as cited:
Grant, Crystal D., et al. "A longitudinal study of DNA methylation as a potential mediator of age-related diabetes risk." Geroscience 39.5-6 (2017): 475-489.

Abstract

DNA methylation (DNAm) has been found to show robust and widespread age-related changes across the genome. DNAm profiles from whole blood can be used to predict human aging rates with great accuracy. We sought to test whether DNAm-based predictions of age are related to phenotypes associated with type 2 diabetes (T2D), with the goal of identifying risk factors potentially mediated by DNAm. Our participants were 43 women enrolled in the Women's Health Initiative. We obtained methylation data via the Illumina 450K Methylation array on whole blood samples from participants at three timepoints, covering on average 16 years per participant. We employed the method and software of Horvath, which uses DNAm at 353 CpGs to form a DNAm-based estimate of chronological age. We then calculated the epigenetic age acceleration, or Δage, at each timepoint. We fit linear mixed models to characterize how Δage contributed to a longitudinal model of aging and diabetes-related phenotypes and risk factors. For most participants, Δage remained constant, indicating that age acceleration is generally stable over time. We found that Δage associated with body mass index ($p = 0.0012$), waist circumference ($p = 0.033$), and fasting glucose ($p = 0.0073$), with the relationship with BMI maintaining significance after correction for multiple testing. Replication in a larger cohort of 157 WHI participants spanning 3 years was unsuccessful, possibly due to the shorter time frame covered. Our results suggest that DNAm has the potential to act as a mediator between aging and diabetes-related phenotypes, or alternatively, may serve as a biomarker of these phenotypes.

Introduction

Worldwide, the population aged 65 years and older is growing rapidly, with a 150% expansion projected over the next few decades.[6] Despite these recent global gains in life expectancy, age-related disease burden and the incidence of chronic disabilities remain high.[9] The healthspan, or years spent in good health, among the aging population remains highly variable, with some maintaining good health throughout their lives while others fall ill.[3] Age itself is the leading risk factor for the development of most diseases and conditions that drive morbidity and mortality and contribute to limited healthspan.[2,3] In many countries, age-related diseases like cardiovascular disease, diabetes, cancer, and neurodegenerative disorders, are among the predominant health problems faced by the population.[4]

A particularly widespread age-related disease adversely impacting the healthspan of millions worldwide is type 2 diabetes (T2D), which is now considered a global epidemic.[161] Due to population growth, increased longevity, and urbanization (which can promote physical inactivity and an unhealthy diet),[162] the global burden of T2D is expected to worsen over time as the prevalence increases from 415 million living with the disease in 2015 to an estimated 642 million in 2040.[161,163] There are many well-documented risk factors associated with the development of T2D, including: weight gain,[164] high body mass index (BMI),[165] high waist circumference,[166] ethnicity,[167] smoking status,[168] high fasting glucose,[169] high fasting insulin,[170] and age.[171,172] Diabetes contributed to approximately 5 million deaths globally in 2015[161] and is itself a risk factor for numerous other co-morbidities. Globally, ~50% of diabetic individuals are unaware of their condition, and subsequently are unaware of their increased risk of diabetes-related complications. Thus, a better marker of early T2D risk could provide mechanistic insights and facilitate earlier identification of high-risk individuals most likely to benefit from targeted lifestyle interventions.[161]

Differential susceptibility to age-related diseases can be attributed to biological differences between individuals, which work to modify disease risk (*reviewed by Feinberg*[12]). Among these biological differences are epigenetic changes, which arise without changes to the underlying DNA sequence and have the potential to modify disease risk through their regulatory influence on gene expression.[30] Additionally, because the major risk factors for T2D are lifestyle factors, such as diet and exercise behavior,[173] an epigenetic mechanism in which these factors can modify underlying genetic predisposition to disease incidence is highly plausible. DNA methylation (DNAm), the presence of a methyl group on the cytosine within a CpG dinucleotide, is the most studied epigenetic modification. The robust and genome-wide changes to DNAm observed with age make it an ideal biomarker of aging.[87,102,103,159,174] Biomarkers of aging are indicators of the biological age of an organism that predict its physiological functioning and disease susceptibility better than its chronological age alone.[22,23] Recently, highly accurate biomarkers of aging have been developed that capitalize on age-related changes to DNAm at a subset of CpGs across the genome to predict chronological age.[99,101] The approach of Horvath[101] uses methylation data from just 353 CpGs to form a multi-tissue, DNAm-based estimate of chronological age (DNAm age). Using DNAm age as a measure of biological age, the difference between a participants' DNAm age and their chronological age can be calculated. This measure is termed the participants' epigenetic age acceleration (Δ_{age}) and may proxy for the general health or rate of aging of the individual.[101] Instances in which the Δ_{age} term is positive indicate an epigenetic age that is higher than the participant's chronological age.

Many studies support the hypothesis that epigenetic Δ_{age} is associated with negative health outcomes, including increased risk of premature mortality,[104,110,114,125,130] early onset of age-related disease,[115,124] and changes in physical and cognitive fitness.[175] These findings indicate that Δ_{age} contributes more predictive information about these health outcomes than chronological age alone. This is consistent with the possibility that Δ_{age} may be acting to mediate the health outcome or risk of disease onset, but also with the possibility that DNAm age may be marking another biological process that is acting as a mediator. Consistent with the adverse health outcomes associated with positive Δ_{age}, a negative Δ_{age} can predict positive outcomes: centenarians in an Italian population and their offspring tended to have a DNAm age that was lower than their chronological age.[120] Taken together, these results support that epigenetics can be important in predicting both negative health outcomes and healthy aging.

Previous studies[176-178] have reported associations between site-specific methylation differences and T2D as well as related phenotypes across several cell types and tissues. Our study aims to assess the potential of 5mC as a mediator between aging and age-related T2D risk phenotypes. To model age-related 5mC patterns, we focus on a well-studied methylation-based biomarker of aging[101] which identified 353 CpG sites as being the most predictive in modeling chronological age. We take advantage of a longitudinal study spanning 16 years to 1) characterize the changes to participants' Δ_{age} over time, and 2) characterize the contribution of DNAm age and Δ_{age} in modeling T2D susceptibility. Given that many T2D risk factors (including high BMI, waist circumference, and fasting glucose and insulin levels) reflect age-related changes, a measure of biological aging may help predict which participants are at a higher risk of T2D incidence throughout the study. Though we do not have the power to model incidence of clinical T2D in our sample, the longitudinal nature of this study allows us to model changes to phenotypes intermediate between age and disease risk. We will use the DNAm-based measure of biological age as a proxy for genome-wide DNAm and other age-related biological processes that may underlie age-related disease risk. We aim to inform future studies by assessing the utility of genome-wide methylation changes and other biological processes as potential mediators between age and risk factors for and indicators of T2D (subsequently referred to as ‘diabetes-related phenotypes’).

Methods

Study population and study design

Participants are a subsample from the 68,132 women who took part in the Women's Health Initiative (WHI) Clinical Trials (CT) Cohort. The WHI was a national study which sought to investigate interventions and treatments for the prevention and management of common causes of morbidity and mortality among older women.[179] All WHI participants were post-menopausal women, aged 50 to 79 years at the time of enrollment, with minority women recruited at the same proportion found in the U.S. population at the time.[180] Women in the WHI were also more likely to be overweight, with three quarters of the women overweight or obese at the time of enrollment.[180]

The study began in 1993 with participants completing questionnaires detailing their: sociodemographic information (including their age and race), current health behaviors (including weekly physical activity and smoking behavior), and current health status (any disease diagnosis and medications or supplements currently prescribed). Participants also attended scheduled clinic visits in which anthropometric measurements were assessed, including: weight, height, and waist circumference; from these measures, body mass index was calculated. Additionally, a 6% minority oversample of participants had blood drawn during these clinic visits from which insulin, glucose, triglyceride, and high-density lipoprotein cholesterol concentrations were measured and buffy coat was archived. "Epigenetic Mechanisms of PM-Mediated CVD Risk" (WHI-EMPC) measured DNAm on a genome-wide scale using DNA extracted from the archived buffy coat in a stratified, random sample (2,200) of the participants who were examined between 1993 and 2001. Among a subset (200) of the 2,200 participants, WHI-EMPC also measured DNAm in buffy coat archived at a second timepoint on average 3.3 years later. Subsequently, a "Longitudinal Study of DNA

Methylation as a Mediator between Age and Cardiovascular Risk" (AS #534) measured DNAm in buffy coat archived at the third timepoint, on average 16.1 years after the first, for a subset (43) of the 200 participants who were followed up as part of the Long Life Study (LLS). These 43 participants are included in our study and described in Table 1.

Data cleaning

Chronological age of the participants was approximated at each timepoint as participant's self-reported age at screening (in years) + 0.5 + number of days between screening and blood sampling / 365.25. Phenotypic measures include: BMI measured as weight (kg) divided by the square of height (m^2), waist circumference (cm), fasting glucose (mg/dL), and fasting insulin (μIU/mL). Homeostasis Model Assessment of Insulin Resistance, termed HOMA-IR, was calculated using the following equation: Insulin (μU/mL) × Glucose (mg/dL)/405.[181] The ratio of plasma triglycerides (mg/dL) to high-density lipoprotein cholesterol concentration (mg/dL), termed TG/HDL-C ratio, was calculated.[182] Lastly, the triglyceride-glucose index, termed the TyG index, was calculated using the following equation: ln[Triglycerides / (Fasting glucose / 2)].[183] Both TG/HDL-C and TyG were included as markers of insulin resistance.

Three unrealistic data points believed to be entered in error were removed. These included BMI measures below 15 kg/m^2, or above 55 kg/m^2; these values were >2 SD away from the participant's mean throughout the study and were flanked by more moderate values measured within 4 years. Additionally, a waist measurement above 150 cm was also removed, as it was 1.7 SD from the participant's mean and was flanked by more moderate measurements within 6 years. Phenotypic data collected from within 30 days of a blood draw were assumed to approximate data that would have been collected at the time of the draw. Additionally, several waist circumference measurements originally recorded in inches were converted to cm, with 1 inch equivalent to 2.54 cm. Insulin measures at the first and second timepoints were ascertained using different but similar methods. All insulin testing for the first timepoint used the radioimmunoassay (RIA) method. For some participants, the second timepoint used an automated ES300 analyzer. Because ES300 and RIA methods gave comparable results at insulin levels below 60 µIU/mL, and because all participants had insulin levels below this cutoff for the first two timepoints, the insulin results were combined into a single variable. The method for measuring insulin concentrations changed again for the third timepoint with the Roche Elecsys 2010 Immunoassay analyzer being used. Measures from the third timepoint were recorded in pmol/L and were converted to µIU/mL, with 6 pmol/L equivalent to 1 µIU/mL.[184] Self-reported smoking behavior, originally recorded as: "Never Smoked," "Past Smoker," and "Current Smoker" were recoded to "Never Smoked" and "Smoked" due to only one participant being classified as a "Current Smoker."

Alcohol intake, total caloric intake, and family history of diabetes were self-reported at the start of the study. Alcohol intake reported was weekly intake of alcoholic beverages. This includes the number of servings per weeks of beer, wine, and/or liquor based on a serving size of 12oz for beer, 6oz for wine, and 1.5oz for liquor. Entries ranged from 0 to 12.4 drinks per week (mean=1.5) with missing data for one participant. Total caloric intake was reported in kilocalories per day, ranging from 660.1 to 3455.2 (mean=1487.4) with data missing for one participant. Participants whose energy intake estimates suggested that they were not properly completing the food frequency questionnaire (i.e. those with daily intake less than 600 kcal or greater than 3500 kcal), were excluded (N=2).[185] In characterizing family history, participants were asked: 'Did your mother, or father, or full-blooded sisters, full-blooded brothers, daughters, or sons ever have sugar diabetes or high blood sugar that first appeared as an adult?'

Participants' responses were: 'Yes' (11 participants), 'No' (31 participants), or 'Unsure' (1 participants). For the model, participants who answered either 'No' or 'Unsure' were combined into 'No or Unsure.' Incident diabetes and incident diabetes treatment occurring within the study period were also characterized as part of the sensitivity analysis. Incident diabetes was defined, according to standards set by the American Diabetes Association,[186] as anyone who fasted for 8 or more hours and has a glucose measure ≥ 126 mg/dL, or anyone who fasted for fewer than 8 hours and has a glucose measure ≥ 200 mg/dL (4 participants). Timepoints occurring after a participant indicated they were prescribed medication to treat diabetes were considered incident treatment with an antidiabetic agent (4 participants).

DNA methylation data

DNA was extracted from buffy coat from participants at each timepoint. DNA (500 ng) was used for the bisulfite conversion with the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA, USA), following the manufacturer's protocol. Once converted and amplified, DNA (15 µL) was fragmented, and hybridized to the Infinium HumanMethylation450 Bead Chip (Illumina Inc., CA, USA). DNAm profiles of >485,000 cytosine-guanine (CpG) sites were measured using the Infinium HumanMethylation450 BeadChip at the Northwestern University Genomics Core Facility in two batches, with the first two timepoints run as part of WHI-EMPC and the third run as part of AS #534. DNA methylation was subject to quality controls: excluding probes targeting CpG sites on the Y chromosome, probes with detection p-values > 0.01 in > 10% of samples, and samples with detection p-values > 0.01 across in > 1% of probes. 484,220 CpG sites passed this quality control step and were eligible for further analysis. Two control DNA samples on each BeadChip were used to assess reproducibility, and duplicates from the first batch were run with the second to account for batch effects. Methylated (M) and unmethylated (U) signals were used to compute estimates of the methylation proportion, β-values, (β=M/(U+M)). Next, beta-mixture quantile normalization (BMIQ), was performed to reduce technical variation and intra-array bias between differing types of probes.[187] Lastly, ComBat, which employs an empirical Bayes method to adjust for batch effects, was used to adjust for differences between the two batches.[188]

Measures of DNA methylation age and Δ_{age}

DNAm age at each timepoint was calculated using the methylation profiles from 353 CpGs and the R pipeline detailed in.[101] The difference between DNAm predicted age and the chronological age of each participant at each of the three timepoints, termed "age acceleration" (Δ_{age}), was calculated at each point.

Testing for association between age and DNA methylation

Using the R package CpGassoc,[189] we performed an epigenome-wide association study (EWAS) to test for association between chronological age and DNAm. For each CpG site we fit a linear mixed model that included a random effect for each participant to account for the repeated measures within participants, and self-reported ethnicity and Illumina chip and row as covariates.

Testing for association between Δ_{age} and diabetes-related phenotypes

Phenotypes analyzed included seven diabetes-related phenotypes: fasting insulin and glucose, HOMA-IR, BMI, waist circumference, TG/HDL-C ratio, and TyG index. Using the R package nlme,[190] we fit longitudinal, mixed effect models with the phenotype as the outcome and a random effect for participants. For each phenotype, two models were fit: the first regressed each phenotype on chronological age and relevant covariates, while the second regressed each phenotype on both chronological age and Δ_{age}, in addition to other covariates. Our goal was to assess whether the additional term Δ_{age} associates independently with the phenotype, indicating that Δ_{age} contributes to our ability to model the phenotype. Thus, for participant (*i*) at timepoint (*j*), the following models were fit:

Model 1:

$$\begin{aligned}&\textit{Diabetes-related phenotype}_{ij}\\ &= \beta_0 + \beta_1 age_{ij} + \beta_2 ethnicity_i + \beta_3 smoking_i + \beta_4 fasting\ hours_{ij} + \nu_i + \varepsilon_{ij}\end{aligned}$$

Model 2:

$$\begin{aligned}&\textit{Diabetes-related phenotype}_{ij}\\ &= \beta_0 + \beta_1 age_{ij} + \gamma\Delta_{age_{ij}} + \beta_2 ethnicity_i + \beta_3 smoking_i + \beta_4 fasting\ hours_{ij} + \nu_i + \varepsilon_{ij}\end{aligned}$$

where $\Delta_{age_{ij}}$ represents age acceleration for individual *i* at time *j*, ν_i represents a random effect (individual-specific error term) for individual *i*, and ε_{ij} represents the error term for individual *i* and timepoint *j*. Significance of the age acceleration coefficient γ in the second model was taken to suggest that the relationship between chronological age and that phenotype could potentially be mediated by methylation or a related biological process, or that Δ_{age} could serve as a biomarker for this phenotype. To adjust for potential confounding, ethnicity, cigarette smoking, and fasting hours (where relevant), were included as covariates. Sensitivity analysis were performed with several well-known T2D risk factors added individually as covariates, including total energy expenditure, total caloric intake, alcohol intake, and family history of diabetes.

Estimation of blood cell proportions based on DNA methylation

A complication in analysis of whole blood samples in aging studies is that cell proportions in whole blood change with age,[93,96] and different subpopulations of blood cells feature different methylation patterns.[95] Together, these can confound the relationship between DNAm and aging, since it is difficult to distinguish DNAm changes in whole blood with age from DNAm changes in blood with disease development if the model does not explicitly account for differences in cell proportions.[191] Houseman's regression-based method[96] was used to estimate the composition of white blood cells in whole blood using DNAm array data. This tool uses DNAm data from 500 CpGs found to be most informative of white blood cell (WBC) type in whole blood. The tool constrains the sum of the 6 blood type proportions (CD4+ helper T cells, CD8+ cytotoxic T cells, granulocytes, monocytes, natural killer cells, B cells) to 100%, then fits a regression model to the DNAm data at the 500 sites. This allows for the estimation of the 6 WBC proportions, which were then included as covariates in the comparisons of Models 1 and 2, with granulocyte proportions excluded as the reference category.

Testing for change in Δ_{age} over time

A mixed effects model, with the year of the participant's clinic visit as a fixed effect and a random effect for participants, was used to test whether there was significant change in Δ_{age} over time.

Results

Sample characteristics

The sample characteristics of our population are detailed in Table 1. Participants were 43 post-menopausal women, between 50 and 76 years of age at enrollment, with a mean age of 61.5 years (sd=6.9). A plurality of our sample was non-Hispanic white (41.9%), about a third were African American (32.6%), and about a quarter were Hispanic or Latino (25.6%). Self-reported smoking behavior indicated that 23 participants (54.8%) were either current or previous smokers; 19 participants report having never smoked (45.2%), while one participant failed to respond. Longitudinal DNAm data were available for three timepoints with the second and third timepoints occurring on average 3.3 and 16.1 years after the first, respectively. At baseline, none of the participants were being treated for diabetes. The distributions of the seven diabetes-related phenotypes in our population are shown in Supplementary Fig. 1.

DNA methylation changes with chronological age

Using DNAm array data, we performed a longitudinal epigenome wide association study as proof of concept that many CpGs display differential methylation associated with participants' estimated chronological ages, a pattern which been well-established in many other datasets (e.g. Alisch et al. 2012; Bollati et al. 2009; Christensen et al. 2009; Teschendorff et al. 2010; Xu and Taylor 2014). In our data, 232 sites showed significant changes with age according to the Holm step-down Bonferroni procedure (p<1.0E-7), while 3,064 sites were found significant by the Benjamini–Hochberg procedure (FDR<.05). Top CpGs are listed in Supplementary Table 1. Supplementary Fig. 2 features a Manhattan plot of p-values reflecting the association between methylation and chronological age. Our results appear consistent with those reported by *Xu and Taylor*,[159] who identified 749 high confidence age-related CpGs in >1000 individuals.

Supplementary Fig. 3 demonstrates a high correlation (r=0.74) between t-statistics across the two studies. Additionally, 11 of our significant sites overlap with the 353 CpGs that make up the epigenetic clock.[101]

DNA methylation age estimates over time

Participants' chronological ages show high correlation with the predicted DNAm ages of our participants (r=0.89) (Fig. 1). The difference between this predicted age and the chronological age of each participant at each of the three timepoints, termed Δ_{age}, is calculated at each point. DNAm age at enrollment ranges from 43.2 to 84.5, while Δ_{age}, at enrollment ranges from -12.3 to 9.0. The median Δ_{age} value across participants is -4.5. Δ_{age} is negative for 109 of the 129 measurements (84.5%), which is consistent with previous reports showing that women tend to have lower Δ_{age} than men.[99,121] The average Δ_{age} at the first timepoint is -3.5 (sd=4.4), -4.9 (sd=4.5) at the second timepoint, and -4.6 (sd=5.2) at the third timepoint (Table 1, Supplementary Fig. 4). According to a Shapiro-Wilk normality test, Δ_{age} is normally distributed at timepoints 1 (p=0.16) and 2 (p=0.87), but not timepoint 3 (p=0.0033). However, with the removal of a single individual with an extreme Δ_{age}, values for timepoint 3 are consistent with a normal distribution (p=0.94).

Δ_{age} is not significantly associated with smoking status (p=0.51) in our data. It is also not significantly associated with chronological age (r=-0.14, p=0.13) (Supplementary Fig. 5), though the negative correlation is consistent with previous reports.[104,130,175] It does vary by ethnicity, with the Hispanic/Latino group having a smaller Δ_{age}, but this difference is not statistically significant in our sample (p=0.39). This observation agrees with recent findings that Hispanic/Latina women participating in the WHI study have a lower Δ_{age} compared to WHI Caucasians,[121] though our study did not have power to detect a significant difference.

Stability of Δ_{age}

Within individuals, very little change in Δ_{age} is observed over time, suggesting that the value of age acceleration remains roughly constant over time among our participants (Fig. 2). On average, Δ_{age} showed a 0.041 decrease each year, which does not differ significantly from a change of zero (p=0.25) (Supplementary Fig. 6). To identify individuals whose Δ_{age} changed significantly during the study, each of the 43 participants' DNAm age was regressed on their chronological age. The mean slope of this regression was close to 1 (mean=0.96, SD=0.29), suggesting that on average, DNAm age increases at a similar rate to chronological age. Five participants (10, 26, 27, 33 and 34) were at least 1.5 standard deviations from the mean, with slope values of 0.46, 1.41, 0.52, 1.71, and 2.02 respectively. To assess whether these changes in Δ_{age} could be influenced by changes in blood cell proportions, we regressed each of six estimated cell type proportions onto the year of the participant's visit, and found that cell proportions did not change significantly over the course of the study (Supplementary Fig. 7).

DNAm age acceleration associates with several diabetes-related phenotypes

Results from our models of diabetes-related phenotypes are listed in Table 2. Δ_{age} has a significant positive association with fasting glucose (p=0.0073), BMI (p=0.0012), and waist circumference (p=0.033). Using a Bonferroni-corrected α of 0.0071 to adjust for the 7 phenotypes tested, the association remains significant for BMI and near-significant for glucose. To assess the robustness of our results to inclusion of covariates, we performed sensitivity analyses that added the following covariates to the model: alcohol intake, total caloric intake, family history of diabetes, incident diabetes during follow-up, and incident treatment with antidiabetic agents. Supplementary Table 2 shows that the addition of each covariate produces similar results to our baseline model. Furthermore, inclusion of a covariate for participants taking medication for incident diabetes suggests that, in addition to Δ_{age} contributing significantly to modeling of BMI, it also contributes significantly ($p<.0071$) to modeling fasting glucose among our participants.

Supplementary Fig. 8 reflects measurements of BMI over the 16-year study period for our participants. Of the five participants with extreme Δ_{age} slope values, three participants (10, 34, and, to a lesser extent, 27) also had extreme changes in BMI during the study. This BMI fluctuation could, perhaps, be linked to changes in DNAm and Δ_{age}. To test whether the relationship between Δ_{age} and BMI, fasting glucose, and waist circumference were driven by these five participants, we removed them in a sensitivity analysis. Supplementary Table 3 includes the results of this analysis in which it appears that our findings are driven by the participants with dynamic Δ_{age}, since the effect sizes decrease substantially upon their removal compared to the original results in Table 2. This loss of an association with the removal of the most dynamic participants suggests that the association may be driven by within-person changes in Δ_{age} and BMI, rather than static differences between individuals.

Replication study in a second WHI subsample

A subset of 200 women from a stratified, random sample of 2,200 WHI-CT participants had two DNAm measurements assessed as part of WHI-EMPC. Our 43 participants with three DNAm timepoints are part of this subset of 200; we attempt to replicate our findings in the remaining 157 participants who had two DNAm timepoints on average 3.7 years apart. The replication cohort's ethnic make-up is fairly similar to our participants, with: 55.4% Non-Hispanic, White, 19.6% Black or African American, 15.92% Hispanic/Latino, 4.46% Asian or Pacific Islander, 3.18% American Indian or Alaska Native, and 1.27% Other. Smoking behavior had a high rate of missingness (85.7% of participants did not provide data on their smoking habits), and thus was not included in regression models. The sample characteristics of our replication population are detailed in Supplementary Table 4. The replication cohort mirrored our finding of female participants having lower DNAm age than their chronological age (mean Δ_{age} is -4.30 years in our data and -3.87 in the replication cohort). However, while the correlation between Δ_{age} and chronological age was not significant in our analysis of 43 participants (r = -0.14, p=0.13), analysis of this larger sample yielded a significant negative correlation (r = -0.20, p=3.9E-6, Supplementary Fig. 9).

Results of the regression of diabetes-related phenotypes on age and Δ_{age} are shown in Supplementary Table 5. We found that Δ_{age} did not contribute significantly to models of our seven diabetes-related phenotypes in our replication group. To test whether the significant findings in the original dataset were due to its longer timespan relative to the replication data, we censored the original dataset so that only the first two timepoints were included in the regression. In Supplementary Table 6, we see that the originally reported associations with BMI and glucose disappear when only two timepoints are used, marked by a substantial drop in the estimated effect size. This suggests that these results may depend on the ability to observe individual changes over a sufficiently long time period.

Discussion

This study supports previous findings on the utility of DNAm-based biomarkers of age in modeling health outcomes. We analyzed longitudinal DNAm data in order to capture the relationship between participants' changes in DNAm age over time and diabetes-related phenotypes. We found that age acceleration contributes significantly to models of diabetes-related phenotypes among our 43 participants. Epigenetic age acceleration is positively associated with longitudinal changes in participants' body mass index. Additionally, epigenetic age acceleration shows a suggestive association with longitudinal changes to participants' glucose, narrowly missing our Bonferroni cutoff for significance. Glucose does in fact reach significance in our sensitivity analysis in which a covariate for incident T2D treatment is included (p=0.0054). This indicates that age acceleration may contribute to longitudinal models of fasting glucose and that more research should be done with a larger sample. Age acceleration does not appear to significantly contribute to longitudinal models of waist circumference, insulin, HOMA-IR measurements, TG/HDL-C ratio, or TyG index. These findings give us leads into which aspects of diabetes-related phenotypes may feature an important epigenetic component. The utility of epigenetic-based biomarkers is that they can offer a more personalized model of an individual's health status than age alone, though this may not be true for all phenotypes. This is evident in the result that Δ_{age} contributes to models of BMI and fasting glucose but that chronological age appears to be a better predictor of fasting insulin, HOMA-IR, TG/HDL-C ratio, and TyG index.

An intriguing finding is that, for most of our participants, a DNAm-based measure of age acceleration remains stable over the course of the study. This indicates that participants who displayed accelerated biological age at the start of the study were likely to display the same degree of epigenetic age acceleration 16 years later. The dynamics of Δ_{age} over time have not been extensively characterized, but this observed stability of Δ_{age} over time among adults is consistent with findings in previous longitudinal studies of age acceleration.[126,175] Additionally, we found that Δ_{age} exhibits a negative correlation with chronological age, which is consistent with previous reports.[104,130,175] While this relationship was not significant in our initial sample, it reaches significance in our larger replication cohort. While this could suggest a non-linear relationship between DNAm age and chronological age over the life course, Δ_{age} did not change significantly over time for the majority of individuals in our study. Thus, the negative correlation appears to result from between-individual differences, and may reflect a selection bias due to biologically "younger" individuals being more likely to survive to old age.[130]

A recent study reported that Hispanic/Latinos from the WHI feature a significantly lower epigenetic age acceleration compared to Caucasians.[121] In our study, Hispanic/Latinos also featured a lower Δ_{age} compared to Caucasians and African Americans, but this was not significant due to our small sample size. Additionally, our findings, that Δ_{age} did not associate significantly with several diabetes-related phenotypes, have been corroborated by another study of Δ_{age} among WHI participants; however, in contrast to our findings, this study did not find a significant association between Δ_{age} and BMI or glucose.[121] Reasons for this difference could lie in our use of longitudinal data over 16 years, while most previous studies of epigenetic age acceleration have relied on cross-sectional data.

A recent publication, which used longitudinal data from an overlapping set of subjects within the WHI, observed a significant association of age acceleration with individual changes in BMI over a 3-year study period.[111] Another study, using longitudinal methylation data, found that an increase in the BMI is significantly associated with an increase in age acceleration.[118] These findings suggest that a longitudinal approach to modeling diabetes-related phenotypes may allow for the detection of associations previously not possible with a cross-sectional study.

The increased ability to detect association between DNAm and the phenotypes tested can be attributed to the length of time between repeated measures. The 3-year study period may explain why our replication sample, though larger, did not reflect the associations between age acceleration and diabetes-related phenotypes noted in our 16-year study.

Longitudinal studies provide a powerful means to identify phenotypic changes associated with within-person changes in DNA methylation, while avoiding potential confounding due to between-person differences. Sensitivity analyses revealed that our observed association between BMI and Δ_{age} was driven by within-individual differences in the participants with the most dynamic Δ_{age} and BMI over the time period studied. We also noted that Δ_{age} was relatively stable over time for most individuals. Based on these observations, to maximize within-person variation in predictors and phenotypes, future longitudinal studies of DNAm and age-related phenotypes should strive to focus on the age ranges that are most dynamic with respect to the phenotypes of interest, and incorporate the widest possible study duration within the relevant age range. In addition, a previous finding that events like menopause can accelerate biological aging in blood[113] imply that perhaps studies of DNAm and/or biological aging could benefit from focusing on post-menopausal women.

Our study had several limitations. Our population of only post-menopausal women potentially limits the generalizability of our findings. More research into the contribution of Δ_{age} to health outcomes in both men and women, and in participants across different age groups is necessary. Furthermore, a disproportionally high number of participants enrolled in the WHI are obese, potentially limiting generalizability to non-obese populations. Additionally, data on smoking behavior, alcohol consumption, exercise habits, and ethnicity were self-reported and thus could be biased, potentially affecting our results. Data on time spent exercising per week was unavailable for the third timepoint, and was thus not included in our models. Because physical activity is known to protect against the development of diabetes,[192] this may inflate the importance in the contribution of DNAm to disease development. Lastly, T2D incidence was included as a covariate in the sensitivity analysis and not analyzed as an outcome because only 4 participants in our study developed T2D over the 16-year time period—which would limit our power to detect associations with disease incidence. Because of this limitation, our focus was on phenotypes associated with the incidence of T2D rather than the incidence itself.

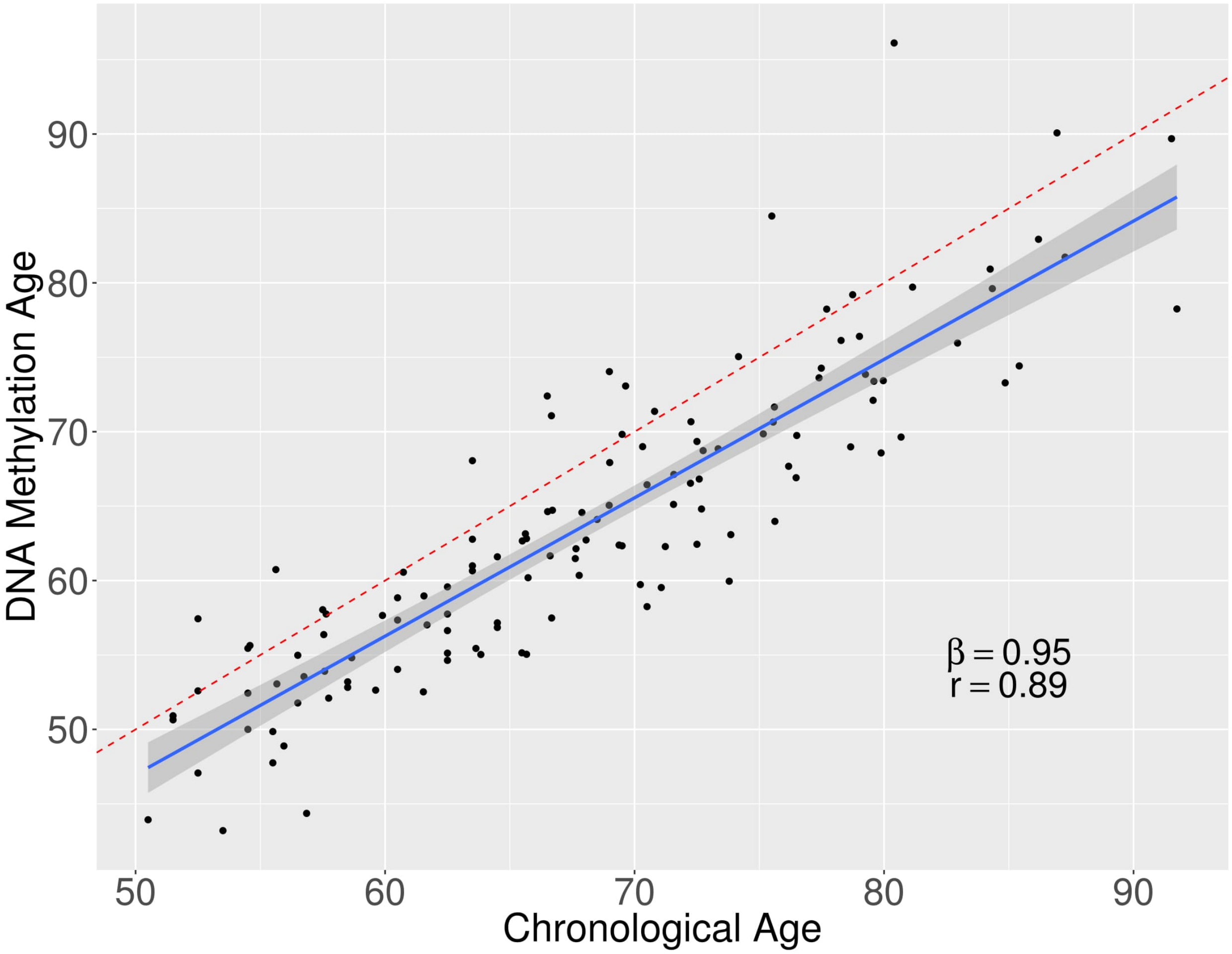
Finally, while our study benefits from a longitudinal design with DNAm spanning an average of 16 years within subjects, the number of subjects is small. Larger studies will be needed to confirm the associations reported here and to investigate mechanisms underlying the associations. Our results are consistent with a scenario in which the relationship between age and these diabetes-related phenotypes may be mediated by DNAm or a related process. However, much larger studies are required to tease out causality in the relationship between epigenetic aging rates and phenotypes associated with diabetes such as high BMI. Recent cross-sectional publications have used Mendelian randomization approaches to assess causality between DNAm and obesity from whole blood.[69,71] Their findings suggest that the majority of obesity-associated differences in DNAm patterns may be a result, rather than a cause, of the development of obesity. Regardless of the direction of causality, our results and others support the potential of DNAm and epigenetic factors as candidates to develop biomarkers for diabetes-related phenotypes.

Conclusions

Diabetes is associated with genetic, lifestyle, and environmental factors, suggesting that the epigenome may be important in determining both susceptibility and progression of the disease. While numerous past studies have noted small scale DNAm changes that accompany diabetes risk and progression, our findings speak to the utility of genome-wide methylation changes in modeling phenotypes associated with diabetes. This contribution of Δ_{age} in modeling diabetes phenotypes also speaks to the ability of DNAm to serve as a potential mediator of the relationship between aging and the phenotypes associated with age-related disease, or alternatively as a biomarker. We believe this pilot study can inform future studies of DNAm-based biomarkers and their potential to predict phenotypes associated with disease.

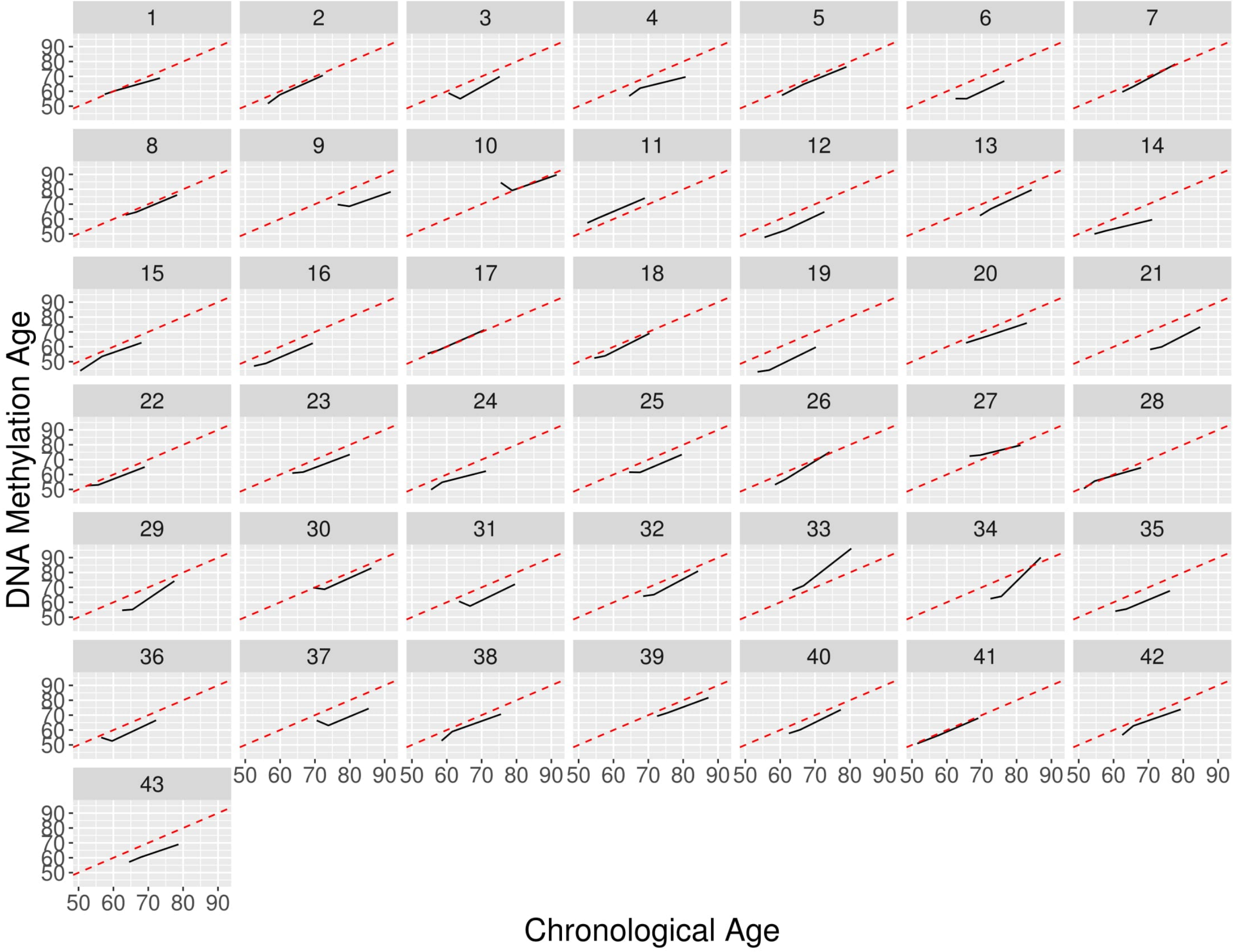
Tables and Figures

Fig. 1 Chronological Age (x-axis) vs. DNA Methylation Age (y-axis)



Each point shows chronological age and DNAm age for one participant at one of the three timepoints. The dotted red line represents the equivalence line, meaning there would be perfect agreement between the computed DNAm age and the approximated chronological age. The blue line represents the regression line obtained from a regression of DNAm age on chronological age with random effects to account for repeated measures within subjects. The shaded grey region around the blue line represents a 95% confidence interval of the regression line

Fig. 2 Chronological Age (x-axis) vs. DNA Methylation Age (y-axis) for Each Participant



Each subplot represents one participant; a solid black line connects the participants' three measures of Δ_{age} across the three timepoints. The dotted red line represents a line of slope = 1, reflecting perfect agreement between DNAm age and chronological age. A participant's black line being nearly parallel to the dotted red line indicates little change in Δ_{age} within a participant over the course of the study, while a black line with a slope other than 1 would reflect changes in Δ_{age} over time. Figure 1 provides a composite view of these data combined across all 43 subjects

Table 1. Demographic and clinical characteristics of study population (N=43)

Variable	**Mean +/- SD or percentage**			
	Baseline (SD)	**Follow-up (SD)**	**LLS (SD)**	**# missing obs.**
Chronological age (years)	61.52 (6.94)	65.05 (6.77)	77.48 (6.48)	0
DNAm age (years)	58.05 (8.05)	60.19 (6.93)	72.85 (7.92)	0
Δ_{age} (years)	-3.47 (4.36)	-4.86 (4.47)	-4.56 (5.20)	0
BMI (kg/m^2)	29.02 (5.23)	29.32 (4.47)	28.34 (5.88)	4
Fasting glucose (mg/dL)	94.88 (8.47)	95.47 (14.22)	100.19 (18.76)	0
Fasting insulin (µIU/mL)	12.00 (5.21)	13.04 (7.58)	18.94 (15.10)[a]	6
HOMA-IR	2.82 (1.23)	3.17 (2.37)	5.01 (4.56)[a]	6
TG/HDL-C ratio	3.09 (1.65)	2.92 (1.91)	2.08 (1.19)	1
TyG index	8.85 (0.47)	8.84 (0.46)	8.57 (0.49)	0
Waist circumference (cm)	87.64 (12.29)	88.87 (11.55)	89.20 (13.37)	10

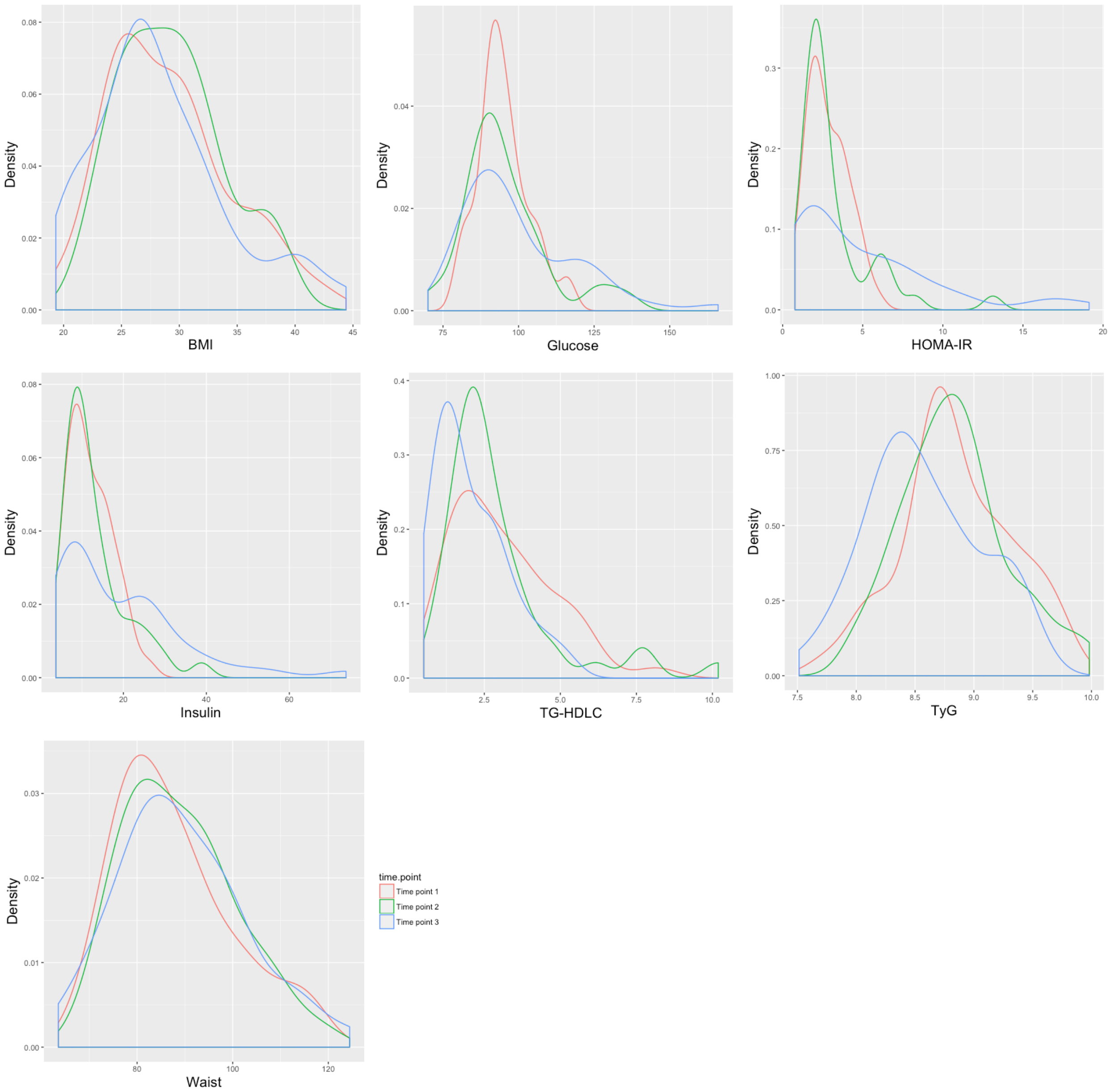
[a] LLS insulin measures were obtained from a different analyzer from the baseline and follow-up measures and units were converted from pmol/L to µIU/mL. We observe higher values and standard errors for this measure. These observed differences between timepoints could reflect a true increase in fasting insulin with age, or could be due to differences in measurement

Table 2. Multivariate regression analysis of diabetes-related phenotypes on age and biological age acceleration

Phenotype	Model 1 Coefficients on chronological age		Model 2 Coefficients on chronological age		Model 2 Coefficients on Δ_{age}	
	Est. (SE)	P-value	Est. (SE)	P-value	Est. (SE)	P-value
BMI	-0.046 (0.030)	0.13	-0.032 (0.029)	0.27	0.29 (0.087)	0.0012*
Fasting Glucose	0.24 (0.14)	0.081	0.30 (0.13)	0.027	0.97 (0.34)	0.0073
Fasting Insulin	0.28 (0.10)	0.0078	0.30 (0.10)	0.0050*	0.26 (0.24)	0.28
HOMA-IR	0.084 (0.028)	0.0043*	0.091 (0.029)	0.0022*	0.089 (0.065)	0.18
TG/HDL-C ratio	-0.051 (0.015)	0.0013*	-0.048 (0.016)	0.0029*	0.048 (0.040)	0.24
TyG index	-0.014 (0.0045)	0.0023*	-0.013 (0.0046)	0.0055*	0.021 (0.012)	0.073
Waist circumference	0.10 (0.077)	0.18	0.13 (0.076)	0.082	0.48 (0.22)	0.033

The model includes the following covariates for the 43 participants: ethnicity, smoking history, age, and estimated cell type proportions. P-values marked with an asterisk (*) are significant at our Bonferroni-corrected α of 0.0071

Fig. S1 Density Plots of Phenotypes Grouped by Timepoint



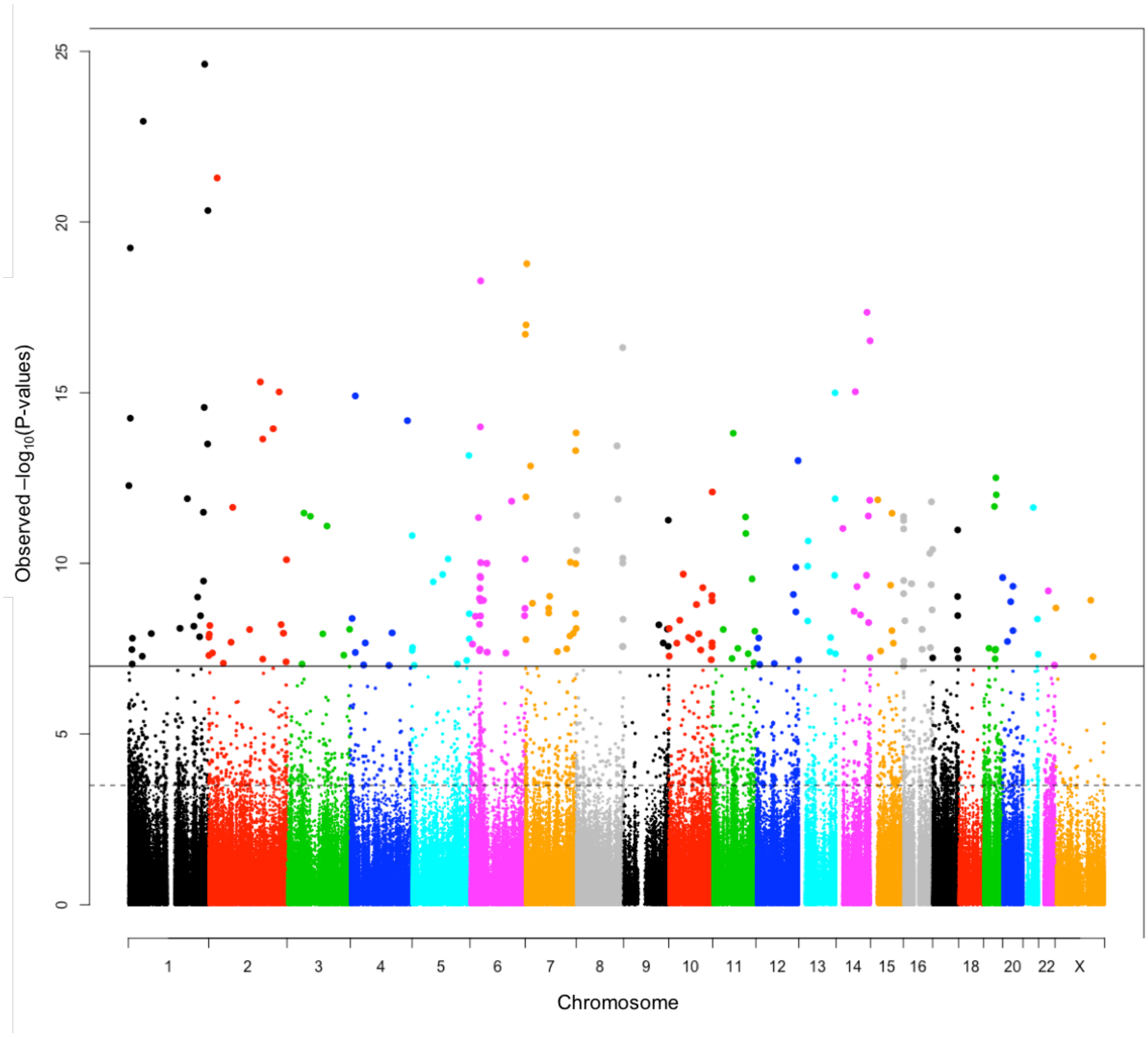
Density plots of our seven diabetes-related phenotypes at each of three timepoints. The first timepoint is red, the second is green and the third is blue. HOMA-IR, and TG/HDL-C ratios are log-transformed

Table S1. Top ten CpG sites reflecting methylation changes with age

CpG site	Est. (SE)	P-value	Associated Gene[a]
cg14252149	-0.0081 (0.00051)	2.37E-25	LGALS8
cg22337626	0.0067 (0.00045)	1.13E-23	MAST2
cg01188578	-0.0065 (0.00047)	5.09E-22	HADHA
cg04246708	-0.0059 (0.00045)	4.63E-21	CNST
cg15075357	0.0058 (0.00046)	5.83E-20	NPHP4
cg09281805	-0.0070 (0.00057)	1.67E-19	FOXK1
cg14782559	0.0059 (0.00049)	5.26E-19	COL11A2
cg18239511	0.0040 (0.00035)	4.43E-18	
cg03122926	0.0056 (0.00049)	1.041E-17	
cg01156747	0.0064 (0.00057)	1.96E-17	

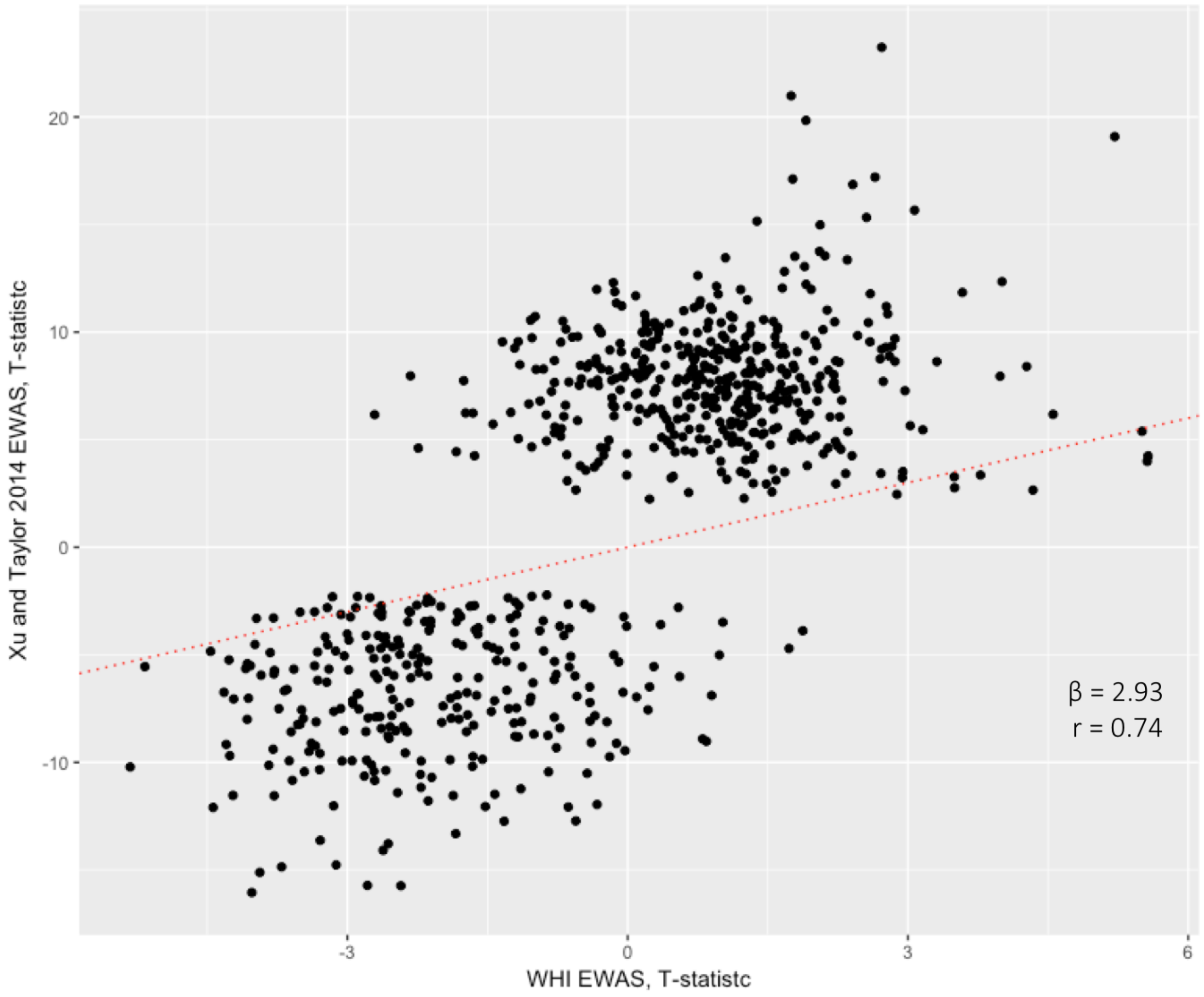
[a] Associated gene according to the Illumina HumanMethylation450K manifest file (https://support.illumina.com/downloads/infinium_humanmethylation450_product_files.html)

Fig. S2 Manhattan Plot for Association Between DNA Methylation and Chronological Age



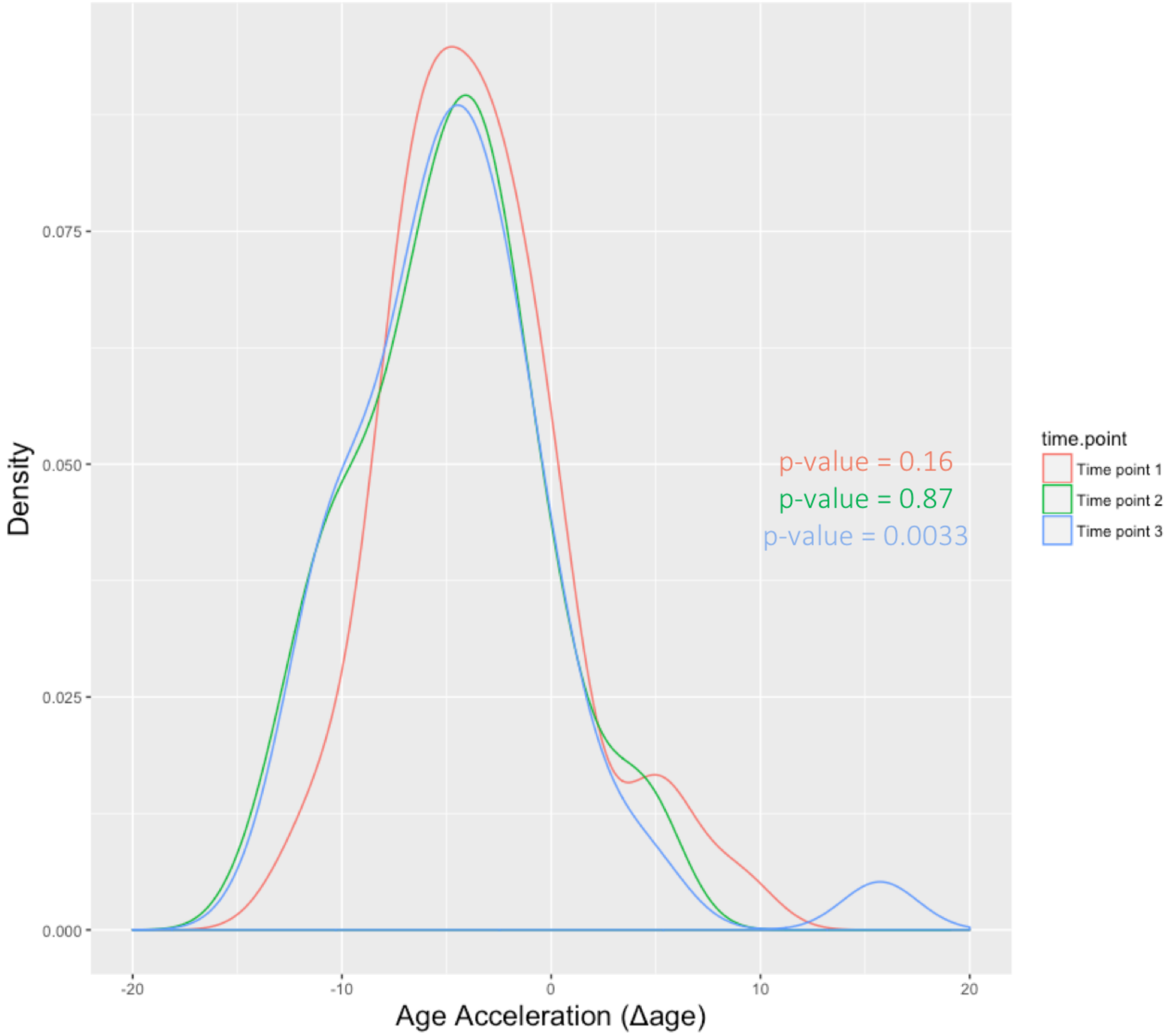
This figure shows the Manhattan plot (−log10 of the p value by genomic location) of the association results. Each dot represents the p-value associated with a CpG. The solid black line indicates the significance threshold of 1E-07. A random effect for participants was included to account for repeated measures within subject

Fig. S3 Comparison of T-statistics of 713 Age-Related CpGs



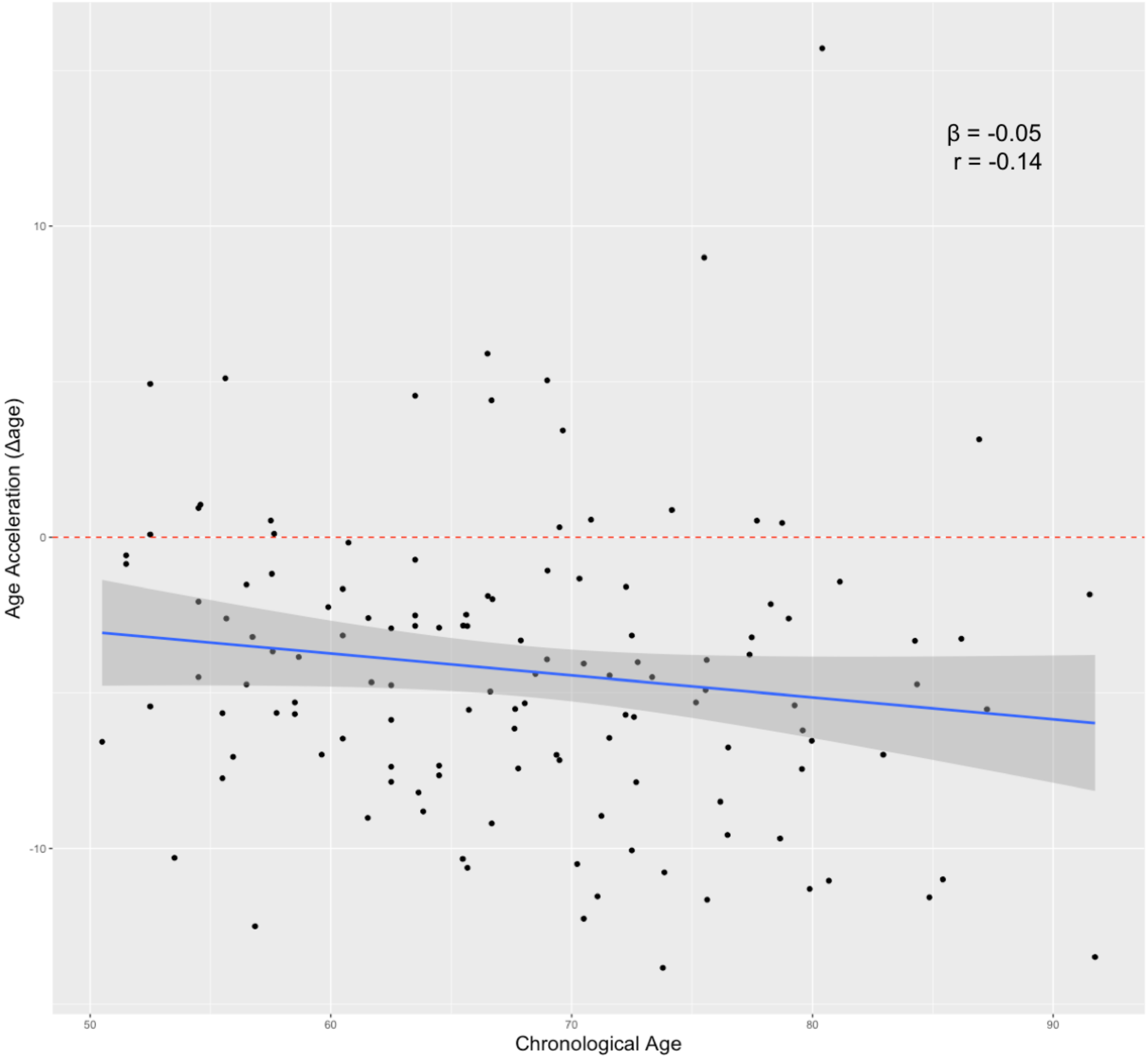
T-statistics of 713 CpGs that were available in both our EWAS and in results presented by Xu and Taylor 2014 were plotted for comparison. The correlation of 0.74 suggests that, though our dataset is smaller, our results are consistent with those identified in a more well-powered study

Fig. S4 Density Plot of Participants' Age Acceleration Grouped by Timepoint



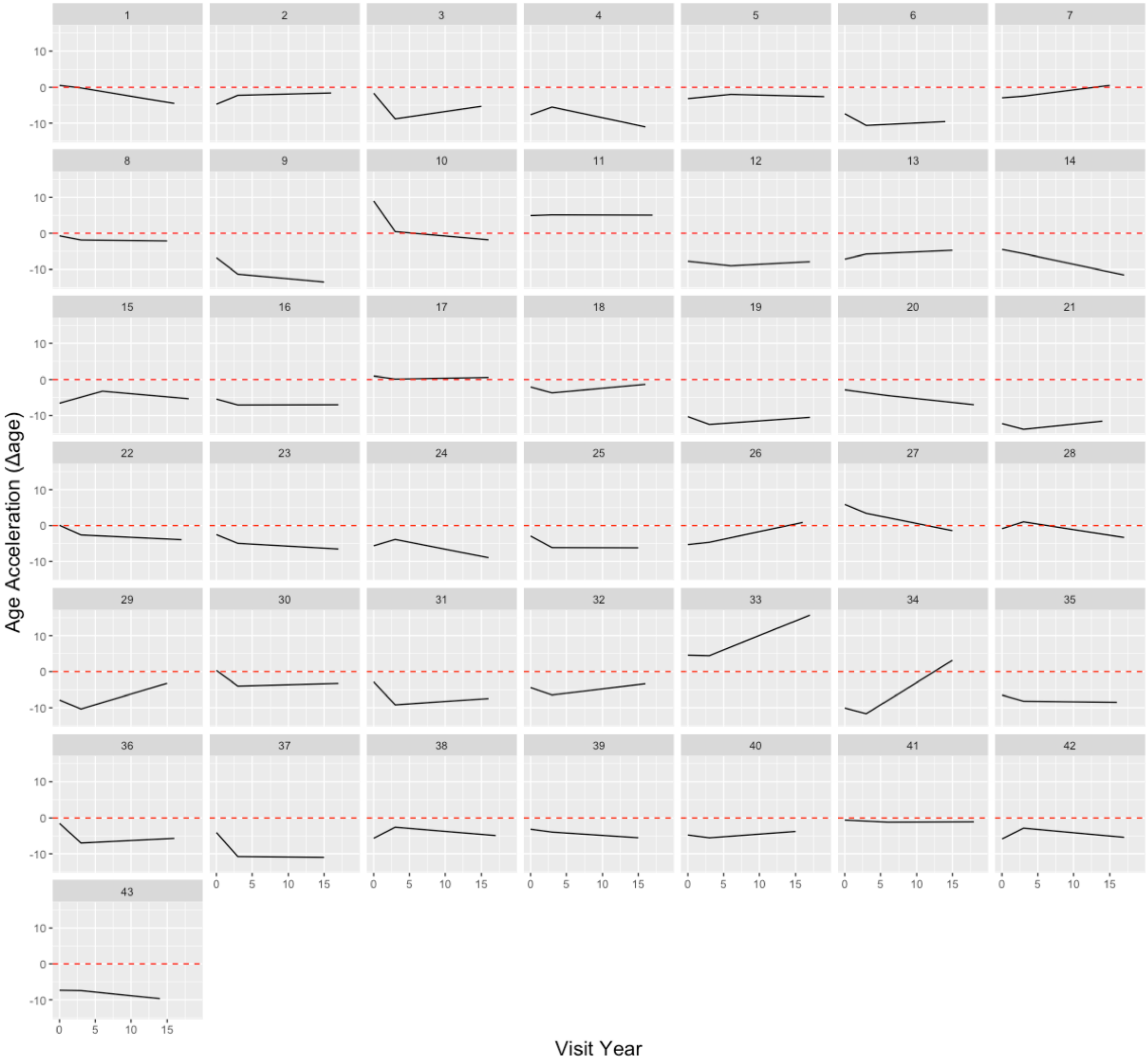
Participants showed substantial variation in their measures of Δ_{age} (range = -13.8 to 15.7 years, mean = -4.3, SD =4.7 years). Red shows the density plot at the first timepoint, green at the second and blue at the third. There was little variation in Δ_{age} across the three timepoints. A Shapiro-Wilk normality test of Delta at each timepoint was performed, timepoints 1 and 2 are normally distributed, while timepoint 3, due to an extreme value, is not normally distributed

Fig. S5 Chronological Age (x-axis) vs. Age Acceleration (y-axis)



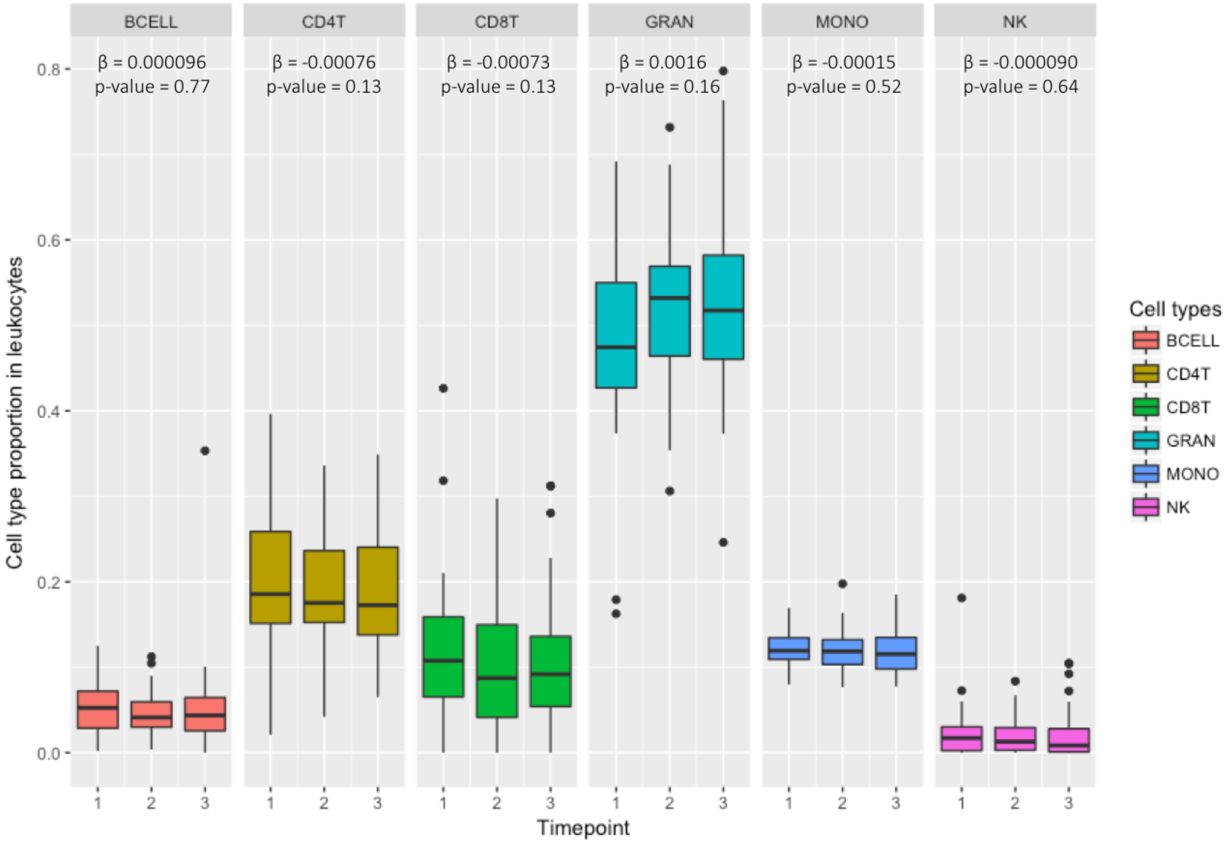
The Δ_{age} term exhibits a negative relationship with chronological age. A mixed effects model was fit that regressed Δ_{age} on chronological age. The dotted red line represents a line of slope = 0, representing no significant relationship between Δ_{age} and chronological age. The solid blue line represents the regression line obtained from the mixed effects model. The line has a negative slope (β= -0.05) but this trend was not significant (p=0.13), suggesting that age acceleration is roughly independent of chronological age in our samples

Fig. S6 Changes to 43 Participants' Age Acceleration Over Time



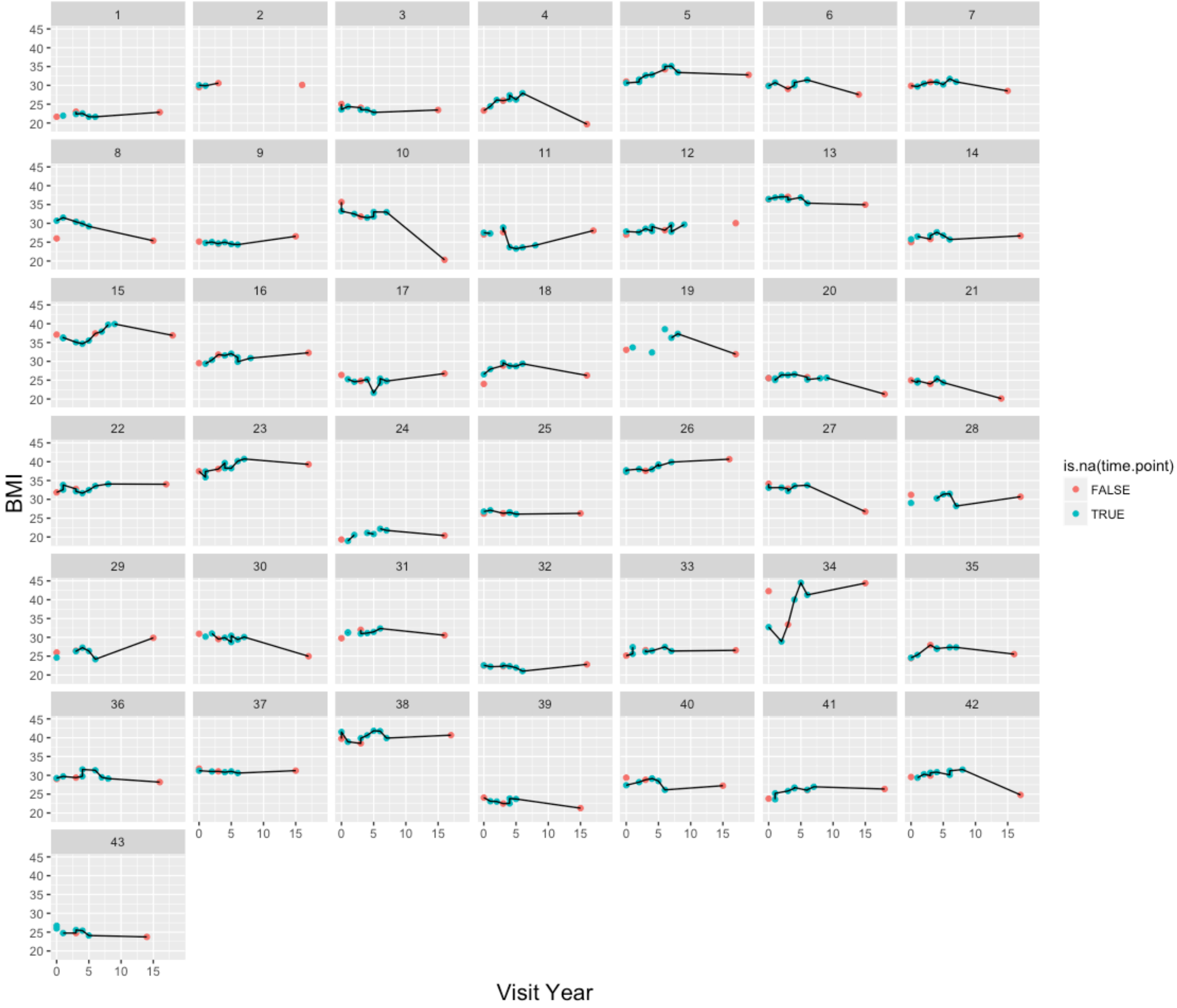
Each subplot represents one participant; a solid black line connects the participants' three measures of Δ_{age} across the three timepoints. The dotted red line represents a line of slope = 0, reflecting no significant relationship between Δ_{age} and years in the study. The observed lack of change in Δ_{age} with time was tested using a mixed effects model, with the year of the participant's clinic visit as a fixed effect and a random effect for participants the slope obtained (-0.041) was not significantly different from the null value of zero (p=0.25). This indicates that age acceleration, on average, does not change significantly over the course of the study for most participants. The lack of significant change in slope over time suggests that 1) Δ_{age} appears stable over time, and 2) our estimates of Δ_{age} are not influenced by possible batch effect

Fig. S7 Estimated Proportions of B cell, CD4T, CD8T, Granulocyte, Monocyte and NK cells in Peripheral Blood



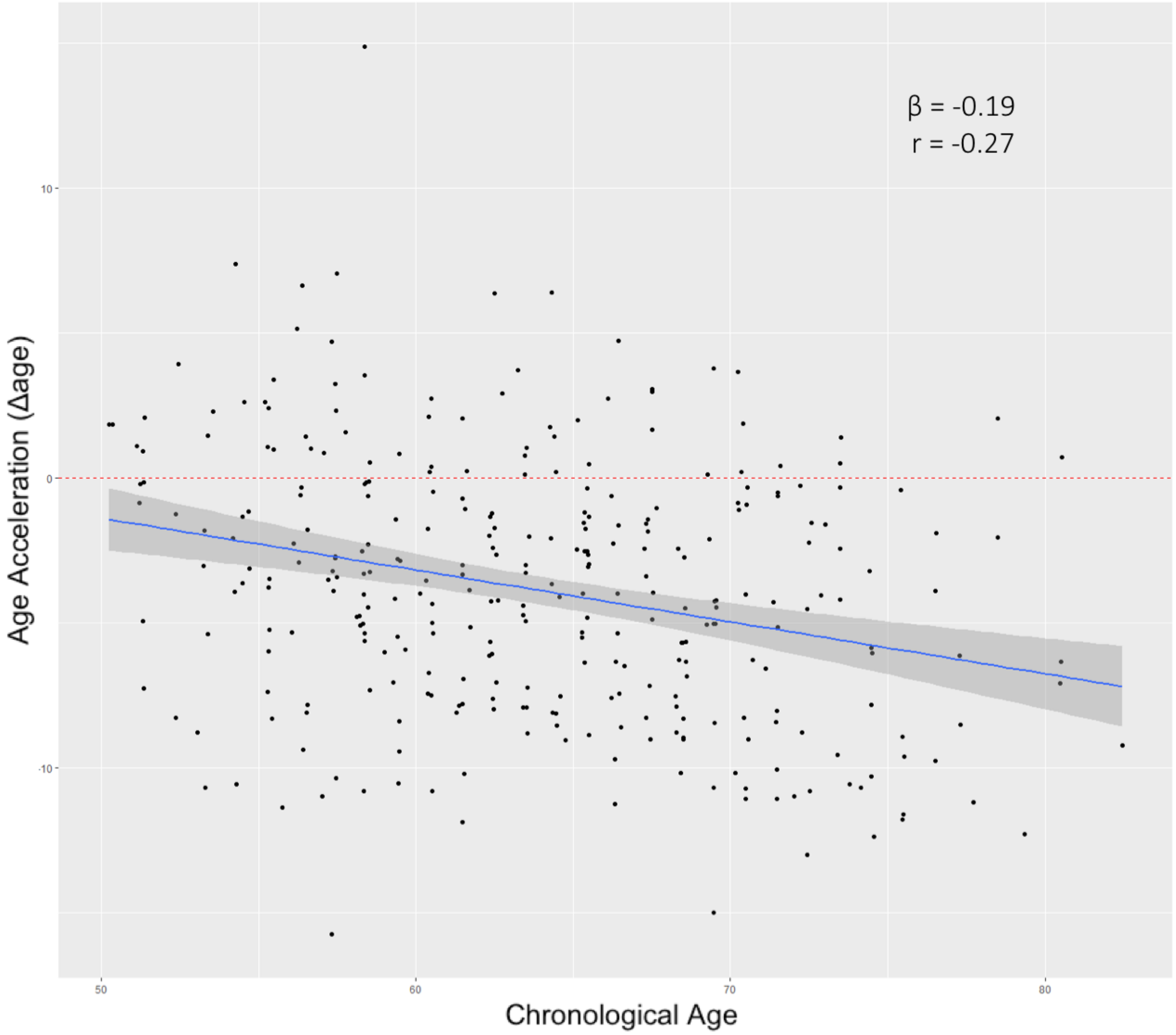
Boxplots of cell type proportions organized in separate panels by cell type and separated by timepoint. Each of the six cell type proportions was regressed on the visit year when blood was drawn and the resulting p-value of the slope is included in the plot, this regression model included a random effect for the participant. We see that the cell proportions do not change significantly across the course of the study

Fig. S8 Changes to 43 Participants' BMI Over Time



Each subplot represents one participant; a solid black line connects the participants' measures of BMI across the study. The red dots represent one of three timepoints when DNAm data in available, other visits are marked by blue dots

Fig. S9 Chronological Age (x-axis) vs. Age Acceleration (y-axis) for replication population



The Δ_{age} term exhibits a negative relationship with chronological age. A mixed effects model was fit that regressed Δ_{age} on chronological age. The dotted red line represents a line of slope = 0, representing no significant relationship between Δ_{age} and chronological age. The solid blue line represents the regression line obtained from the mixed effects model. The line has a negative slope (β= -0.19) and this trend was significant (p=0.0000039), suggesting that age acceleration is inversely related to chronological age in our replication cohort. This could be a true underlying relationship due to a non-linear relationship between chronological age and DNAm age or, perhaps, a result of selection bias with subjects who survived to advanced age being more likely to show lower Δ_{age}

Table S2. Sensitivity Analysis: Multivariate regression analysis of diabetes-related phenotypes on age acceleration

Phenotype	**Model 2** Δ_{age}[a]		**Model 2** Δ_{age} + alcohol consumption[b]		**Model 2** Δ_{age} + total caloric intake[c]		**Model 2** Δ_{age} + family history of diabetes[d]		**Model 2** Δ_{age} + incident diabetes[e]		**Model 2** Δ_{age} + incident diabetes treatment[f]	
	Est. (SE)	**P-value**	**Est. (SE)**	**P-value**	**Est. (SE)**	**P-value**	**Est. (SE)**	**P-value**	**Est. (SE)**	**P-value**	**Est. (SE)**	**P-value**
BMI	0.29 (0.087)	0.0012*	0.26 (0.086)	0.0032*	0.32 (0.090)	0.0007*	0.28 (0.085)	0.0015*	0.26 (0.089)	0.0044*	0.29 (0.088)	0.0015*
Fasting Glucose	0.97 (0.34)	0.0073	0.83 (0.33)	0.015	0.75 (0.36)	0.043	0.87 (0.33)	0.0091	0.61 (0.30)	0.045	0.96 (0.34)	0.0054*
Fasting Insulin	0.26 (0.24)	0.28	0.11 (0.23)	0.62	0.22 (0.25)	0.38	0.18 (0.23)	0.44	0.23 (0.24)	0.34	0.27 (0.24)	0.27
HOMA-IR	0.089 (0.065)	0.18	0.050 (0.063)	0.42	0.073 (0.069)	0.29	0.072 (0.065)	0.27	0.068 (0.064)	0.29	0.090 (0.066)	0.17
TG/HDL-C ratio	0.048 (0.040)	0.24	0.036 (0.038)	0.34	0.012 (0.036)	0.75	0.047 (0.039)	0.23	0.046 (0.041)	0.27	0.049 (0.041)	0.23
TyG index	0.021 (0.012)	0.073	0.018 (0.011)	0.10	0.0082 (0.012)	0.48	0.023 (0.011)	0.052	0.018 (0.012)	0.013	0.023 (0.012)	0.053
Waist circumference	0.48 (0.22)	0.033	0.39 (0.22)	0.079	0.53 (0.22)	0.021	0.42 (0.22)	0.056	0.37 (0.22)	0.10	0.46 (0.22)	0.043

[a] Model 2 Δ_{age} results from Table 2

[b] Weekly consumption of alcoholic beverages was included as a covariate in model

[c] Daily dietary energy intake for participants in kilocalories was included as a covariate in model

[d] Family history of diabetes was included as a covariate in model, with n = 11 participants responding ‘Yes’

[e] Incident diabetes was included as a covariate in model, with n = 4 participants having incident diabetes in at least one time period

[f] Incident treatment with antidiabetic agents (including pills or insulin shots) was included as a covariate in model, with n = 4 participants under treatment in at least one time period

Table S3. Multivariate regression analysis of diabetes-related phenotypes on age and biological age acceleration, 5 outliers removed

	Model 1		**Model 2**		**Model 2**	
Phenotype	**Coefficients on chronological age**		**Coefficients on chronological age**		**Coefficients on Δ_{age}**	
	Est. (SE)	**P-value**	**Est. (SE)**	**P-value**	**Est. (SE)**	**P-value**
BMI	-0.022 (0.023)	0.34	-0.018 (0.024)	0.45	0.049 (0.094)	0.60
Fasting Glucose	0.14 (0.14)	0.33	0.20 (0.15)	0.18	0.60 (0.42)	0.16
Fasting Insulin	0.31 (0.11)	0.0079	0.35 (0.12)	0.0039*	0.40 (0.31)	0.20
HOMA-IR	0.086 (0.031)	0.0081	0.098 (0.032)	0.0038*	0.11 (0.083)	0.18
TG/HDL-C ratio	-0.042 (0.016)	0.0099	-0.043 (0.017)	0.011	-0.015 (0.048)	0.75
TyG index	-0.014 (0.0049)	0.0064*	-0.014 (0.0051)	0.0066*	-0.0060 (0.015)	0.6915
Waist circumference	0.082 (0.077)	0.29	0.098 (0.082)	0.23	0.18 (0.28)	0.52

The model includes the following covariates for the 43 participants: ethnicity, smoking history, age, and estimated cell type proportions. The 5 outliers have been removed

Table S4. Demographic and clinical characteristics of replication population (N=157, 314 observations)

Variable	Mean +/- SD or percentage		
	Baseline (SD)	Follow-up (SD)	# missing obs.
Chronological age (years)	62.10 (6.62)	66.04 (6.63)	0
DNAm age (years)	58.63 (6.99)	61.78 (6.95)	0
Δ_{age} (years)	-3.47 (4.54)	-4.27 (4.47)	0
BMI (kg/m^2)	29.20 (5.86)	29.73 (6.06)	30
Fasting glucose (mg/dL)	99.69 (24.61)	98.39 (23.55)	0
Fasting insulin (uIU/mL)	10.86 (5.72)	11.74 (6.35)	12
HOMA-IR	2.80 (2.10)	2.99 (2.38)	12
TG/HDL-C ratio	3.26 (4.75)	2.99 (2.69)	0
TyG index	8.80 (0.61)	8.79 (0.58)	0
Waist circumference (cm)	88.07 (12.95)	89.79 (13.55)	54

Table S5. Multivariate regression analysis of diabetes-related phenotypes on age and biological age acceleration among replication population

Phenotype	Model 1 Coefficients on chronological age		Model 2 Coefficients on chronological age		Model 2 Coefficients on Δ_{age}	
	Est. (SE)	P-value	Est. (SE)	P-value	Est. (SE)	P-value
BMI	0.016 (0.043)	0.71	0.026 (0.047)	0.58	0.038 (0.065)	0.56
Fasting Glucose	-0.0090 (0.23)	0.97	-0.052 (0.24)	0.83	-0.17 (0.33)	0.61
Fasting Insulin	0.062 (0.060)	0.31	0.087 (0.065)	0.18	0.096 (0.093)	0.30
HOMA-IR	0.024 (0.022)	0.28	0.037 (0.024)	0.13	0.049 (0.033)	0.15
TG/HDL-C ratio	-0.041 (0.037)	0.28	-0.033 (0.040)	0.41	0.029 (0.058)	0.62
TyG index	0.0026 (0.0054)	0.62	0.0031 (0.0057)	0.59	0.0019 (0.0079)	0.81
Waist circumference	0.095 (0.12)	0.42	0.17 (0.12)	0.18	0.30 (0.17)	0.088

The model includes the following covariates for the 157 participants: ethnicity, age, and estimated cell type proportions

Table S6. Multivariate regression analysis of diabetes-related phenotypes on age and biological age acceleration, third timepoint censored

	Model 1		Model 2		Model 2	
Phenotype	Coefficients on chronological age		Coefficients on chronological age		Coefficients on Δ_{age}	
	Est. (SE)	P-value	Est. (SE)	P-value	Est. (SE)	P-value
BMI	0.080 (0.071)	0.27	0.099 (0.074)	0.19	0.081 (0.10)	0.43
Fasting Glucose	0.23 (0.21)	0.27	0.26 (0.22)	0.24	0.16 (0.33)	0.62
Fasting Insulin	0.13 (0.095)	0.18	0.13 (0.098)	0.19	-0.017 (0.15)	0.91
HOMA-IR	0.036 (0.026)	0.18	0.036 (0.027)	0.20	-0.0052 (0.043)	0.90
TG/HDL-C ratio	0.023 (0.031)	0.47	0.029 (0.032)	0.38	0.043 (0.049)	0.39
TyG index	0.0081 (0.0082)	0.33	0.0094 (0.0084)	0.27	0.0090 (0.013)	0.49
Waist circumference	0.10 (0.18)	0.57	0.11 (0.18)	0.55	0.038 (0.26)	0.88

The model includes the following covariates for the 43 participants: ethnicity, smoking history, age, and estimated cell type proportions. The third timepoint has been censored, so only the first two timepoints are included

Chapter III. Modeling the impact of age-related epigenetic variability on aging-related phenotypes

Crystal D. Grant, Thomas H. Jonkman, BIOS Consortium, Riccardo E. Marioni, Ian J. Deary, Bastiaan T. Heijmans

Abstract

DNA methylation (DNAm) data has been used to identify CpGs across the genome that reflect increased variability with age, termed age-related variably methylated cytosines (aVMCs). Interestingly, methylation at these aVMCs was linked to changes in the expression of genes with key roles in biological pathways implicated in aging. We hypothesized that participants' epigenetic variability associates with aging-related phenotypes, and thus can serve as a biomarker of aging. We analyzed DNAm at 412,373 CpGs in whole blood samples from 4,112 participants. Using Double Generalized Linear Models to test for age-related changes in variance, we discovered 11,524 aVMCs ($P < 10^{-7}$). Of these, 3,151 replicated in two independent data sets of whole blood (n=643) and purified monocytes (n=1,187). Functional annotation of the replication aVMCs found them enriched in repressive regions of the genome, and more specifically, enriched at developmental genes. We used these aVMCs to construct a composite score indicating participants' methylomic variability relative to the young (<30 years) group. This variability score was then evaluated for its link to known aging-related phenotypes in an external longitudinal dataset from the Lothian Birth Cohort (LBC1921 n = 469, LBC1936 n = 1,055). Though these sites appear functionally linked to aging pathways, the score does not appear to contribute significantly to longitudinal models of aging-related phenotypes nor does it contribute to predictions of all-cause mortality. Thus, while modeling epigenetic variability in aging may be informative, an alternate approach to the development of this score is necessary.

Introduction

DNA methylation (DNAm), shows robust, genome-wide changes with age.[87,102,103] Capitalizing on these changes with age, researchers have used DNAm data at a subset of age-related differentially methylated cytosines (aDMCs)[99,101,104] to estimate chronological age. This estimate, based on DNAm patterns (DNAm age), can have a strong correlation with individuals' chronological ages. Some studies have used the difference between predicted DNAm age and chronological age, the residual, as an indicator of biological age. An organism's biological age predicts its physiological functioning and disease susceptibility better than chronological age alone.[22] However, a recent study[128] found that the residual can be influenced by the sample size used in training the predictor. This suggests that the use of this residual as an indicator, or biomarker, of biological age may not be the optimal approach.

Considering the variation in DNAm age among people of the same chronological age, a relevant alternative biomarker may take into account environmental and stochastic changes to an individual's epigenome over time. This model would, in theory, better capture biological age as well as the variability among humans in the process of aging. We hypothesize that a more informative model of biological aging should feature CpGs undergoing changes in variability over time instead of those reflecting linear age-related methylation change. An example of such a CpG, termed an age-related variably methylated cytosine (aVMC) in contrast to an aDMC can be observed in **Supplementary Figure 1**. This hypothesis is supported by studies among monozygotic twins observing variability of DNAm increasing with age,[78,129] and with discordant disease development.[138] Studies have found that this increase in variability appears to occur especially at age-associated CpGs,[134] and that these CpGs appear near to genes involved in the aging process.[132]

In a recent publication from our group,[106] Slieker *et al.* identified 6,366 aVMCs, across the genome that showed increased variability with age. Moreover, methylation at these aVMCs was found to strongly associate with changes in gene expression of genes with central roles in biological pathways associated with aging, including: apoptosis, DNA repair, and lymphocyte activation. These sites represent a distinct class of CpGs indicating which regions of the genome are undergoing epigenetic drift. Several studies have characterized such CpGs using differing methods of identification. Slieker *et al.*[106] used the Breusch–Pagan test for heteroscedasticity[193] which first models linear changes in the relationship between DNAm and age, then tests the resulting squared residuals for a relationship between their variance and age (6,366 CpGs identified). This approach can also be used on longitudinal DNAm data where the residuals are modeled as the independent variable in a mixed effect model with a random intercept to account for repeated measures in the same individual (570 CpGs identified).[131] Similarly, also in longitudinal data, a random slope model can be applied to the CpGs in the 450K array to find sites whose slopes reflect significant change over time (1507 CpGs identified).[132] Another approach is to use Double Generalized Linear Models (DGLM)[194] in which the linear and variance relationships between DNAm and age are modeled simultaneously. This approach was used on our cross-sectional cohort.

In the current study, we aimed to create an updated catalogue of aVMCs using a robust statistical approach, and to combine the DNAm data at these sites into a composite score of methylomic variability (MV). We hypothesize that this MV score, because of the links between aVMCs and biological pathways important in aging, will be a useful biomarker of aging. To test the utility of this biomarker, we apply it to longitudinal models of important age-related phenotypes[28] using longitudinal DNAm and phenotypic data from the Lothian Birth Cohorts.

Methods

BIOS Cohort Information

DNA methylation and RNA-seq data were generated within the Biobank-based Integrative Omics Studies Consortium (BIOS Consortium). The BIOS Consortium, detailed in [106], encompasses data from six Dutch biobanks: Cohort on Diabetes and Atherosclerosis Maastricht (CODAM),[195] LifeLines (LL),[196] Leiden Longevity Study (LLS),[197] Netherlands Twin Registry (NTR),[198] Rotterdam Study (RS),[199] and the Prospective ALS Study Netherlands (PAN).[200] To ensure that analysis was performed only on unrelated participants, a random twin was chosen from each twin pair from the Netherlands Twin Registry.

BIOS Data Cleaning

For each participant, whole blood samples were obtained. Briefly, 500 ng of genomic DNA was bisulfite converted using the EZ DNA Methylation kit (Zymo Research, Irvine, CA, USA) and hybridized to Illumina Infinium 450k arrays according to the manufacturer's protocols. Signal intensities were measured using an Illumina iScan BeadChip scanner. Quality control (QC) on the DNAm data was performed using the R package MethylAid.[201] DNAm was available at 481,388 CpGs; of these, 59,232 CpGs were ambiguously mapped probes and were removed from future analysis as recommended by Zhou *et al.*[202] The remaining 422,156 CpGs were used in future analysis. Probes with a high detection P value (>0.01), probes with a low bead count (<3 beads), and probes with a low success rate (missing in >95 % of the samples) were set to missing. Probes mapping to chromosomes X and Y were excluded from future analyses. Functional normalization, as implemented in the minfi R package,[203] was used per cohort.

Out of 4,386 samples passing QC, 263 of these participants did not have their age available and were removed from future analysis (n=4,123). The getSex.DNAmArray() function from the DNAmArray R package,[204] which uses DNAm data on the X chromosome to impute sex, was used to impute sex for participants for whom sex was not reported by the biobank (of the 4123, n=2). This step also identified participants (n=11) for whom the biobank's reported sex appeared to be incorrect. These 11 participants were removed from analysis; imputed sex was then used in future analysis for all participants (n=4,112). Residual batch effects were removed using ComBat,[205] with biobank as batch and gender and age as outcome variables. The sample identities were confirmed by comparing DNAm data to genotype data using MixupMapper.[206]

Cell count data of the whole blood samples (neutrophils, lymphocytes, monocytes, eosinophils, and basophils) were available for the majority of samples (>69%). Previous studies have emphasized the importance of accounting for changes in cell type proportions during aging,[96,207] cell counts were imputed for each individual and these counts were included as a covariate. To impute these missing cell counts, a prediction model of cell composition was fitted using the subset of participants for whom cell counts were available. A multivariate partial-least-squares model was fit to the normalized DNAm data, and using the fitted model, cell composition was imputed for all samples. This step was performed using the wbccPredictor R package, pipeline available on Github,[208] and imputed cell type percentages were used as covariates in future analysis.

Identifying aVMCs

Two methods were used to identify aVMCs then their overlap compared in **Supplementary Figure 3.** The first method was the Breusch–Pagan (BP) test for heteroscedasticity[193] (as used in Slieker *et al.*) and the second was Double Generalized Linear Models (DGLM).[194] Both models were applied to DNAm in M-values, in contrast to Slieker *et al.* which used Beta values. M-values are considered more statistically valid for analyses when compared to Beta values because they are considered approximately homoscedastic,[209] thus this approach is an improvement over previous aVMC selection. In applying the BP test for heteroscedasticity, first a linear regression model accounting for changes in age, blood cell composition (lymphocytes, monocytes, basophils, eosinophils), and gender was applied. Next, squared residuals were tested for an association with age, again adjusting for blood cell composition and gender. CpGs that showed a significant association between squared residuals and age were considered aVMCs, using a Bonferroni-adjusted alpha level of 0.05/(412,373 tests).

In an alternative method, aVMCs were identified by using DGLM, as implemented in the R package dglm.[210] Briefly, DGLM works by simultaneously modeling both the mean and dispersion in the relationship between variables–in this case the relationship between age and DNAm–until convergence. This approach is thought to be advantageous over the use of BPtest because it avoids some of the issues associated with overdisperson and the mean value affecting the measure of variances. The linear model of the mean included age, blood cell type proportions, and gender as covariates. To identify sites reflecting changes only in variability with age, only age was included in the dispersion model and the p-value on this age term was extracted and used to identify which sites were aVMCs (using a Bonferroni corrected alpha of 0.05). This method identified 11,524 aVMCs. For four CpGs (cg09370299, cg04045327, cg24139216, cg07035145), the DGLM model failed to converge and these sites were removed from analysis. For both methods, validation of aVMCs was performed in two external datasets of whole blood (n=643) and monocyte datasets (n=1,187). Discovery of aVMCs was performed on M-values, other analyses were performed on Beta values for easier visualization of the DNAm data; M-values were transformed to Beta values using the lumi package.

The aVMCs were validated in two external datasets, whole blood and monocyte datasets. For the whole blood data, IDAT files from Hannum *et al.*[99] underwent the same quality and normalization procedures outlined above. After quality control, 643 samples were used in subsequent analysis. For the monocyte data, normalized data were obtained from GEO[211] (accession number GSE56046).[212] For the whole blood data, the BP test and DGLM outlined above were used with the same cell type proportions used as covariates. For the monocyte data, imputed cell type proportions were used as covariates in the model (B cells, T cells, natural killer (NK) cells, and neutrophils).

Smoking sensitivity analysis in aVMC discovery

Studies have suggested that smoking can influence DNAm patterns.[73,213] Self-reported smoking behavior was available for a subset of participants from BIOS (n=3,222). EpiSmokEr,[214] was used to impute smoking behavior among all participants. This smoking exposure score was then included as part of a sensitivity analysis in aVMC discovery. Briefly, EpiSmokEr uses DNAm data at 187 CpGs to compute a numeric score indicating smoking exposure, with a higher score indicating more exposure. Among the BIOS data, participants ranged in their smoking score from -12.44 to 23.17 (mean=-0.01, sd=5.34). Agreement between participants' self-reported smoking behavior and calculated smoking exposure scores can be observed in **Supplementary Figure 8A**. This smoking score was then included as a covariate in the identification of aVMCs, with almost all (>96%) of the same aVMCs being identified, confirming that smoking behavior does not influence aVMC discovery **(Supplementary Figure 8B, C**).

Functional annotation of aVMCs

Chromatin state annotations were obtained from the Epigenomics Roadmap project,[33] which contains tissue-specific data on 15 chromatin states in 127 tissue types, imputed from 5 histone marks using a hidden Markov model. Though our data were generated in whole blood samples, the peripheral blood mononuclear cell (PBMC) annotation was used as proxy for our tissue. Additionally, data on 6 histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me3, H3K27me3) in PBMCs were obtained.

For genic annotations, genomic ranges of all protein-coding genes were obtained using the R package *ensembldb*.[215] Nearest genes were annotated to the aVMCs as well as five gene-centric features: 1) Distal promoter: 10kb - 1.5 kb upstream of the coding sequence of each gene; 2) Proximal promoter: 1.5 kb upstream to 0.5 kb downstream of the coding sequence of each gene; 3) Gene body: the coding sequence of each gene, minus the overlap from the proximal promoter; 4) Downstream: 5000 bp downstream of the coding sequence of each gene; and 5) Intergenic: none of the above.

For the enrichment analysis, odds ratios (OR) were calculated for aVMCs overlapping an annotation compared to non-aVMCs (all other CpGs from the array). Enrichments were expressed as odds ratio on a log2 scale. Fisher's exact test was then performed to test for enrichment of a feature in aVMCs vs. non-aVMCs. Gene ontology (GO) enrichment of the nearest annotated genes was performed using the R package STRINGdb.[216] GO terms were exported to the web-based GO tool WEGO,[217] to investigate enrichments of high-level biological processes. In WEGO, the full human genome was used as a background, and only GO terms of Biological Process of level 2 were visualized; all GO-terms tagged as "obsolete" were removed from analysis. For visualization of enrichments of GO terms, only enrichments with a p-value smaller than $0.05/15,631 = 3.19877\times10^{-6}$ were considered statistically significant, as 15,361 GO terms were tested for enrichment.

LBC Descriptions and Data Cleaning

The Lothian Birth Cohort (LBC) studies feature two birth cohorts of 1921 (LBC1921) and 1936 (LBC1936), and include children born in 1921 and 1936 respectively in the Lothian region of Scotland (which includes Edinburgh and its surrounding areas).[128] As part of national testing for almost every child attending school in the region at around age 11, participants completed a test of cognitive ability (the Moray House Test); this was done on June 1, 1932 (n=89,498) and June 4, 1947 (n=70,805) for participants in LBC1921 and LBC1936 respectively.[218] In both studies, participants were re-contacted later in life and have been followed-up on at subsequent waves. For LBC 1921, the first wave of data collection occurred between 1999 and 2001; for LBC 1936, the first wave of data collection occurred between 2004 and 2007—where waves of testing were spaced roughly 3 years apart. LBC1921 participants at Wave 1 are on average at age 79 and there have been four additional follow-up waves for data collection at average ages of 83, 87, 90, and 92. LBC1936 participants at Wave 1 are on average at age 70 and there have been four additional follow-up waves for data collection at average ages of 73, 73, 76, and 79. Both cohorts have been deeply phenotyped during these later-life waves, including: white blood cell counts, blood biomarkers, cognitive testing, and psychosocial, lifestyle, and health measures. In addition to these phenotypes being measured, whole blood was drawn from participants for DNAm analysis using the 450k array. DNAm was measured in LBC1921 participants as Waves 1, 3, and 4; DNAm was measured in LBC1936 participants at all four waves.

The phenotypic data from LBC used in this analysis focused on established risk factors for age-related disease detailed in MARK-AGE,[28] including: C-reactive protein, MMSE, 6 meter walk time (in seconds), grip strength (both right and left were assayed in LBC136 but only right hand was used in analysis), serum creatinine, glycated hemoglobin/HbA1C, systolic blood pressure (measured 3 times while seated with the means of the measures taken in LBC136), serum albumin, total cholesterol, forced expiratory volume (FEV), serum urea nitrogen, body mass index (BMI), height, and weight. Mortality data for participants and their age in days at death was available for both cohorts (LBC1921 n = 519, 91.37%; LBC1936 n =

277, 25.39%). Age was measured in days since birth and divided by 365.25 to find age in years (used in subsequent analysis). Observations recorded as -999 or 999 in phenotypic data (n=10 in LBC 1936) were set to missing in future analysis.

DNA was extracted from whole blood samples for LBC1921 and LBC1936, using standard methods. Raw intensity data were background-corrected and normalized using internal controls, and methylation M-values were generated using the R minfi package.[203] DNAm M-values were regularized by constraining them to be in the interval between − 9.96 and 9.96 (corresponding to the interval 0.001 to 0.999 of the Beta-value). Participants with DNA methylation three standard deviations above and below the mean M value were excluded as outliers for each probe. Covariates including sex, age, and cell counts (CC), and batch effects including position in array (PIA), hybridization date (HD), set ID (SI), plate ID (PI) and array ID (AI, both PI and AI were regarded as random effects), were corrected for each probe. M-values were then transformed to Beta values using the lumi package.[219] Additional protocols and quality control steps are detailed in Zhang *et al.*[132]

Creating an MV score

To characterize the directionality and the degree to which LBC participants' DNAm values differ from the "Young" (aged >30 years) group of the BIOS consortium, the Z score at each of the 3,151 aVMCs was calculated for each individual relative to the mean and standard deviation within the Young group:

$$Z_{ij} = \frac{(DNAm_{ij} - \overline{DNAm_{\iota y}})}{\sigma_{iy}}$$

where Z_{ij} is the Z-score of participant *j* and aVMC *i*; $DNAm_{ij}$ is the Beta value of participant *j* and aVMC *i*; $\overline{DNAm_{\iota y}}$ the average Beta value in young participants (*y*) at aVMC *i*; and σ_{iy} is the standard deviation in young participants (*y*) at aVMC *i*. These Z-scores are then meant to capture variability as measured in standard deviations from the young DNAm value, with the assumption that values farther from the young mean represent aberrant DNAm among some older participants.

3,183 of the 3,151 sites were present in the LBC studies, so only these sites were used in constructing the score. MV scores were calculated for each participant in the LBC by totaling the number of aVMCs at which the Z score was greater than or equal to 2 or less than or equal to -2. This was done at each wave to measure change in the number of sites showing extreme variability over time in each participant.

Applying MV score to Longitudinal Data

For each aVMC, a linear mixed model that included a random effect for each participant was fit to account for repeated measures within participants. Covariates included in all models were sex, age in years (centered), and the square of the centered age value. These were included in the model to account for effects on phenotypes driven by linear or nonlinear changes in age. Thus, for participant (i) at timepoint (j), the following model was fit:

$$Age - related\ phenotype_{ij} \sim Age_{ij} + Age^{2}{}_{ij} + Sex_{i} + MV\ score_{ij}$$

The coefficients on each term were then presented in **Tables 1 and 2**.

Survival analysis using MV score

Quartiles were calculated for the MV score values (25%: 11; 50%: 53; 75%: 206) and 235 of the participants whose MV scores placed them in either the first and fourth quartile were compared for differences in their risk of mortality during the study. Five LBC121 participants were recorded at having died but did not include age in days at time of death, thus age at last follow-up was used as age at death. Cox models were used for analyzing the censored survival time data (from the age in days at blood draw until age in days at death or last follow-up). We regressed the censored survival times on covariates using Cox regression models from the R function coxph in the survival package. Sex was not found to have a significant effect on survival time (p-value = 0.67) and thus was not included as a covariate for overall survival.

Results

Identification of aVMCs

To model methylomic variability and characterize its link to age-related phenotypes, whole blood samples from 4,112 participants aged 18 to 87 were analyzed at 412,373 CpGs (**Supplementary Figure 2**). Building on the findings from Slieker *et al.* in which 3295 BIOS consortium participants were analyzed, in this study, an additional 817 participants were included. CpGs showing increased variability with age, independent of an average change in DNAm or changes in blood cell composition with age, termed age-related variably methylated cytosines (aVMCs), were identified. In contrast to Slieker *et al.*, DNAm M-values were analyzed instead of Beta values because of the benefit they offer in identifying differential sites. Additionally, Double Generalized Linear Models (DGLM) were implemented in identifying aVMCs, compared to Slieker's approach of using the Breusch–Pagan (BP) test. To ensure that the identification of aVMCs wasn't influenced by different cell compositions in blood with age, imputed cell type proportions were included as covariates in the model.

DGLM identified 178,597 sites (43% of CpGs tested) reflecting linear changes in DNAm with age, termed age-related differentially methylated cytosines (aDMCs) (Figure 1A); while the BP test identified 165,493 aDMCs (40% of CpGs tested). For sites reflecting changes in variability in DNAm with age, DGLM identified 11,524 aVMCs ($P < 10^{-7}$, Figure 1B); while the BP test identified 22,499 aVMCs ($P < 10^{-7}$). A subset of aVMCs (8.0% for BP test aVMCs and 53.5% for DGLM aVMCs) also reflected changes in their mean methylation with age in addition to age-related changes in variance, thus they were also considered aDMCs. The overlap of aVMCs identified using the two methods were compared in **Supplementary Figure 3**, finding that many aVMCs (11,376/22,647, 50.2%), especially those with the highest effect sizes, overlap between the two methods. Additionally, the majority of sites (99.97% for

DGLM, 99.88% for BP test) appear to reflect increased methylomic variability with age, with very few showing decreasing variability, consistent with previous findings.[106,131]

The aVMCs were validated in two large public datasets consisting of whole blood (n = 643, ages range 19–101)[99] and purified monocytes, (n = 1,187, ages range 44–83).[212] This validation step supports that aVMCs used in future analysis were not influenced by discovery in a Dutch population or by any age-related cellular composition changes in whole blood. Additionally, in monocytes, predicted blood cell subtype fractions (CD8+ T cells, CD4+ T cells, natural killer (NK) cells, B cells, and granulocytes)[96] were included as covariates in aVMC discovery. External replication of the identified aVMCs using each method was performed and the final number of aVMCs identified was 3,151 using DGLM **(Figure 1C)** and 9,038 using BP test **(Supplementary Figure 4)**. Overlap between the previously characterized 6,366 sites and the newly discovered sites was noted in **Figure 1D**, finding that each method results in the identification of unique aVMCs. Only the aVMCs identified using the more conservative DGLM method were used in future analysis.

aVMCs are depleted in active regions and enriched in repressed regions

Functional annotation of the 3,151 aVMCs was performed using chromatin state annotations from the Epigenomics Roadmap[33] in peripheral blood mononuclear cells (PBMCs) (used as a proxy for the whole blood collected in the BIOS consortium). The aVMCs appear to be present throughout the genome (**Figure 2A)**. Similar to the findings of Slieker *et al.*,[106] aVMCs appear to be enriched in regions featuring repressive chromatin states (**Figure 2B)**. Specifically, aVMCs are enriched in the regions marked by repressed polycomb (6.1-fold enrichment, P<0.0001), weak repressed polycomb (2.3-fold enrichment, P<0.0001), and bivalent enhancers (4.9-fold enrichment, P<0.0001). Conversely, aVMCs appear depleted in regions marked by strong transcription (0.05-fold enrichment, P<0.0001) and regions undergoing active transcription (0.23-fold enrichment, P<0.0001). Overall, 1,943 aVMCs (61.7%) mapped to segments marking repressed DNA. The aVMCs are also marked by histone modifications consistent with repressive chromatin marks **(Figure 2C)** and depleted in regions containing proximal promoters or gene bodies **(Figure 2D)**. This finding was supported by a weak enrichment of aVMCs in binding sites of the PcG repressive complex 2 (PRC2) protein EZH2 in the ENCODE blood cell line GM12878 (1.4-fold enrichment, P = 0.03). Gene ontology (GO) analysis of cis-associated genes was performed finding an enrichment for genes involved in developmental processes (P < 0.0001). (**Figure 3**)

DNAm at aVMCs remains relatively stable over time for most individuals

Longitudinal data from the Lothian Birth Cohorts (LBC) (detailed in **Figure 4**) was used to characterize changes over time in DNAm among the aVMCs discovered in the cross-sectional BIOS consortium. DNAm was collected in either 3 waves for LBC 1921 or 4 waves for LBC 1936, with waves spaced approximately 3 years apart. The LBC data features participants progressing through to old age, with ages between the first and final waves ranging from 79.1 to 90.1 years for LBC 1921 and 67.6 to 80.9 years for LBC 1936. **Supplementary Figure 5** shows an example of an aVMC in participants from LBC 1936 with data at all 4 waves. Though the Beta values appear to change over time, a simple linear regression for DNAm on age tested for each person reveals the slope is not significantly different from zero—indicating relative stability over time. This suggests that DNAm does not change significantly over time even at sites that are becoming more variable on a population-wide level. Of note in **Supplementary Figure 5,** are the participants whose slopes deviate significantly from zero (in red).

In contrast to participants whose Beta values appear to be significantly diverging from the Young value with time, in **Supplementary Figure 5** participants whose DNA remains stable but is relatively far from the mean Beta value in Young across all waves can be observed. We hypothesized that participants with this higher deviation are experiencing more epigenetic drift and that this drift is linked to negative phenotypic outcomes. In order to test the relationship between methylomic variability (MV) and phenotypic outcomes, the degree of MV across these aVMCs must be summarized into a single measure for each participant.

Development of methylomic variability (MV) score

Methods for creating this MV score were first explored using the BIOS dataset. Because many of the Beta values of the aVMCs appear weakly correlated **(Supplementary Figure 6A)**, it was hypothesized that the dimension reduction technique, Principal Component Analysis (PCA), could be used to summarize the aVMCs into a single measure. The mean and standard deviations of DNAm values among the Young (<30 years) group in the BIOS data were measured. Next, Z-scores relative to the Young group were calculated for each BIOS participant. PCA was performed on these Z-scores in order to capture underlying contributors of variance. The MV score would ideally itself resemble an aVMC, however, the first 5 PCs together capture a low (0.1672) proportion of the variance and PC1 does not resemble an aVMC **(Supplementary Figure 6B)**. Instead of this approach, the count of aVMCs at which a participant appears very different from Young were used as an MV score.

To examine whether epigenetic drift is linked to phenotypic outcomes, the degree to which DNAm changes throughout the life course must be characterized. To accomplish this, the Young (<30 years) population from the BIOS data, considered to be the standard for a healthy DNAm profile, was compared to the aged population in the LBC. The mean and standard deviations of DNAm values among the Young group were measured; Z-scores relative to the Young group were calculated in the LBC data for each participant at each aVMC at each wave. 3,138 of the 3,151 aVMCs discovered in the BIOS data were also present in the

LBC data, so only these sites were used in constructing the score. To create a composite indicator of MV, the number of sites at which the 3,138 Z-scores were above a value of 2 or below -2 (representing 2SD from the Young mean) was totaled for each LBC participant at each wave. This score captures the degree to which, at each wave, the participant's DNAm differs from that of a young, healthy participant.

MV score does not appear informative in modeling age-related phenotypes or mortality

In order to test whether this measure of methylomic variability contributes to longitudinal models of age-related phenotypes, MV scores were calculated at each wave. LBC participants with data for at least two waves (n= 160 LBC1921, n = 810 LBC196) were used in analysis (data available at each wave detailed in **Figure 4**). A density plot of this MV score can be observed in **Supplementary Figure 7A-B**, showing that the measured methylomic variability is low for many participants. The MV score does not appear to be correlated with age in either LBC1921 (r = 0.04. p-value = 0.25) or LBC1936 (r = 0.005, p-value = 0.81), suggesting that variability at the CpGs used to construct the score is not driven solely by an increase in participants' chronological age.

Established risk factors for age-related disease, detailed in MARK-AGE,[28] were available for analysis. These included: C-reactive protein, MMSE, 6 meter walk time, grip strength, serum creatinine, glycated hemoglobin (HbA1C), systolic blood pressure (BP), serum albumin, total cholesterol, forced expiratory volume (FEV), serum urea nitrogen, Body mass index (BMI), height, and weight. A linear mixed model with a random effect for participant was used to assess whether the inclusion of a covariate for the MV score improved longitudinal models of aging-related phenotypes; results from the model are listed in **Tables 1 and 2**. Both the Age and the Age^2 terms appear significantly informative in the model (Bonferroni corrected α of 0.0045 and 0.0036 for LBC1921 and LBC1936 respectively) for some of the aging phenotypes. Among the phenotypes in which age were informative in our models were: serum album, BMI, cholesterol, FEV, 6 meter walk time, and hemoglobin A1C. The phenotypes in which the Age^2 term was informative in our models (indicating a non-linear relationship between the phenotype and age) were: creatine, hemoglobin A1C, BMI, cholesterol, grip strength, and weight. This finding underlies the necessity in accounting for the non-linear effects of age when modeling aging phenotypes. The MV score, however, does not appear to contribute significantly to modeling any of the aging phenotypes surveyed as the coefficient does not reach significance.

Mortality data on whether a participant had died and their age in days at death was available for both cohorts (LBC1921 n = 519, 91.37%; LBC1936 n = 277, 25.39%). Kaplan–Meier survival curves for MV score quartiles at the Wave closest to death are presented in **Figure 5** for the LBC1921 cohort (because this cohort has higher rates of mortality). There does not appear to be a significant association between survival rates and MV scores (p-value = 0.92).

Discussion

We aimed to characterize CpGs reflecting increased variability with age and to test whether these sites would be informative in modeling age-related phenotypes and mortality. 3,151 aVMCs were identified and found to have statistically significant increases in variability among participants, in addition to replicating in two external cohorts. These sites appear to be present throughout the genome but are enriched in repressed regions and developmental genes and depleted in regions undergoing active transcription This finding suggest that sites that should remain repressed see that repression in dysregulation with as the DNAm maintenance processes degrading with age. Additionally, the aVMCs that replicated all feature increasing variance perhaps reflecting increased epigenetic drift with age.

Changes in DNAm relative to the Young in the discovery cohort were used to indicate both how different older participants' DNAm was from younger participants, and how DNAm at these aVMCs change over time. We hypothesized that participants with higher amounts of drift relative to the young mean, as evidenced by a higher MV score, would also have worse phenotypic outcomes including early risk of mortality. A linear mixed model was used to test the contribution of the MV score to longitudinal models of aging related phenotypes. The findings do not support the hypothesis that this particular implementation of the MV score functions as a meaningful biomarker of aging-related phenotypes as it does not appear to significantly contribute to these models. Additionally, using the LBC1921 cohort mortality data, the MV-score does not appear to significantly linked to risk of mortality.

A possible reason for this finding is that all 3,151 aVMCs may not be biologically meaningful in the aging-related phenotypes tested. Future directions of this study will include limiting the composition of the MV score to only sites found to be biologically meaningful in aging-related pathways. In an alternative approach to generating the MV score, 400 of the aVMCs identified in the cross-sectional BIOS data were also identified in a previous study using the LBC data to be longitudinally increasing in variability in DNAm.[132] Sites like these may be more informative in an MV score since they have been found to act as aVMCs in both cross-sectional and longitudinal data from their BIOS and LBC cohorts respectively. Alternatively, an analysis of differentially expression genes linked to the aVMCs may better inform an MV score, as these sites would serve as a direct link between variability in DNAm and changes in transcription with age or age-related phenotypes. RNA-seq expression profiles are available for 3,377 participants for whom DNA methylation data is available in BIOS, allowing for the discovery of these potentially more informative sites in developing a meaningful score of methylomic variability.

Tables and Figures

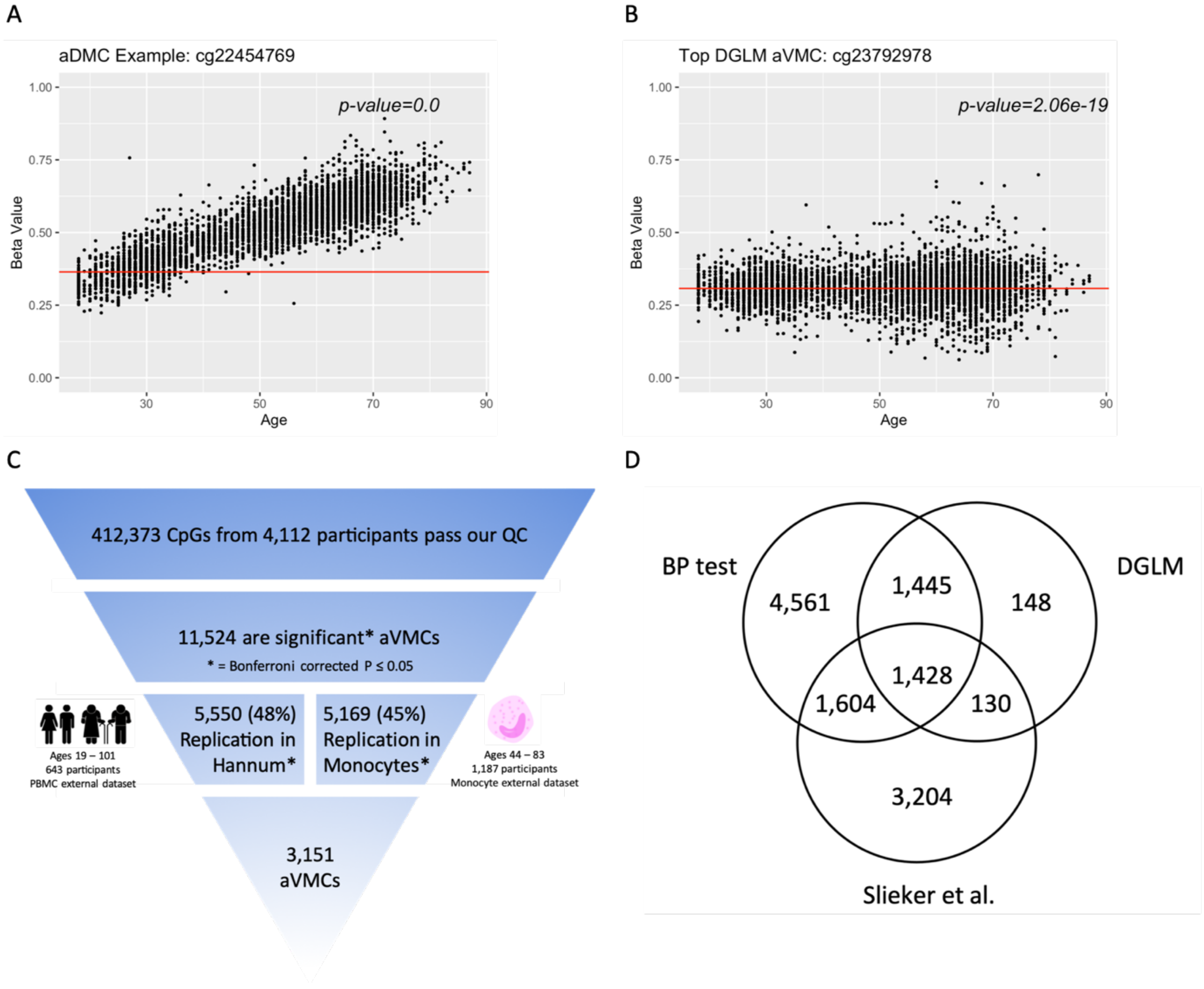


Figure 1: Discovery of aVMCs using Double Generalized Linear Models (DGLM)
A) An example of an aDMC identified from DGLM (p-value = 0.0). Each point represents a participant, the y-axis is the methylation Beta value, and the x-axis is age. The red horizontal line represents the mean methylation value among the young participants. **B)** The example aVMC pictured is the CpG with the lowest p-value (6.19 × 10-23) from DGLM. **C)** Flow chart of 3,151 CpGs identified as aVMCs, featuring replication rates in external datasets. **D)** Overlap between the previously characterized 6,366 sites from Slieker *et al.* and the newly discovered sites using DGLM (n=3,151) and BP test (n=9,038).

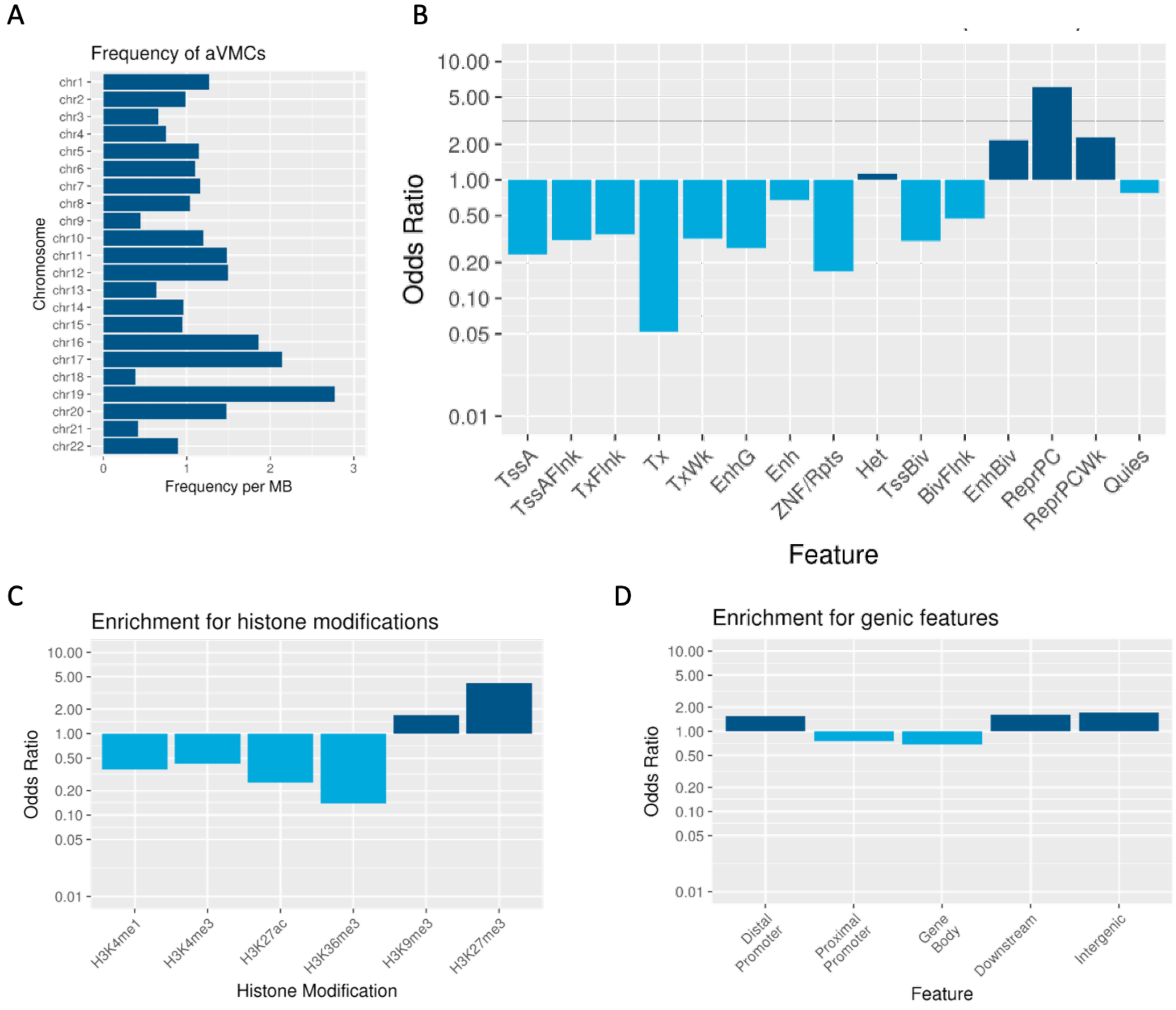


Figure 2: Characteristics of genomic regions featuring aVMCs

A) The frequency of aVMCs (x-axis) on each of the autosomal chromosomes (y-axis). **B)** The enrichment (odds ratio, y-axis) of aVMCs in chromatin state segments (x-axis) in peripheral blood mononuclear cells. **C)** The enrichment (odds ratio, y-axis) of aVMCs in histone modifications (x-axis). **D)** The enrichment (odds ratio, y-axis) of aVMCs in genic features (x-axis).

A

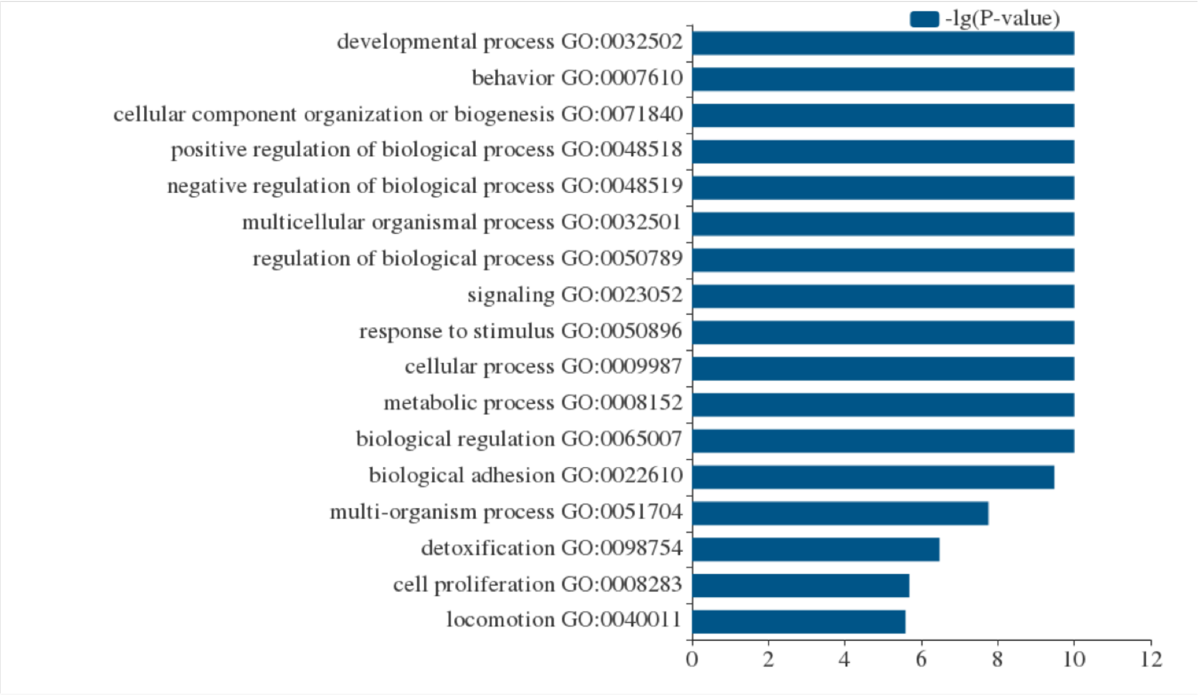
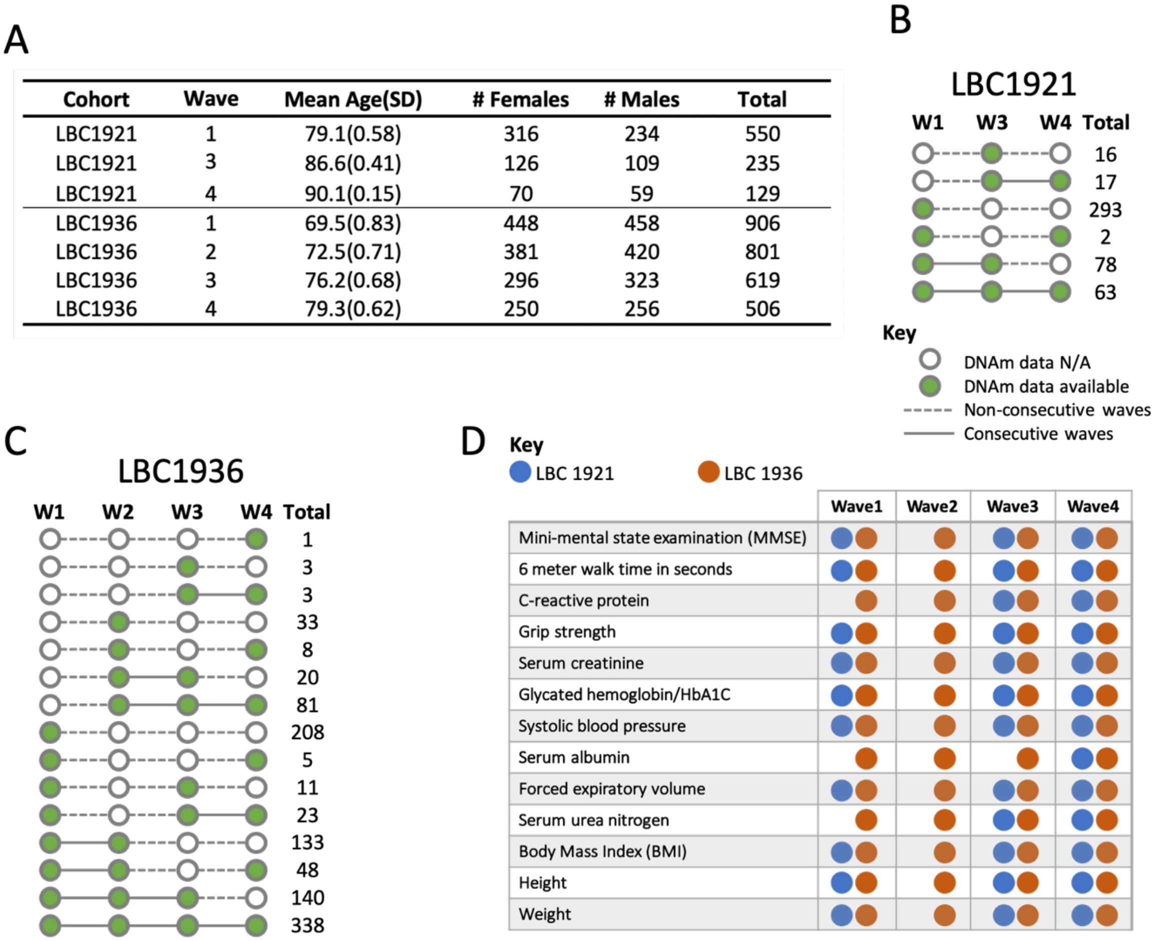


Figure 3: Annotation of genes associated with aVMC methylation
GO categories of regions enriched for aVMCs.



A

Cohort	Wave	Mean Age(SD)	# Females	# Males	Total
LBC1921	1	79.1(0.58)	316	234	550
LBC1921	3	86.6(0.41)	126	109	235
LBC1921	4	90.1(0.15)	70	59	129
LBC1936	1	69.5(0.83)	448	458	906
LBC1936	2	72.5(0.71)	381	420	801
LBC1936	3	76.2(0.68)	296	323	619
LBC1936	4	79.3(0.62)	250	256	506

B
LBC1921

W1	W3	W4	Total
○	●	○	16
○	●	●	17
●	○	○	293
●	○	●	2
●	●	○	78
●	●	●	63

Key

○ DNAm data N/A
● DNAm data available
----- Non-consecutive waves
—— Consecutive waves

C
LBC1936

W1	W2	W3	W4	Total
○	○	○	●	1
○	○	●	○	3
○	○	●	●	3
○	●	○	○	33
○	●	○	●	8
○	●	●	○	20
○	●	●	●	81
●	○	○	○	208
●	○	○	●	5
●	○	●	○	11
●	○	●	●	23
●	●	○	○	133
●	●	○	●	48
●	●	●	○	140
●	●	●	●	338

Key
🔵 LBC 1921 🟠 LBC 1936

	Wave1	Wave2	Wave3	Wave4
Mini-mental state examination (MMSE)	🔵🟠	🟠	🔵🟠	🔵🟠
6 meter walk time in seconds	🔵🟠	🟠	🔵🟠	🔵🟠
C-reactive protein	🟠	🟠	🔵🟠	🔵🟠
Grip strength	🔵🟠	🟠	🔵🟠	🔵🟠
Serum creatinine	🔵🟠	🟠	🔵🟠	🔵🟠
Glycated hemoglobin/HbA1C	🔵🟠	🟠	🔵🟠	🔵🟠
Systolic blood pressure	🔵🟠	🟠	🔵🟠	🔵🟠
Serum albumin	🟠	🟠	🟠	🔵🟠
Forced expiratory volume	🔵🟠	🟠	🔵🟠	🔵🟠
Serum urea nitrogen	🟠	🟠	🔵🟠	🔵🟠
Body Mass Index (BMI)	🔵🟠	🟠	🔵🟠	🔵🟠
Height	🔵🟠	🟠	🔵🟠	🔵🟠
Weight	🔵🟠	🟠	🔵🟠	🔵🟠

Figure 4: Data available from the Lothian Birth Cohorts
A) Table of characteristics of participants with phenotypic data available. For LBC 1921, 4 waves of data were collected, but DNAm data was only available for Waves 1, 3, and 4. **B)** A figure showing which LBC 1921 participants have DNAm data that passed QC for multiple waves; a green box indicates DNAm data is available at that wave. The Total number of participants with data available is displayed on the right. **C)** A figure showing which LBC 1936 participants have DNAm data available for multiple waves. **D)** A figure of aging-related phenotype data available for each LBC cohort across waves.

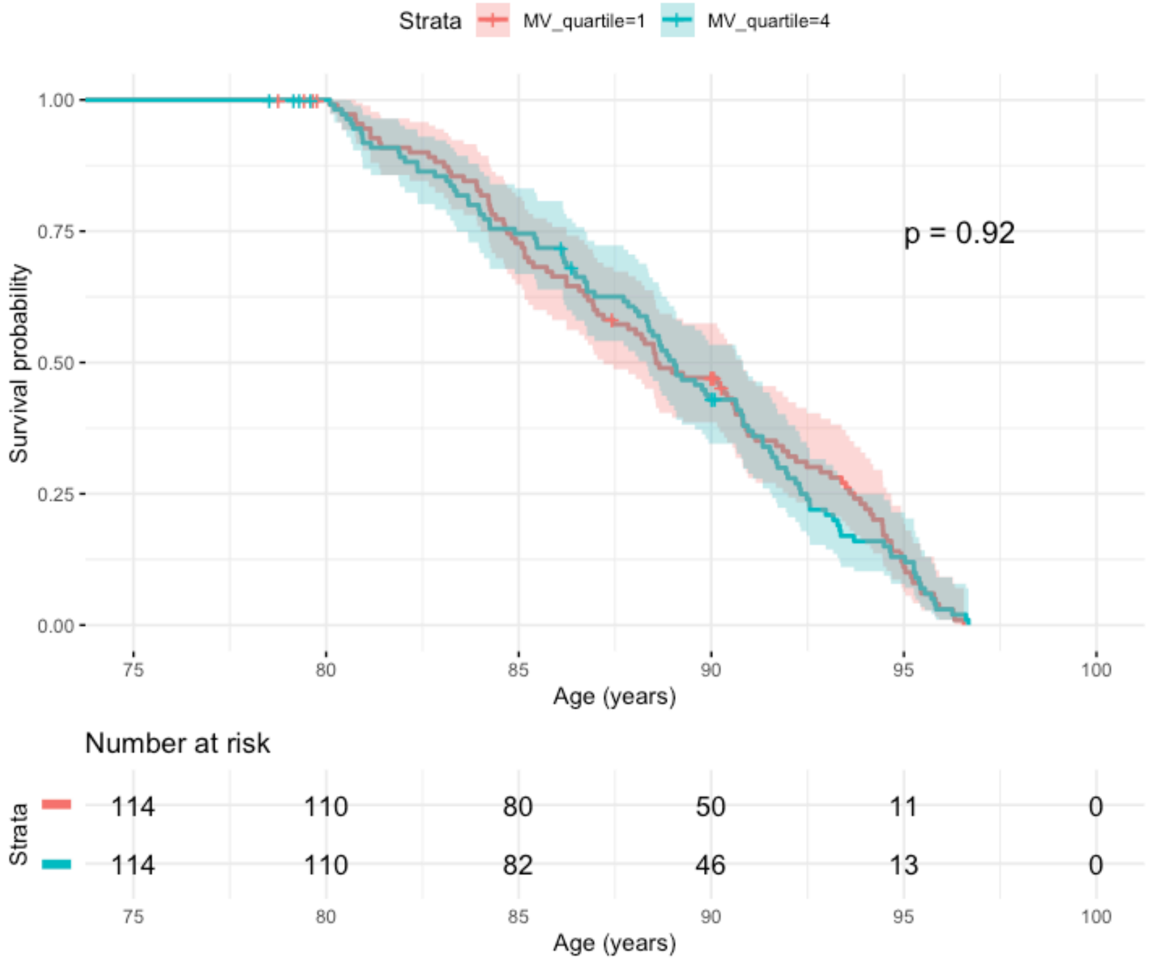


Figure 5: Survival probability by quartiles of MV score in LBC1921 by MV score
A figure indicating the survival probability (y-axis) of subjects by Age in years (x-axis) by MV score quartile (1st quartile is in red and 4th is in blue).

	MV score	MV score	Age	Age	Age^2	Age^2
Phenotype	**Estimate**	**p value**	**Estimate**	**p value**	**Estimate**	**P value**
BMI	-2.79E-04	0.645	0.0594	0.0350	-0.0131	0.0492
Creatine	1.81E-03	0.673	1.650	1.75E-09*	-0.190	3.94E-03*
FEV	-2.43E-05	0.736	-0.0236	3.38E-11*	-0.00134	0.101
Grip Strength	-1.14E-03	0.269	-0.829	1.62E-32*	0.0435	0.0436
Hemoglobin A1C	2.35E-06	0.987	-0.0410	2.06E-05*	0.00722	4.41E-05*
Height	-3.53E-05	0.963	-0.4280	1.02E-33*	0.0180	0.0117
MMSE	-2.00E-04	0.550	-0.0617	9.83E-03	-0.00487	0.420
6 Meter Walk Time	7.89E-05	0.877	0.2630	2.63E-10*	-0.00814	0.443
Systolic BP	3.67E-04	0.931	-1.480	4.55E-06*	0.0935	0.253
Urea	6.4E-04	0.480	-2.11E-03	0.969	0.0366	0.323
Weight	-3.21E-04	0.838	-0.1860	6.78E-03	-0.019	0.237

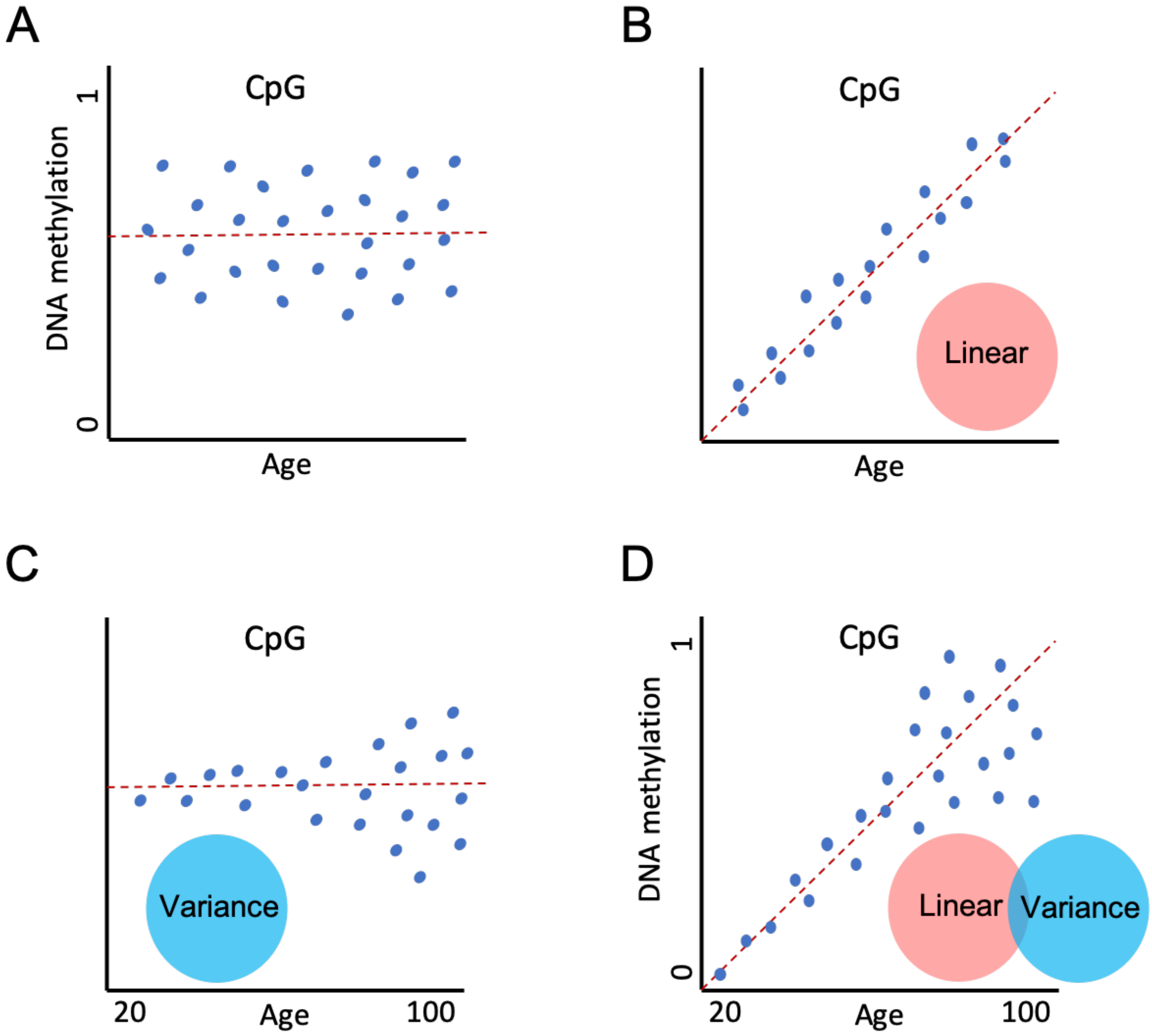
Table 1: Multivariate regression analysis of age-related phenotypes on age and MV score in LBC1921

The model includes the following covariates for the LBC1921 participants: age, age^2, and sex. p values marked with an asterisk (*) are significant at our Bonferroni-corrected α of 0.0045.

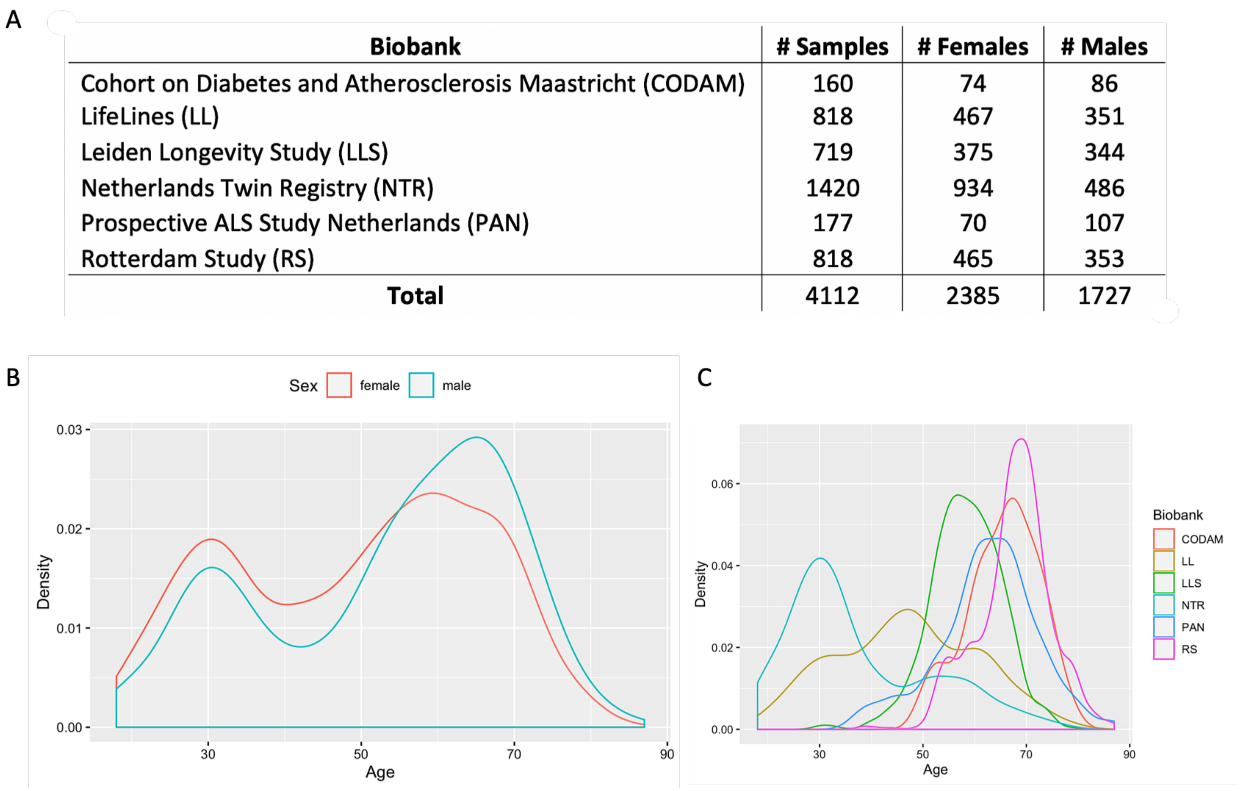
Phenotype	MV score Estimate	MV score p value	Age Estimate	Age p value	Age^2 Estimate	Age^2 p value
Album	-2.42E-04	0.360	-0.78500	0.00E+00*	-0.0108	0.0102
BMI	-1.39E-04	0.573	-0.0016	0.840	-0.0124	2.69E-08*
Cholesterol	-6.65E-05	0.499	-0.0613	1.48E-43*	0.00692	3.37E-08*
Creatine	3.72E-03	0.0311	0.1600	0.0247	0.00999	0.625
C-reactive Protein	8.74E-04	0.163	-0.1930	5.20E-07*	-0.0174	0.126
FEV	-5.31E-05	0.191	-0.0407	2.34E-143*	0.00164	8.60E-05*
Grip Strength	5.04E-04	0.363	-0.3090	2.33E-42*	-0.0217	6.31E-04*
Hemoglobin A1C	-5.99E-05	0.354	-0.00424	0.217	0.00548	6.65E-07*
Height	-3.46E-04	0.142	-0.170	8.58E-111*	-0.00215	0.282
MMSE	-1.33E-04	0.344	-0.0387	1.65E-07*	-0.00171	0.427
6 Meter Walk Time	3.26E-05	0.803	0.1500	5.96E-118*	0.0023	0.193
Systolic BP (mean)	-6.88E-04	0.678	-0.2910	3.27E-04	-0.0318	0.175
Urea	4.67E-04	0.0463	0.02690	0.0485	-0.00354	0.380
Weight	-5.27E-04	0.431	-0.1580	1.30E-13*	-0.0347	6.41E-09*

Table 2: Multivariate regression analysis of age-related phenotypes on age and MV score in LBC1936

The model includes the following covariates for the LBC1936 participants: age, age^2, and sex. p values marked with an asterisk (*) are significant at our Bonferroni-corrected α of 0.0036.

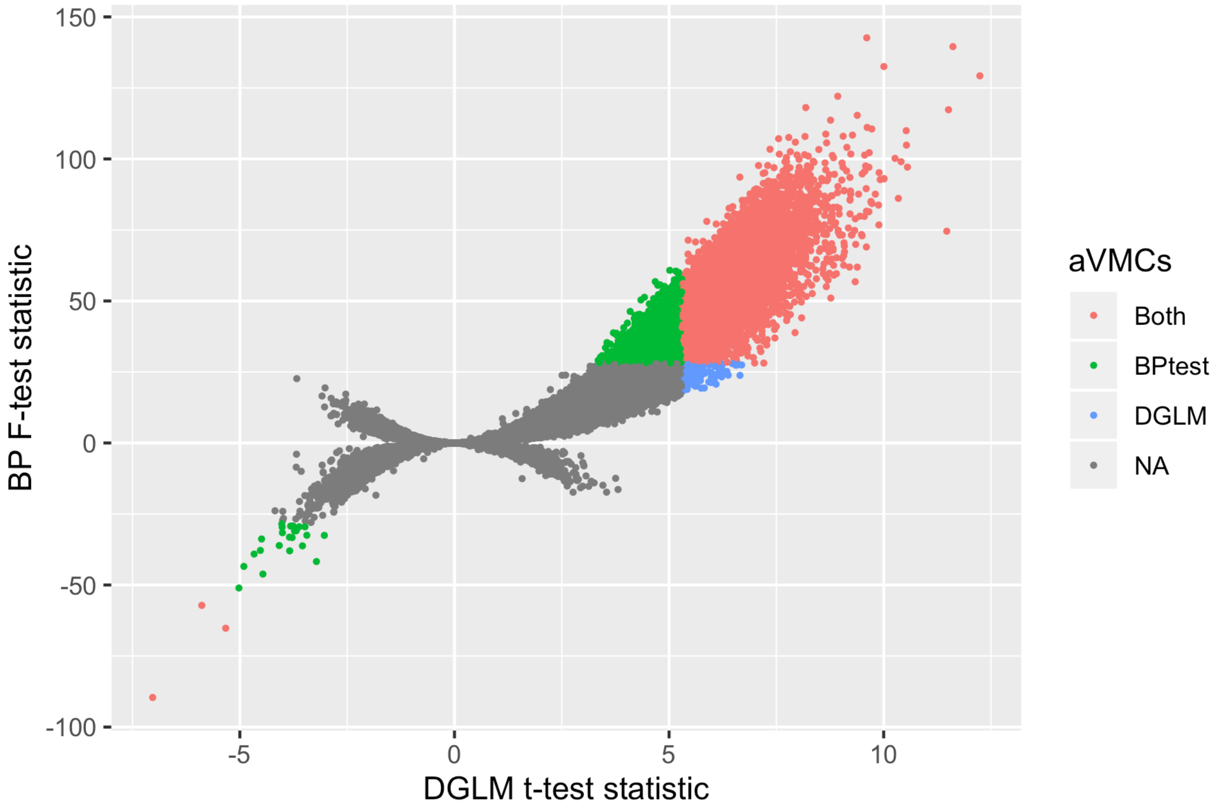


Supplementary Figure 1: Examples of CpGs Changing with Age
A) Figure shows an example of a CpG where DNAm (y-axis) does not change with age in years (x-axis). **B)** Figure shows an example of an aDMC, a CpG at which DNAm and age have a linear relationship. **C)** Figure shows an example of an aVMC, a CpG at which the variance in DNAm and age have a linear relationship. **D)** Figure shows an example of an aVMC at which DNAm changes with age both linearly and in its variance.

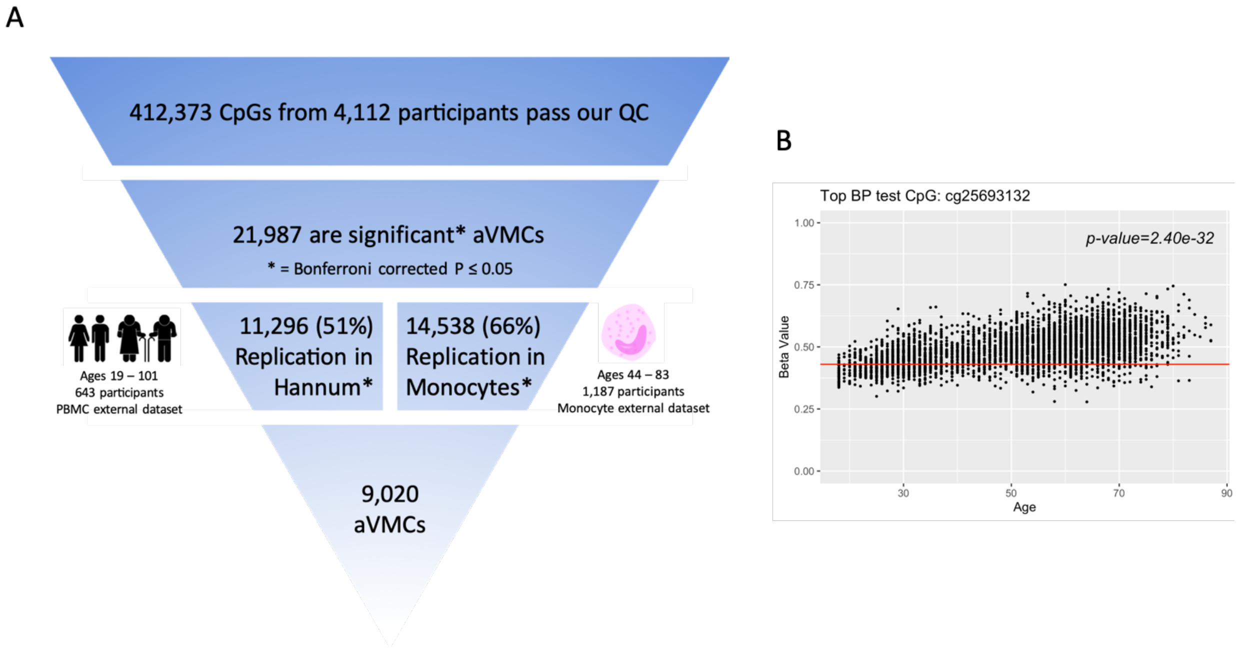


Biobank	# Samples	# Females	# Males
Cohort on Diabetes and Atherosclerosis Maastricht (CODAM)	160	74	86
LifeLines (LL)	818	467	351
Leiden Longevity Study (LLS)	719	375	344
Netherlands Twin Registry (NTR)	1420	934	486
Prospective ALS Study Netherlands (PAN)	177	70	107
Rotterdam Study (RS)	818	465	353
Total	4112	2385	1727

Supplementary Figure 2: Information about biobanks in the BIOS Consortium
A) Data on each of the six biobanks in the BIOS Consortium by number of samples and sex. **B)** A density plot of all participants' (n=4,112) age by sex. **C)** A density plot of all participants' age by biobank.

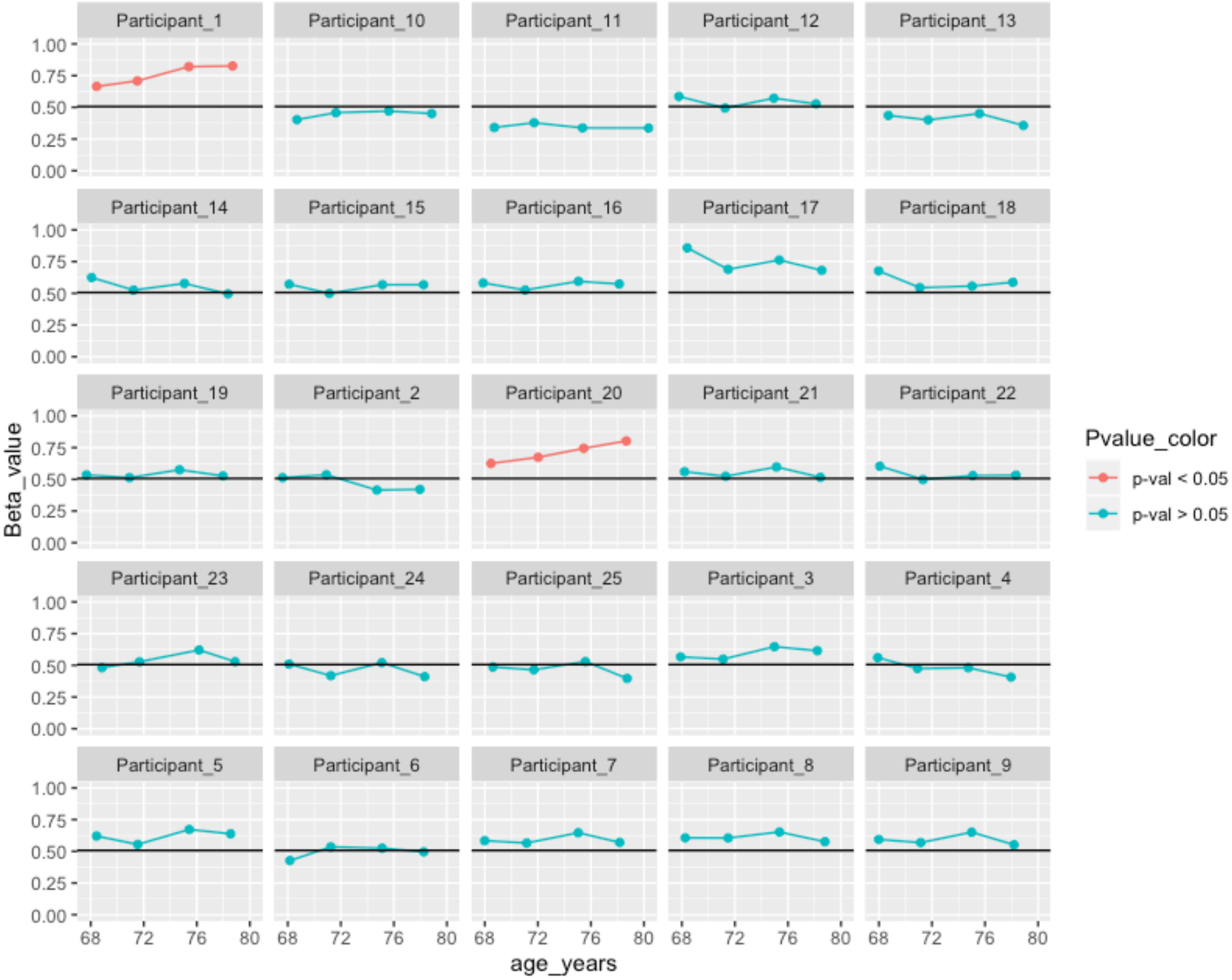


Supplementary Figure 3: Overlap of aVMCs identified using two statistical tests
Test statistics from DGLM (x-axis) are plotted against the test statistics from the BP test (y-axis), where each point represents one of the 412,373 CpGs tested. The BP test statistics have been multiplied by -1 for sites where there was negative change in the variance with age. Sites in grey (labeled NA) are not aVMCs according to a Bonferroni corrected P value < 0.05. Sites in blue are aVMCs according only to DGLM, sites in green are aVMCs according only to BP test, sites in red are aVMCs according to both tests. Sites in the upper right quadrant reflect CpGs reflecting increased variability with age in both tests, while sites in the lower left quadrant reflect CpGs reflecting decreased variability with age in both tests.

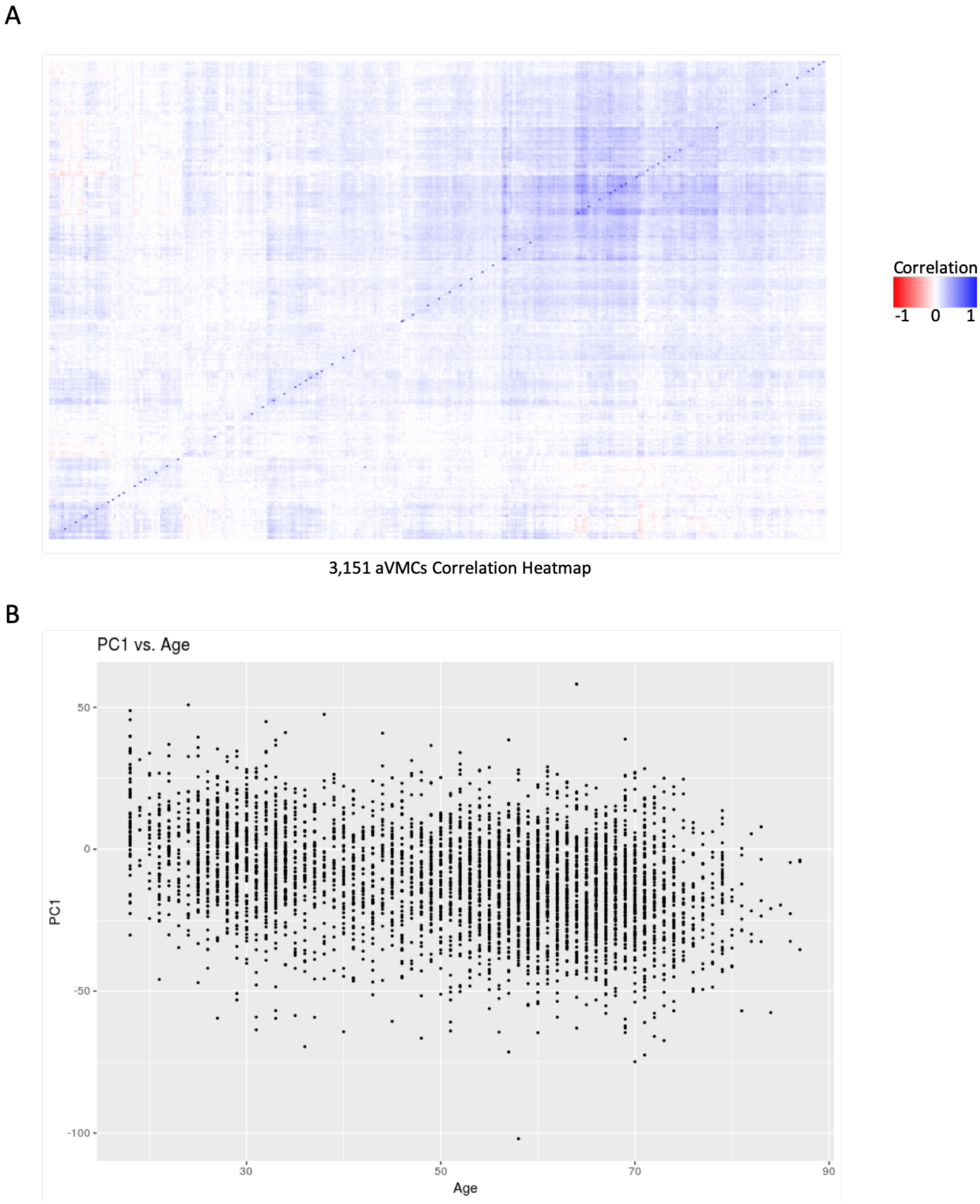


Supplementary Figure 4: Discovery and replication of aVMCs using Breusch–Pagan test (BP test)

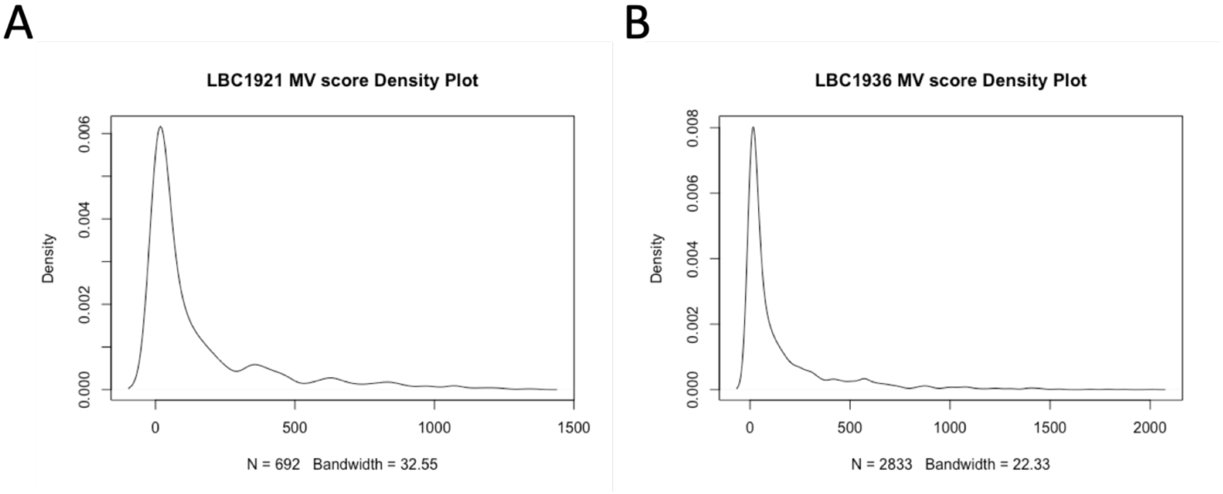
A) Flow chart of 9,020 CpGs identified as aVMCs, featuring replication rates in external datasets. **B)** The example aVMC pictured is the CpG with the lowest p-value (2.40 × 10-32) from BP test. Each point represents a participant, the y-axis is the methylation Beta value, and the x-axis is age.



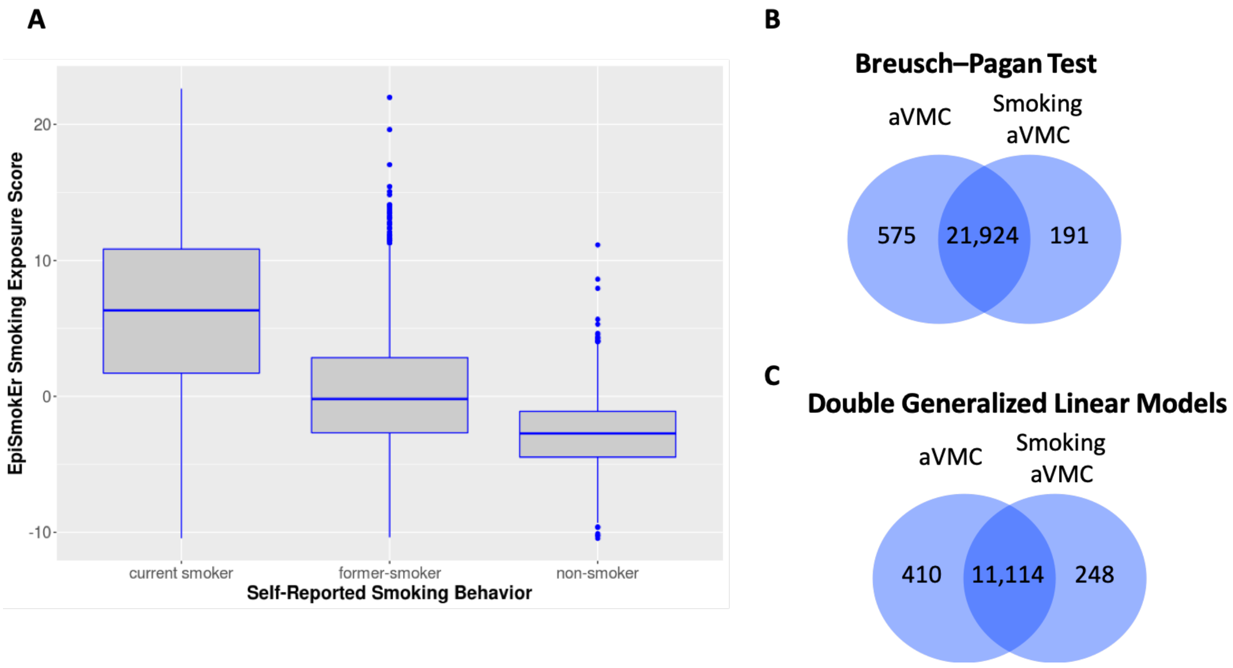
Supplementary Figure 5: DNAm patterns do not change significantly over time
Individual participants (n=25) from LBC1936 have their ages (x-axis) plotted against their Beta values (y-axis) for an aVMC chosen at random ("cg04884090"). A black horizontal line indicates the mean Beta value (0.506) among the Young group in BIOS at this aVMC. Points represent the Beta value corresponding to the age at that wave; points are connected in a line representing the trajectory of change at the aVMC over time. A simple linear regression for each participant was performed in which their Beta values were regressed on age at each wave; line color indicates the p-value of the slope of the age term (blue indicates a p-value > 0.05, red indicates a p-value ≤ 0.05). Though the DNAm patterns appear variable for some, a simple linear regression reveals the slopes are not significantly different from zero for the majority of participants, this significance disappears after correction for testing multiple participants.



Supplementary Figure 6: Methods for creating an MV score in BIOS
A) Correlation matrix showing the Spearman correlations between the 3,151 Beta-values. **B**) The ages of participants in BIOS (x-axis) were plotted vs. the first Principal Component (y-axis).



Supplementary Figure 7: Visualizing the methylomic variability score
A) Density plot of the LBC1921 MV scores. **B**) Density plot of the LBC1936 MV scores.



Supplementary Figure 8: Sensitivity analysis, effect of smoking on aVMC discovery
A) For the individuals (n=3,222) who self-reported smoking behavior (current, former, or non-smoker), a boxplot of their EpiSmokEr smoking exposure score (x-axis) is plotted against their self-reported smoking behavior (y-axis). **B**) The overlap in aVMCs identified between the aVMCs with smoking score included as part of a sensitivity analysis show that smoking doesn't affect aVMCs selected using the Breusch–Pagan test or **C**) Double Generalized Linear Models.

Chapter IV. Profiling Chromatin Accessibility and Histone Modification Changes in Aging: An Integrative Approach

Crystal D. Grant, Nicholas D. Johnson, Kasey J. Brennan, Hao Wu, Yujing Li, Yun Li, Trenell J. Mosley, Tessa R. Bloomquist, Douglas P. Kiel, Eric A. Whitsel, Peng Jin, Joanne M. Murabito, Andrea A. Baccarell, Karen N. Conneely

Abstract

Aging is marked by widespread alterations in the epigenome. Epigenomic studies have reported age-related changes in DNA methylation across the genome, while changes in chromatin accessibility and histone modifications are not as well-characterized. We performed ATAC-Seq, ChIP-Seq (H3K27me3, H3K4me3, and H3K27ac), and RNA-Seq on peripheral blood mononuclear cells from 20 healthy, Caucasian women (10 aged 23-30 and 10 aged 68-76). We identified 23 ATAC-Seq peaks of age-related differential chromatin accessibility (FDR<0.05); 19 of these reflected increased accessibility in younger women. We did not observe significant age-related differences in H3K27me3 or H3K4me3, while 49,742 sites showing age-differential H3K27ac were found across the genome. When sites were annotated to nearest genes, a weak positive correlation (r=0.17) was observed between log2-fold-change for age-differential H3K27ac and expression at 124 age-differentially-expressed genes. Genes annotated to regions showing increased H3K27ac in younger participants were enriched for biological processes related to development and neurogenesis, while those annotated to regions showing increased H3K27ac in older participants were enriched for immune response and metabolic processes. In conclusion, we report age-related changes in chromatin accessibility and H3K27ac, though these changes do not appear to recapitulate each other, and our findings suggest that H3K27ac may be useful in modeling age-related changes.

Introduction

Aging is associated with altered biological functioning and increased risk of morbidity and mortality.[1] Among the molecular hallmarks of aging are epigenetic changes,[13] in which reversible, heritable changes occur without changes to the underlying genetic sequence. Similar to disease, aging is marked by widespread, reproducible alterations to the epigenome thought to be characteristic of epigenetic dysregulation.[80] Because of the stability of DNA methylation (DNAm) and the relative ease with which it can be assayed, numerous studies have characterized the robust changes to DNAm with age in whole blood in humans.[87,99,101,102] In contrast, other structural levels of the epigenome, like broad histone post-translational modifications (PTMs) and overall chromatin structure, remain relatively poorly understood in how they change with age in blood.

The interaction between sequence-level DNAm and higher-level covalent histone PTMs are vital to proper chromatin structure and function.[142,220] Regions in the genome that feature differential methylation with age are enriched for histone PTMs associated with repression, such as H3 lysine 27 trimethylation (H3K27me3) associated with Polycomb repression, and H3 lysine 9 trimethylation (H3K9me3) associated with heterochromatin regions.[143] These modifications have the ability to induce changes in chromatin structure as well as recruit proteins important in chromatin regulation and gene expression.[37] Histone PTMs have also been found to reflect the aging process, with increases in H3K9me3,[155] H3K27me3,[157] and H4K20me3[156] (associated with constitutive heterochromatin) observed with age. In addition to their local contribution to transcription, histone PTMs have important functional consequences in establishing global chromatin environments.[37] The accessibility of chromatin domains is also related to DNAm patterns, with transcriptionally inactive heterochromatin showing an enriched for DNAm.[147] Because of their importance in chromatin function and their interrelatedness to DNAm, a better understanding of both histone modifications and chromatin accessibility will be informative in understanding a complex biological process like aging.

The assay for transposase-accessible chromatin with sequencing (ATAC-Seq)[43] allows for genome-wide profiling of chromatin accessibility. The locations of broad histone modifications can be profiled using chromatin immunoprecipitation with sequencing (ChIP-Seq).[52] The functional relevance of chromatin structure changes indicated by ATAC-Seq and ChIP-Seq can be quantified through observed transcriptional changes in RNA sequencing (RNA-Seq).[221] Two recent studies performed ATAC-Seq and RNA-Seq on blood samples among populations of different ages in order to elucidate the aging signature in peripheral blood mononuclear cells (PBMCs).[222,223] Focusing on CD8+ T cell subsets purified from PBMCs, Moskowitz *et al.* detailed how aging is accompanied by alterations in chromatin accessibility among naïve and central memory cells.[222] Using PBMCs, Ucar *et al.* found less accessibility with age at promoters and enhancers associated with T-cell signaling, as well as an increase in accessibility at quiescent and repressed sites thought to reflect stochastic epigenetic changes with age.[223] While these studies have shed light on the effects of aging on chromatin accessibility and the subsequent influence on gene expression, age-related changes at the level of histone modifications were not assayed.

Here, we present the results of a pilot study with the aim of uniting all three assays (ChIP-Seq, ATAC-Seq, and RNA-Seq) to leverage the different information contributed by overall chromatin accessibility and three histone modifications. The histone modifications profiled include H3K4 trimethylation (H3K4me3, associated with active promoter), H3K27 acetylation (H3K27ac, enhancer regions), and H3K27 trimethylation (H3K27me3, a repressive mark).[66] The aim is to better understand age-related changes in these marks and to elucidate possible functional roles through their associations with gene expression in peripheral blood mononuclear cells (PBMCs).

Results

Profiling Epigenetic Changes with Age

PBMCs were isolated from 20 healthy Caucasian women: 10 participants aged 23 to 30 years ("Young") and 10 participants aged 68 to 76 years ("Old"). ATAC-Seq, ChIP-Seq, and RNA-Seq were performed (detailed in **Methods** and **Figure 1**). The histone modifications profiled via ChIP-Seq were H3K4 trimethylation (H3K4me3), H3K27 acetylation (H3K27ac), and H3K27 trimethylation (H3K27me3).

Age-Related Changes in Chromatin Accessibility

ATAC-Seq data was used to generate genome-wide maps of chromatin accessibility. A total of 4,430 distinct chromatin accessibility peaks were identified that were present in more than one participant. Differential analysis by age group ("Young" vs "Old") was performed for samples passing QC (**Figure 1** and **Supplementary Tables 1-4**). 23 differentially accessible regions (DARs) due to differences in age group were identified (FDR < 0.05). Comparing regions more accessible in Young and Old groups, 19 DARs are observed to be more accessible in the Young group and 4 more accessible in the Old group (**Figure 2A**). The ATAC-Seq peaks occur throughout the genome and the predicted chromatin states of the peaks, obtained from the Roadmap Epigenomics Project,[224] can be observed in **Figure 3A**. In order to test for enrichment in chromatin states among DARs, a relaxed criterion (FDR<0.2) was used to define differentially accessible regions. The DARs that are less accessible with age appear more likely to occur at enhancer regions (OR=1.47) and sites associated with transcription (OR=1.50), while DARs that are more accessible with age appear more likely to occur at quiescent or repetitive regions (OR=1.68) (**Figure 3B**). However, given the small numbers of DARs, the p-values of enrichment tests do not indicate significance (Fisher's exact test; $0.09<P<1$).

Age-Related Changes in Histone Modifications

Next, using ChIP-Seq, the relationship between chromatin accessibility and chromatin modifications was examined for H3K4me3, H3K27me3, and H3K27ac. For samples passing QC (**Supplementary Tables 2-4**), differential analysis by age group was performed. For H3K4me3, 108,629 peaks were called in at least two participants, and these regions were tested for age-differential enrichment of H3K4me3. No regions were considered significantly age-differentially represented (FDR<0.05). Similarly, for H3K27me3, 401,265 peaks were called in at least two participants and no regions were significantly age-differential after correction for multiple testing.

For H3K27ac, 306,204 peaks were called in more than one person; comparing regions more accessible in Young or Old groups, 49,742 regions were found to be significantly differentially represented between the Old and Young groups (FDR < 0.05). Of these, 14,687 showed an enrichment of H3K27ac in the Young group, while 35,055 were more enriched in the Old group (**Figure 2B**).

Comparing Age-Related Changes Across Assays

To indicate the functional relevance of the chromatin openness and chromatin organization patterns identified, age-differential peaks were integrated with age-differential gene expression results previously estimated by our group for the same 20 samples.[225] Among the 124 differentially expressed genes (DEGs) by age group that were identified, 79 sites showed increased expression in the Young group and 45 sites showed increased expression in the Old group. Data from the three assays, ATAC-Seq, H3K27ac ChIP-Seq, and RNA-Seq can be observed in **Figure 4**, where only regions found to show significant age-related differences are plotted for each assay.

ATAC-Seq peaks were annotated to their nearest gene using HOMER (results of the ATAC-Seq peaks can be observed in **Table 1**). Of the genes annotated to our 23 DARs, none corresponded to the 124 DEGs—suggesting that, though there are age differential changes in chromatin accessibility especially enriched at promoters, they are not driving significant age-related differences in gene expression Two genes annotated to DARs – AKAP7 and DOLPP1 – were identified as DEGs in a larger (N=14,893) study that identified 1,497 age-related DEGs in whole blood.[127] Expression decreased with age in both genes but the changes in accessibility reflect a decrease with age in DOLPP1 and an increase in AKAP7.

To integrate the differential H3K27ac peaks with DARs, H3K27ac peaks found to be age-differential were also annotated to their nearest gene using HOMER. Additionally, a test for overlaps between the two types of peaks was performed, finding that none overlapped. was used to test for overlap between the two types of peaks—finding that none overlapped. This is likely because the locations assayed comprise non-overlapping chromatin states in the genome, with the ATAC-Seq peaks often occurring in promoters and the H3K27ac peaks occurring in enhancers. 81 of the ChIP-Seq peaks were annotated to 9 unique genes also annotated to DARs, including *ADAM22, AKAP7, ASB7, CYLD, CYTH2, ELOVL5, FAM117A, FBXO11,* and *TLE4*. For *CYTH2* both ChIP-Seq and ATAC-Seq feature an enrichment of peaks in the Young group; for *AKAP7* and *ADAM22*, both assays feature an enrichment in the Old group. For the other six genes, the test statistics from the two assays did not agree in direction, suggesting that H3K27 acetylation and accessibility, mostly at promoter regions, do not co-occur in these regions.

To investigate whether peaks reflecting age-differential H3K27ac annotate to genes reflecting differential expression in aging, annotated genes from H3K27ac were compared to the 124 DEGs. The DEG and H3K27ac peaks were merged, identifying 203 H3K27ac peaks corresponding to 58 unique DEGs (detailed in **Supplementary Table 5**). A weak positive correlation was observed between the test statistics of genes linked to each of the assays (r=0.174, p=0.0129; **Supplementary Figure 1**), suggesting that the H3K27ac mark may be associated with expression for some of these regions. Gene ontology (GO) analysis showed that the 4,418 unique gene annotations reflecting increased H3K27 acetylation in the Young group were significantly enriched for genes reflecting 545 unique GO terms, many of which are involved in biological processes in development and neurogenesis. The 7,183 unique gene annotations reflecting increased H3K27 acetylation in the Old group were enriched for 380 unique GO terms, many of which were involved in immune response and metabolic processes. The top ten GO terms for each of the two groups are shown in **Table 2**.

Discussion

This is a pilot study with the aim of elucidating which levels of the epigenome can be used to model age-related changes. Because of its stability and the ease with which it can be assayed, DNAm is usually at the center of studies detailing changes with age in the epigenome. In this study, we characterized changes at the level of overall chromatin accessibility and individual histone modifications, confirming that age-related changes happen at several levels of the epigenome. We found that some of the modifications assayed appear to undergo widespread changes with age, while others, namely H3K27 trimethylation and H3K4 trimethylation, do not appear to show significant changes with age.

Among the 23 DARs annotated to genes, many have been previously linked to aging phenotypes and age-related disease in changes in their gene expression or their DNAm patterns. Among these was *CYLD* (log2 fold change = -1.56, *P*=0.0087), with the region being significantly more open in the Young group. The *CYLD* gene encodes an enzyme, the CYLD lysine 63 deubiquitinase,[226,227] which acts as a tumor suppressor. Interestingly, a loss of function of this enzyme has been linked to premature aging in mice.[228] Another peak found to be more accessible in the Young group was annotated to *HSPA5* (fold change = -1.304, p-value = 7.25× 10^{8}). Aging leads to lower levels of HSPA5 in mice, leading to ER stress; it is also thought to play a role in the development of age-related disease.[229-232]

A peak found to be more accessible in the Young group was annotated to Elongation of very long chain fatty acids 5 (*ELOVL5*). Interestingly, this gene belongs to the same family as ELOVL2, which shows robust, age-related changes in DNAm across tissues.[92,100] Also among these sites, the gene corresponding to the ATAC-Seq peak significantly more open in the Old group was *AKAP7* (fold change =-1.07, p-value= 0.0418). The expression of A-kinase anchoring protein 7 (AKAP7) in blood was recently found to serve as a reliable biomarker for patients at risk of post-stroke blood brain barrier disruption.[233] Eight of the 23 ATAC-Seq peaks found to be age-differential in this study were annotated to genes also identified as age differentially accessible in a larger ATAC-Seq study in blood by Ucar *et al.*[223] (these sites are marked with an asterisk in **Table 1**).

Similar to the study approach by Ucar *et al.*, we performed ATAC-Seq and RNA-Seq in PBMC samples of different ages. Though we had a smaller sample size (n=20 compared to n=77) we were able to replicate some of the findings of regions of age-related chromatin accessibility changes. Our replication of only some of these findings may result from our smaller sample size and limited power, which represents a major limitation of our study. Another limitation is that our assay was performed in PBMCs. Without accounting for the composition of different cell types in the data analyses, it is possible that some of the observed age-related differences are related to the change of cell composition. Interestingly, several genes linked to premature aging or age-related diseases like cancer were identified by us to reflect differential accessibility despite not accounting for age-related changes in cell types. Some of these genes have previously only been linked to these phenotypes in mice, suggesting that similar mechanisms may be at work in humans. Another potential limitation is that our cross-sectional (rather than longitudinal) study design leaves open the possibility that observed differences could reflect cohort effects, mortality selection, or inter-individual differences. Our study did focus on a set of individuals that were matched on sex, broad ancestral group (all Caucasian), and geographic location, which will limit confounding due to inter-individual heterogeneity; however, this may also limit generalizability of our findings.

A key finding of our study is the widespread changes in H3K27 acetylation with age in blood. While the next step should be replication in a set of independent samples, this finding may help inform future aging studies as to which histone modifications are useful to assay in developing models of age-related epigenetic changes. This is reinforced by the findings from the GO analysis that the sites marked by more H3K27ac in Young are involved in developmental processes while those in the Old are involved in immunological responses, supporting the functional relevance of this modification. In addition, the changes at the sites marked by this modification, which often marks active enhancers, could shed light on the importance of enhancers in mediating age-related changes in a cell. Lastly, considering the inter-relationship between DNAm and histone modifications, our findings of widespread age-related changes at H3K27ac suggest that future studies of aging, including DNAm studies, would benefit from focusing on age-related changes occurring at enhancers.

Methods

Human subjects

Whole blood samples were collected at iSpecimen, an independent, contract research organization, after receiving approval by Quorum Review IRB. Donors were healthy non-smoking Caucasian females not currently taking medication. They belonged to two age groups, a younger group aged 20-30 years and an older group aged 68-76 years. Donors from each group were asked to provide their age, medical history, including past conditions and treatments, and current medications to determine whether they qualified. Ten recruited donors from each group were scheduled and blood was drawn between 2-3pm and shipped overnight in refrigerated shipping containers. PBMCs were isolated from whole blood samples using density gradient media and centrifugation.

ATAC-Seq library generation and preprocessing

ATAC-Seq libraries were generated and preprocessing was performed according to the protocol previously described.[43] All samples were frozen in a DMSO/DMEM medium; cells were thawed for 1 minute in a 37 °C warming cabinet and transferred to a 50 mL Falcon tube. Cells were then suspended in 20 mL of warm 1x PBS followed by a 5 minute centrifugation at 1000 rpm to pellet the cells. Cells were lysed with a cold lysis buffer and spun to pellet the nuclei. Two million unfixed nuclei were tagged using a Tn5 transposase reaction mix (Illumina Nextera DNA Library Prep Kit) for 30 min at 37 °C, and the resulting library fragments were purified using the Qiagen Minelute kit. Libraries were amplified with 10-12 PCR cycles and finally sequenced on an Illumina HiSeq 2500, generating paired end 50 bp reads.

ATAC-Seq data analysis was performed using the following tools and versions: Samtools v1.5, Picard v2.6.0, Bowtie2 v2.2.6, macs2 v2.1.1.20160309, and bedtools v2.25.0. Adapters were trimmed from the reads using trimmomatic.[234] Paired end trimmed reads were then aligned to hg38 using bowtie2,[235] with standard parameters and a maximum fragment length of 2 kb (-X 2000). Duplicate reads were removed using Picard. De-duplicated reads were then filtered for alignment quality (MAPQ ≥ 30) and were required to be properly paired (Samtools flag 0 × 2). Reads mapping to the mitochondria, unmapped contigs, and sex chromosomes were removed and not analyzed.

To identify regions of accessibility in the ATAC-Seq data, peak calling was performed individually on each of the 20 samples using MACS2.[236] MACS2 was run with the model building option turned off (--nomodel), 100-bp shift, 200-bp extension, and only peaks called with a peak score (q-value) of 1% or better (--q 0.01) were kept for each sample. Peaks overlapping regions from the consensus excludable ENCODE blacklist,[66] designed to remove regions that show artifacts due to deficiencies in the genome assembly, were removed from future analysis. Additionally, only peaks called on autosomal chromosomes were used in this study.

ChIP-Seq library generation and preprocessing

PBMC samples had nuclei prepared and were crosslinked and frozen. ChIP was then performed in parallel using antibodies to each of three histone modifications, H3K4me3, H3K27me3, and H3K27ac. Libraries were generated using the NEBNext ChIP-Seq Library Prep Reagent Set for Illumina according the manufacturer's protocol. Fifty-cycle single-end sequencings were performed using Illumina HiSeq 2000. Samples that failed QC were recaptured and re-sequenced in a different batch. Single end reads were then aligned to hg38 using bowtie2,[235] with standard parameters. Sorted BAM files were filtered for alignment quality (MAPQ ≥ 30). Reads mapping to the mitochondria, unmapped contigs, and sex chromosomes were removed and not analyzed. Duplicate reads were removed using Picard. To be eligible for analysis, samples were required to have: at least 10 million reads sequenced, a minimum of 50% reads mapping, and a maximum duplication rate 50%; details on the number of samples passing this quality control (QC) step are available in **Supplementary Tables 1-4**.

Peak calling was performed with MACS2[236] on single end BAM files using the parameters –nomodel, –nolambda, –broad, –keepdup all, in order to identify regions of enrichment. MACS2 was used to identify three types of regions: (1) narrow peaks passing a p-value threshold (-p) of 0.01, termed narrowPeaks; (2) broader regions of enrichment passing an additional broad-peak cutoff p-value of 0.1 (-p 0.01; --broad; --broad-cutoff) termed broadPeaks; and (3) gapped regions of enrichment defined as broadPeaks containing at least one strong narrowPeak, termed gappedPeaks. Following the approach of Kellis *et al.*,[237] the gappedPeak representation was used for the histone marks with relatively compact enrichment patterns (H3K4me3 and H3K27ac). The broadPeak representation was used for more diffuse histone marks (H3K27me3). Peaks overlapping blacklisted regions were removed from future analysis, as were peaks called on non-autosomal chromosomes.

RNA-Seq library generation and preprocessing

Total RNA was extracted from PBMCs for each subject using the QIAGEN RNeasy Kit. Next, cDNA was synthesized using 1ug of the extracted mRNA with Invitrogen Oligo(dT)20 primers, the cDNA was then amplified using PCR. RNA-Seq libraries were generated using 0.5 μg of cDNA via the TruSeq RNA Sample Preparation Kit v2 (Illumina). Libraries were validated by DNA Chips via the Agilent 2100 Bioanalyzer. Libraries were then sequenced using 50-cycle single-end runs via the Illumina HiSeq 2000. RNA-Seq reads were aligned to hg38 using the STAR aligner.[238] Reads shorter than 50 bps and with quality scores below 20 were discarded.

Differential analysis

First, peaks called for the 4 assays (ATAC, H3K4me3, H3K27me3, H3K27ac) were merged into a consensus peakset for each assay using the BEDTools multiinter tool.[239] Peaks present in only one subject were removed from future analysis and a peak was required to have a mean read count of 5 to be included in differential analysis. For quality control in the RNA-Seq data, read count matrices were generated and regions with fewer than 10 reads in 5 or more samples were excluded. Next, reads overlapping consensus peaks were counted and these counts compared in an age-differential analysis using DESeq2.[240] Briefly, DESeq2 models read counts as a negative binomial distribution with a normalized mean to account for different library sizes across individuals. It then uses an Empirical Bayes approach to estimate log fold change and dispersion. A Wald test is used to test for significantly differential read counts between the two groups. Either batch (in the case of ATAC-Seq) or read length (in the case of ChIP-Seq) was included as a covariate in the model (the RNA-Seq was performed as one batch). Age group (Old or Young, with Young as the reference) was included as the predictor variable and read counts as the outcome. A p-value for each region was calculated, indicating confidence that the peak is differentially present in the Young and Old groups. Age differential sites were defined based on a false discovery rate criterion (FDR<0.05), using the Benjamini-Hochberg procedure. Lastly, the circularize package in R[241] was used to generate a circos plot of all the sites found to be significantly age-differential.

Peak annotation and downstream analysis

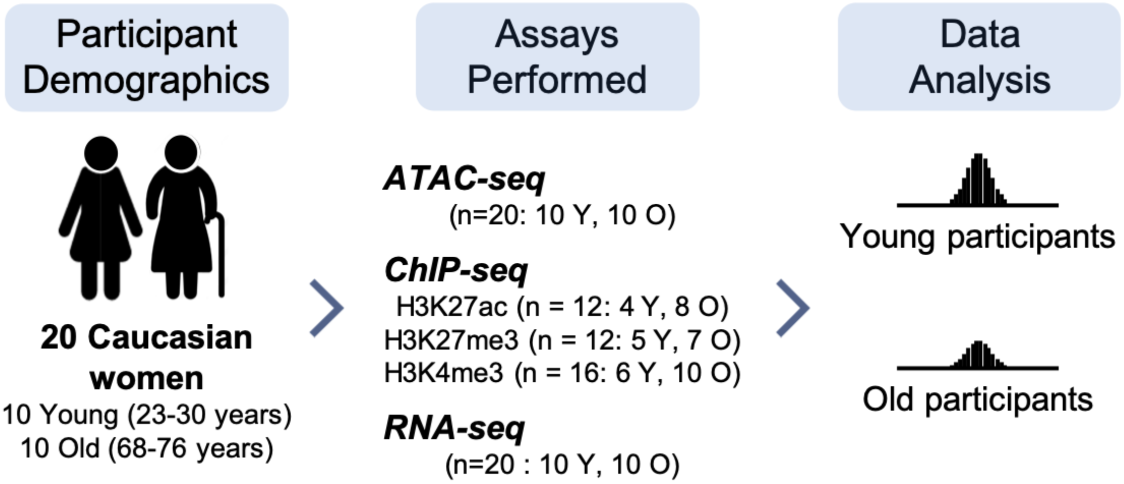
For functional annotation of peaks, the 18-state ChromHMM chromatin state annotations for PBMCs were obtained from Roadmap Epigenomics.[224] Overlaps between the peaks and these annotated regions of the genome were then identified. ChromHMM chromatin states "1_TssA," "2_TssFlnk," "3_TssFlnkU," "4_TssFlnkD," and "14_TssBiv," are labeled as Promoters. Chromatin states labled "5_Tx," and "6_TxWk," are labeled as

Transcription. Chromatin states “7_EnhG1,” “8_EnhG2,” “9_EnhA1,” “10_EnhA2,” “11_EnhWk,” and “15_EnhBiv” are labeled as Enhancers. Chromatin states “12_ZNF/Rpts,” and “18_Quies,” are labeled as Quiescent/Repetitive. Chromatin states “13_Het,” “16_ReprPC,” and “17_ReprPCWk,” are labeled as Repressed/Heterochromatin.

Peaks found to be age-differential were annotated to the hg38 genome using HOMER.[242] Rsamtools, GenomicFeatures, and GenomicAlignments R Bioconductor packages were used to count reads of RNA-Seq regions found to reflect age-differential expression changes overlapping each Ensembl-annotated gene of hg38.[243] The subsetByOverlaps command from the GenomicRanges package[243] was used to test for overlaps between different assays. When large sets (>100) of age-differential regions were identified, GO analysis on genes annotated to these regions was performed using the ideal and goseq Bioconductor packages in R.[244,245] Because 15,971 GO terms were tested, a Bonferroni corrected α of 3.13×10^{-06} was applied to identify significantly enriched GO terms.

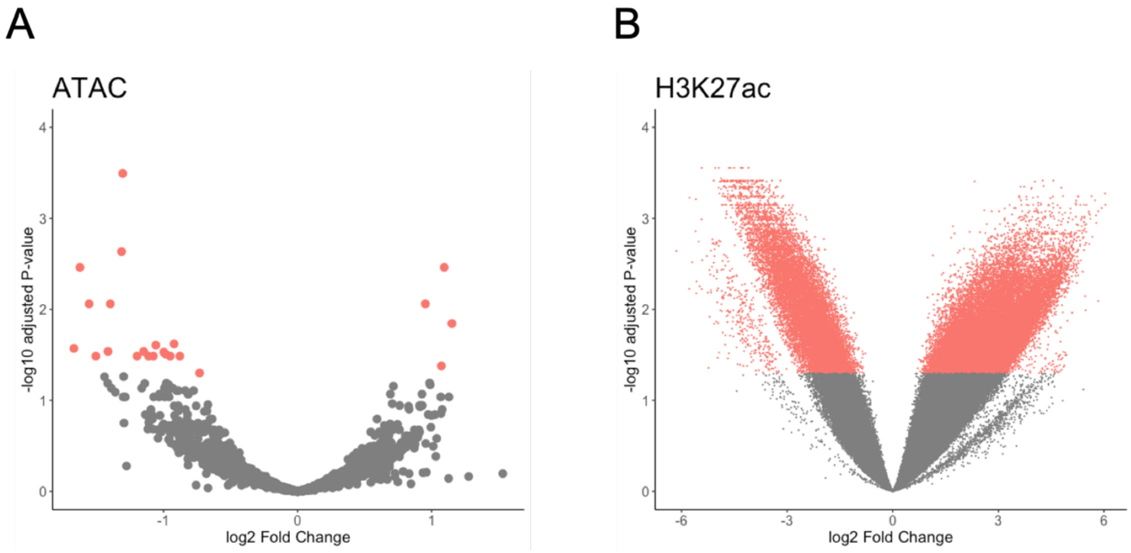
Tables and Figures

Figure 1: Overview of Assays and Analyses Performed



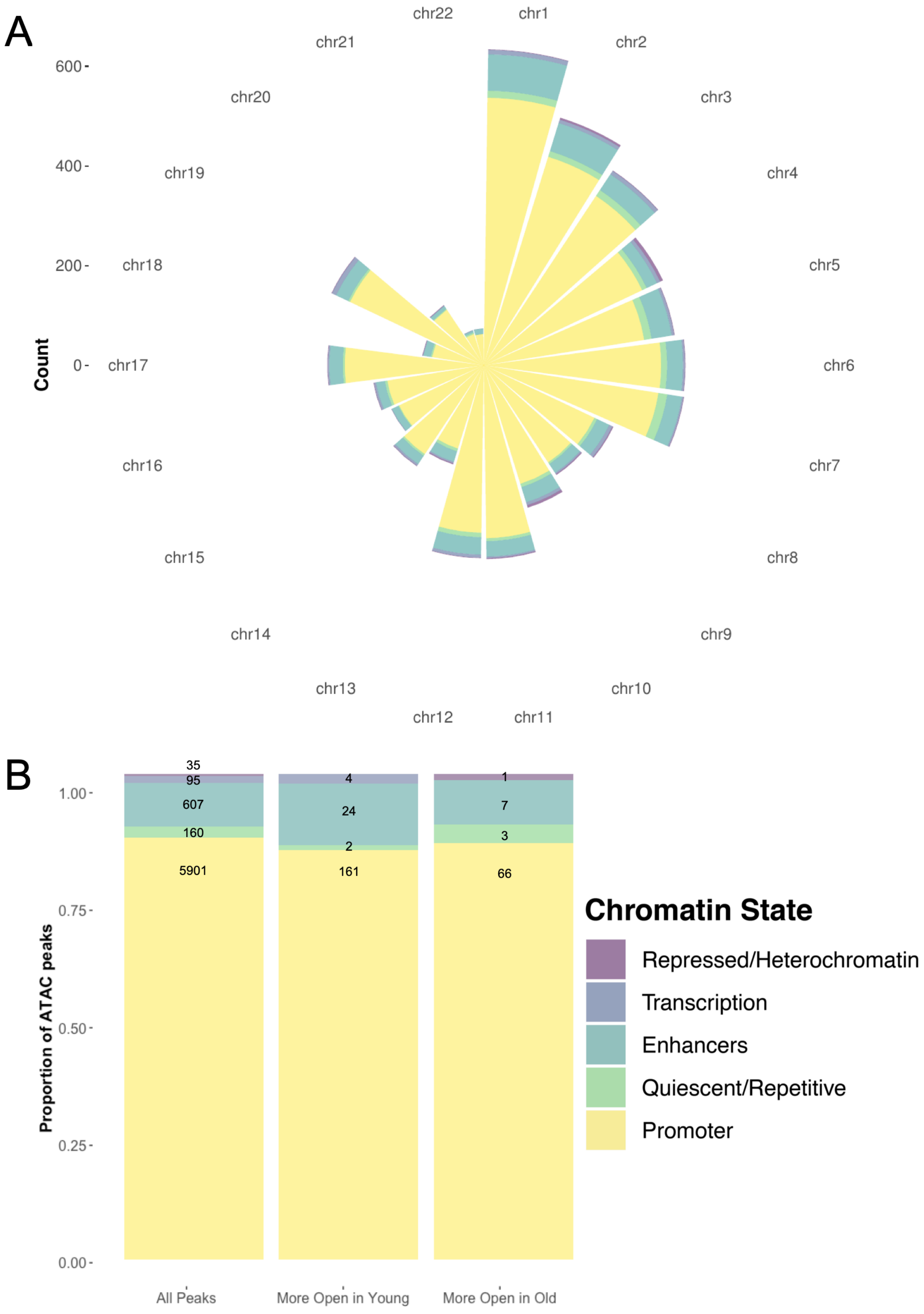
A schema summarizing the sample participants, high-throughput sequencing assays performed, and data analysis in this study. The total number of participant samples (n) in each assay are detailed as well as the specific number of young (Y) and old (O) participants.

Figure 2: Age-Differential Regions of Accessibility and H3K27 Acetylation



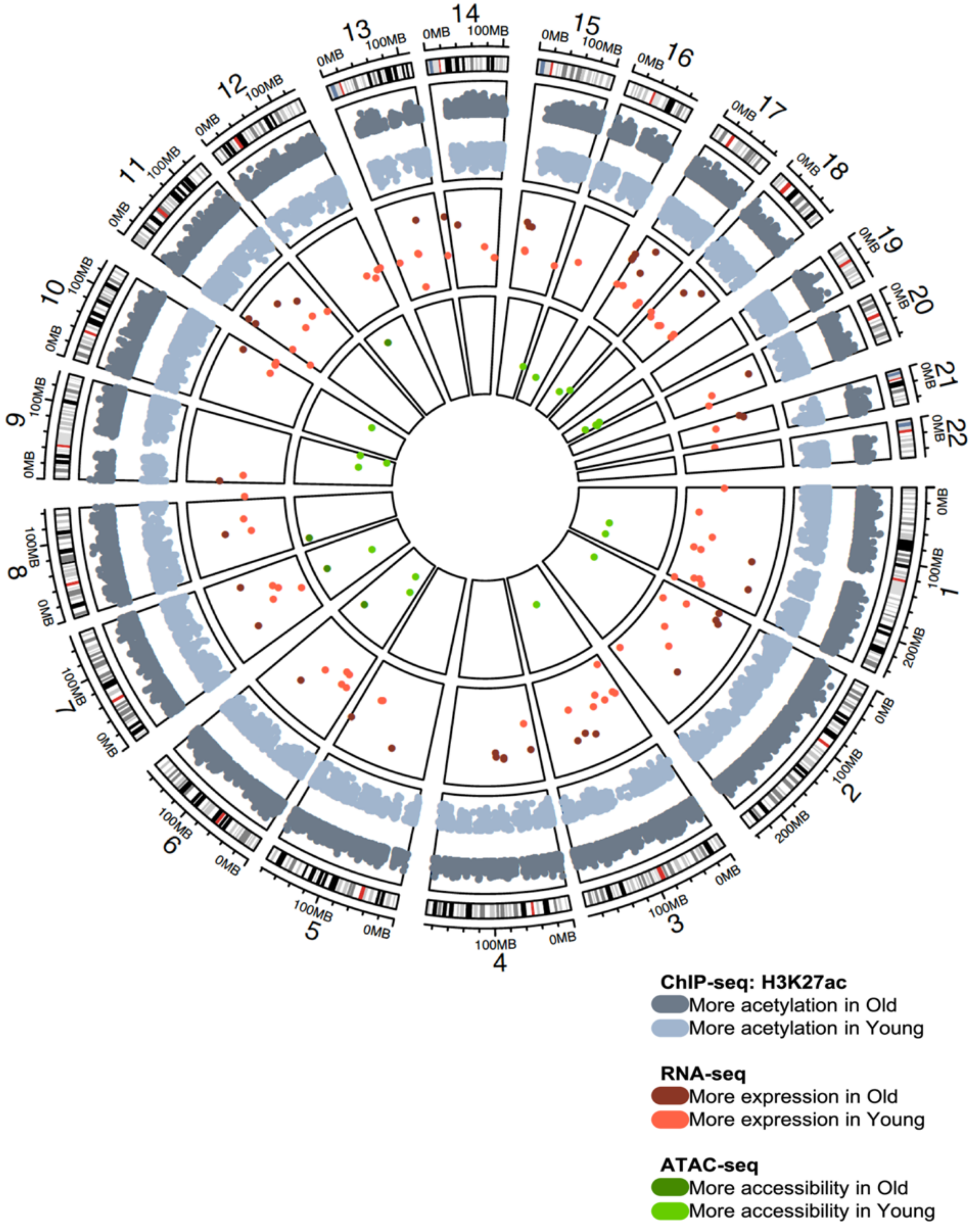
A) A volcano plot of the 4,430 peaks called by ATAC-Seq, where each point represents a peak present in at least 2 particpants. The points in grey represent peaks not found to be age differential while the points in red represent the 23 peaks found to be age differential after multiple testing correction. **B)** A volcano plot of the 306,204 peaks called by the H3K27ac ChIP-Seq. The points in grey represent peaks not found to be age differential while the points in red represent the 49,742 peaks found to be age differential after multiple testing correction.

Figure 3: Genomic Locations and Chromatin States of ATAC-Seq peaks



A) Genomic locations by chromosome of the 4,430 consensus ATAC-Seq peaks by Roadmap annotated chromatin state. Bar height indicates the number of peaks present on the chromosome. **B)** Proportions of the functional states of ATAC-Seq peaks for: all consensus ATAC-Seq peaks ('All Peaks'), peaks found to be less accessible with age ('More Open in Young'), and peaks found to be more accessible with age ('More Open in Old'). For the enrichment analysis in this figure, a relaxed criterion (FDR<0.2) was used to define differentially accessible peaks. Counts of chromatin state features represented by each part of the bar are noted in that portion of the bar.

Figure 4: Circos plot of Age-Differential Sites Across the Genome by Assay



49,742 sites were identified throughout the genome showing age-related differential H3K27 acetylation (blue). 14,687 sites were more enriched in the Young group (light blue) and 35,055 sites were more enriched in the Old group (dark blue). 124 sites showed differential expression (red) by age group throughout the genome. 72 sites showed increased expression in the Young group (light red) and 39 sites showed increased expression in the Old group (dark red). 23 sites were found to show differential accessibility (green) by age group. 19 sites showed increased accessibility in the Young group (light green) and 4 sites showed increased accessibility in the Old group (dark green).

Table 1: Gene Annotations of the 23 Age-Differential ATAC-Seq Peaks

Gene Name	Gene Description	log2 FoldChange	Annotation	Relationship to Aging Phenotypes
ACTR1A	ARP1 actin-related protein 1 homolog A, centractin alpha	-1.148	intron	
ADAM22	ADAM metallopeptidase domain 22	0.953	promoter-TSS	
AKAP7*	A-kinase anchoring protein 7	1.072	intron	Higher expression among patients at risk of post-stroke complications[233]
ASB7	ankyrin repeat and SOCS box containing 7	-0.922	promoter-TSS	
CIC	capicua transcriptional repressor	-1.623	intron	
CNOT3	CCR4-NOT transcription complex subunit 3	-1.414	promoter-TSS	Levels are reduced in aging-induced osteoporosis[246]
CYLD	CYLD lysine 63 deubiquitinase	-1.556	promoter-TSS	Tumor suppressor, loss of function leads to premature aging[228]
CYTH2	cytohesin 2	-1.058	promoter-TSS	NA
DCBLD2	discoidin, CUB and LCCL domain containing 2	-1.668	promoter-TSS	Differential expression with age in skin;[247] Suggested to have a suppressor role in some cancers[248]
DOLPP1	dolichyldiphosphatase 1	-1.076	promoter-TSS	
ELOVL5*	ELOVL fatty acid elongase 5	-1.313	promoter-TSS	
FAM117A*	family with sequence similarity 117 member A	-1.397	promoter-TSS	
FAM72A	family with sequence similarity 72 member A	-0.981	promoter-TSS	
FBXO11*	F-box protein 11	-1.198	promoter-TSS	

HSPA5	heat shock protein family A (Hsp70) member 5	-1.304	promoter-TSS	Aging leads to declining levels in mice causing ER stress; thought to be linked to age-related disease[229-232]
KRT74*	keratin 74	1.151	exon	
NEU1*	neuraminidase 1	-0.878	promoter-TSS	
P4HB	prolyl 4-hydroxylase subunit beta	-1.111	promoter-TSS	
SHC1	SHC adaptor protein 1	-1.505	promoter-TSS	Linked to longevity in mice[249]
TLE4*	transducin like enhancer of split 4	-0.998	promoter-TSS	
TMEM160*	transmembrane protein 160	-0.731	promoter-TSS	
TYW1B	tRNA-yW synthesizing protein 1 homolog B	-0.950	promoter-TSS	
ZDHHC2	zinc finger DHHC-type containing 2	1.094	promoter-TSS	Tumor suppressor;[250] Expression is associated with cancer metastasis[251]

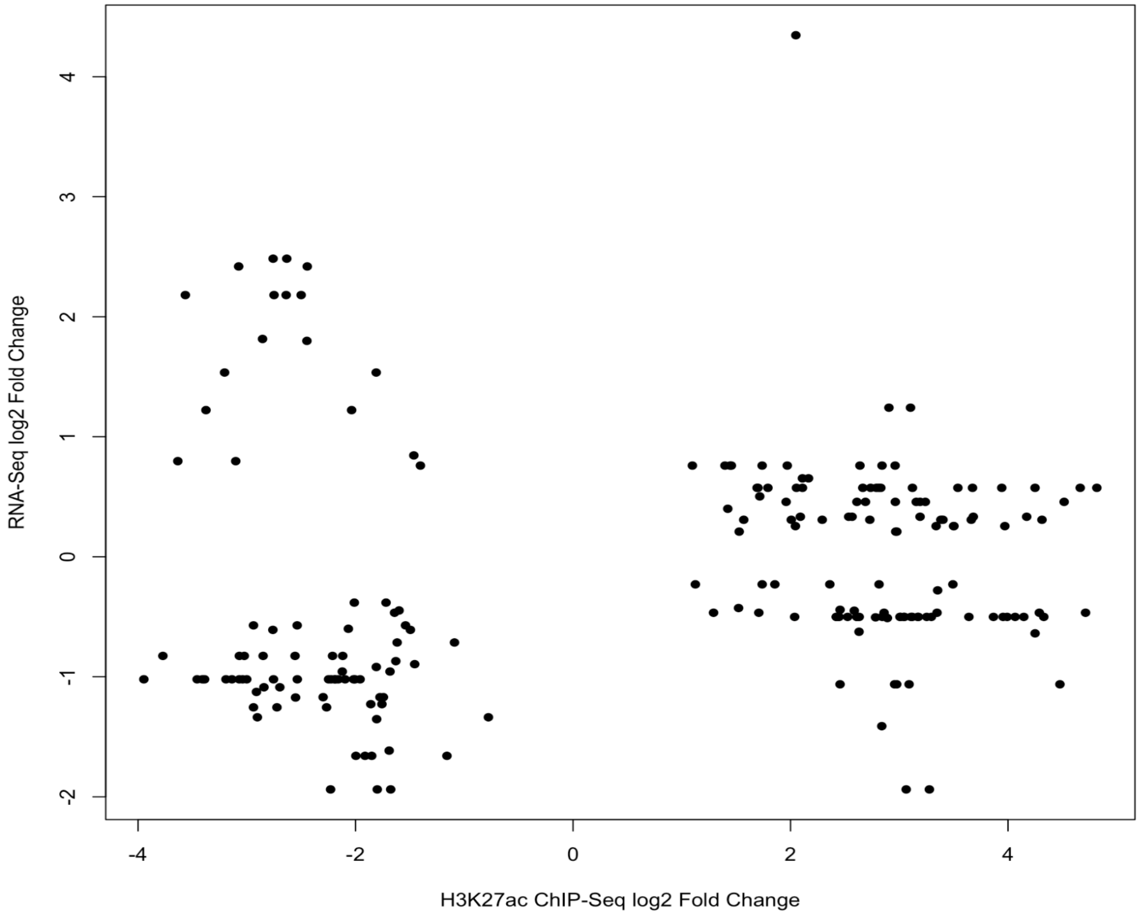
Gene names marked with an asterisk (*) are also found to show age-differential accessibility in blood by Ucar *et al*.

Table 2: GO Analysis of Biological Processes Associated with H3K27ac Peaks

Young			Old		
GO Term	**Biological Process**	**p-value**	**GO Term**	**Biological Process**	**p-value**
GO:0048731	system development	1.50E-86	GO:0048731	immune system process	3.12E-43
GO:0007275	multicellular organism development	8.28E-84	GO:0006955	immune response	1.19E-41
GO:0032501	multicellular organismal process	8.69E-83	GO:0044237	cellular metabolic process	2.17E-34
GO:0048856	anatomical structure development	1.78E-82	GO:0045321	leukocyte activation	1.26E-29
GO:0009653	anatomical structure morphogenesis	1.95E-81	GO:0002252	immune effector process	2.57E-29
GO:0007399	nervous system development	1.91E-79	GO:0044248	cellular catabolic process	9.05E-27
GO:0032502	developmental process	1.61E-73	GO:0002682	regulation of immune system process	1.74E-26
GO:0009887	animal organ morphogenesis	1.70E-67	GO:0001775	cell activation	2.63E-26
GO:0022008	neurogenesis	1.54E-64	GO:0002757	immune response-activating signal transduction	4.81E-26
GO:0048699	generation of neurons	6.21E-64	GO:0008152	metabolic process	2.83E-25

A GO analysis was performed on peaks annotated to their nearest gene. This was done for the 4,418 unique peaks with more H3K27 acetylation in Young group (on the left) and in the 7,183 unique peaks in the Old group (right). The top 10 sites for each of the two groups can be observed.

Supplementary Figure 1: Comparison of DEGs and Differential H3K27ac Regions



The relationship between the log2 fold change from the H3K27ac ChIP-Seq (x-axis) and the RNA-Seq (y-axis) are plotted. There is a positive correlation between the test statistics of genes linked to each of the assays (r=0.174, p=0.0129).

Supplementary Table 1: Sample information from ATAC-Seq

ID	Age	Age Group	Batch	# of total reads	% of reads mapping	% of unique reads	QC
NS0001	73	Old	2	35485108	86.45	89.85	Passed
NS0002	70	Old	2	30523168	84.97	91.99	Passed
NS0003	24	Young	2	21465194	83.18	88.39	Passed
NS0004	74	Old	2	25574713	88.64	93.05	Passed
NS0005	27	Young	1	33918017	73.42	82.98	Passed
NS0006	27	Young	1	35547266	84.92	80.39	Passed
NS0007	23	Young	2	50460486	81.41	85.1	Passed
NS0008	70	Old	2	37608809	80.92	81.34	Passed
NS0009	24	Young	2	40728137	84.98	67.79	Passed
NS0010	30	Young	2	39927602	77.94	80.47	Passed
NS0011	30	Young	1	36493122	82.89	85.21	Passed
NS0012	73	Old	1	34025313	81.88	81.77	Passed
NS0013	71	Old	1	36028394	81.78	82.24	Passed
NS0014	24	Young	1	32155280	80.04	81.27	Passed
NS0015	23	Young	1	38753173	78.67	85.09	Passed
NS0016	76	Old	1	39644091	80.77	79.53	Passed
NS0017	69	Old	1	51056153	78.81	58.82	Passed
NS0018	76	Old	1	46222752	83.88	56.69	Passed
NS0019	68	Old	1	32642193	87.64	56.26	Passed
NS0054	24	Young	1	33500794	75.32	64.21	Passed

Supplementary Table 2: Sample information from H3K27ac ChIP-Seq

ID	Age	Age Group	Batch	read length	# of total reads	% of reads mapping	% of unique reads	QC
NS0001	73	Old	4	51	20619199	94.38	73.2051	Passed
NS0002	70	Old	3	51	35410959	93.91	87.1	Passed
NS0003	24	Young	3	51	12038209	82.13	52.5	Passed
NS0004	74	Old	4					Failed
NS0005	27	Young	3	51	27566908	91.04	6.04	Failed
NS0006	27	Young	3	51	4775829	87.75	23.61	Failed
NS0006	27	Young	4	51	13458562	89.79	12.5577	Failed
NS0007	23	Young	1	151	38805698	98.16	75.61	Passed
NS0008	70	Old	3	51	4548117	92.15	10.59	Failed
NS0008	70	Old	4	51	2569339	92.56	14.15	Failed
NS0009	24	Young	1	151	30292529	94.19	45.07	Failed
NS0010	30	Young	3	51	24506026	91.74	6.58	Failed
NS0011	30	Young	3	51	996752	64.39	23.26	Failed
NS0011	30	Young	4	51	1151725	67.98	19.8561	Failed
NS0012	73	Old	1	151	34985622	98.27	59.83	Passed
NS0013	71	Old	1	151	35101421	95.44	85.11	Passed
NS0014	24	Young	3	51	18554584	94.06	49.18	Failed
NS0015	23	Young	2	51	34514616	97.58	67.97	Passed
NS0016	76	Old	2	51	31056944	97.9	87.18	Passed
NS0017	69	Old	2	51	29726115	98.69	89.85	Passed
NS0018	76	Old	2	51	30465602	98.78	90.1	Passed
NS0019	68	Old	2	51	28622019	98.48	88.85	Passed
NS0054	24	Young	2	51	23915846	98.48	88.89	Passed

Supplementary Table 3: Sample information from H3K27me3 ChIP-Seq

ID	Age	Age Group	Batch	read length	# of total reads	% of reads mapping	% of unique reads	QC
NS0001	73	Old	4	51	57677288	97.32	84.5583	Passed
NS0002	70	Old	3	51	26802741	94.29	80.82	Passed
NS0003	24	Young	3	51	514593	55.85	63.14	Failed
NS0003	24	Young	4	51	31516817	97.88	88.8968	Passed
NS0004	74	Old	4					Failed
NS0005	27	Young	4					Failed
NS0006	27	Young	4					Failed
NS0007	23	Young	1	151	20791778	89.05	52.93	Passed
NS0008	70	Old	3	51	28847930	86.85	4.36	Failed
NS0009	24	Young	1	151	16959523	96.13	81.8	Passed
NS0010	30	Young	3	51	2130215	91.4	19.73	Failed
NS0010	30	Young	4	51	1312928	90.74	23.9674	Failed
NS0011	30	Young	3	51	36440223	91.41	3.76	Failed
NS0012	73	Old	1	151	31614393	93.29	49.78	Passed
NS0013	71	Old	1	151	7716320	21.6	92.98	Failed
NS0014	24	Young	3	51	21725828	91.84	7.01	Failed
NS0015	23	Young	2	51	26088736	98.24	82.32	Passed
NS0016	76	Old	2	51	25172405	96.03	67.33	Passed
NS0017	69	Old	2	51	32536797	96.39	89.4	Passed
NS0018	76	Old	2	51	30975417	98.61	89.84	Passed
NS0019	68	Old	2	51	31915403	98.2	84.75	Passed
NS0054	24	Young	2	51	27845713	98.01	80.47	Passed

Supplementary Table 4: Sample information from H3K4me3 ChIP-Seq

ID	Age	Age Group	Batch	read length	# of total reads	% of reads mapping	% of unique reads	QC
NS0001	73	Old	4	51	30696913	95.89	66.5361	Passed
NS0002	70	Old	3	51	31828166	91.9	56.53	Passed
NS0003	24	Young	3	51	48254065	97.44	36.92	Failed
NS0004	74	Old	3	51	26373865	94.12	58.45	Passed
NS0005	27	Young	4					Failed
NS0006	27	Young	3	51	26702859	93.17	5.21	Failed
NS0007	23	Young	1	151	49872915	97.87	64.36	Passed
NS0008	70	Old	3	51	45265509	98.75	60.16	Passed
NS0009	24	Young	1	151	36108251	97.41	58.6	Passed
NS0010	30	Young	3	51	39914720	97.23	68.72	Passed
NS0011	30	Young	3	51	38801691	98.14	81.09	Passed
NS0012	73	Old	1	151	30216046	98.17	86.28	Passed
NS0013	71	Old	1	151	28048815	94.26	65.31	Passed
NS0014	24	Young	3	51	21429947	93.32	8.89	Failed
NS0015	23	Young	2	51	25715387	98.26	83.48	Passed
NS0016	76	Old	2	51	34965552	98.09	71.31	Passed
NS0017	69	Old	2	51	35803126	98.17	82.66	Passed
NS0018	76	Old	2	51	27985935	98.03	83.63	Passed
NS0019	68	Old	2	51	54566585	92.72	80.5	Passed
NS0054	24	Young	2	51	38026067	98.51	59.37	Passed

Supplementary Table 5: Results from genes annotated to both H3K27ac ChIP-Seq and DEGs

ChIP-Seq peak	log2FoldChange_CHIP	RNA-Seq region	log2FoldChange_RNA	Gene Name	Gene Description
chr2:42496858-42498331	-2.0659492	chr2:42494569-42756947	-0.6002609	MTA3	metastasis associated 1 family member 3
chr1:234499988-234500402	2.89048395	chr1:234391313-234479103	-0.5104962	TARBP1	TAR (HIV-1) RNA binding protein 1
chr2:219434990-219435338	-3.0690829	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr2:219441913-219442300	-2.0177913	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr2:219496156-219496746	-3.136465	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219441693-219441913	-2.7542898	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr2:219493031-219493505	-2.1938028	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219435571-219436076	-2.1582653	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr2:219483647-219483993	-3.3878274	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219451903-219452711	-2.2239712	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr2:219496848-219497096	-2.2479082	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219483400-219483647	-3.4081299	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219448765-219449084	-3.9466279	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr2:219443289-219443978	-2.0055026	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr2:219490838-219492544	-1.9580753	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219484900-219485211	-3.0382794	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219483993-219484272	-2.5355782	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219495450-219495907	-2.1840844	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219494591-219495450	-2.0959153	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219484463-219484706	-3.4567466	chr2:219434846-219498287	-1.0211249	SPEG	CAVP-target protein-like
chr2:219448362-219448704	-3.1913029	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr2:219445176-219446086	-3.0006451	chr2:219434846-219498287	-1.0211249	SPEG	SPEG complex locus
chr16:180401-180725	-3.2041283	chr16:180453-181181	1.53510878	HBQ1	hemoglobin subunit theta 1

chr16:180725-181637	-1.8099627	chr16:180453-181181	1.53510878	HBQ1	hemoglobin subunit theta 1
chr14:30620776-30622899	1.52803495	chr14:30622112-30735812	0.20939892	SCFD1	sec1 family domain containing 1
chr14:30691531-30692753	2.97749024	chr14:30622112-30735812	0.20939892	SCFD1	sec1 family domain containing 1
chr14:30632735-30633653	2.96745401	chr14:30622112-30735812	0.20939892	SCFD1	sec1 family domain containing 1
chr12:54276048-54277002	3.35285529	chr12:54230940-54280133	-0.279898	CBX5	chromobox 5
chr10:114499019-114499598	4.7128821	chr10:114431113-114685003	-0.4663786	ABLIM1	actin binding LIM protein 1
chr10:114540753-114541598	1.70848242	chr10:114431113-114685003	-0.4663786	ABLIM1	actin binding LIM protein 1
chr10:114526216-114527946	1.29203414	chr10:114431113-114685003	-0.4663786	ABLIM1	actin binding LIM protein 1
chr10:114500753-114501462	3.34719292	chr10:114431113-114685003	-0.4663786	ABLIM1	actin binding LIM protein 1
chr10:114494612-114495719	2.8601631	chr10:114431113-114685003	-0.4663786	ABLIM1	actin binding LIM protein 1
chr10:114498787-114499019	4.28776742	chr10:114431113-114685003	-0.4663786	ABLIM1	actin binding LIM protein 1
chr10:114631655-114632344	-1.641289	chr10:114431113-114685003	-0.4663786	ABLIM1	actin binding LIM protein 1
chr14:24643720-24644131	-1.4034763	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24681172-24682125	2.96171651	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24694961-24696013	1.44455383	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24678335-24679549	1.39692987	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24693936-24694849	1.96995212	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24680084-24680891	1.45610968	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24673378-24674454	1.73899232	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24673231-24673378	2.84143239	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24631259-24633957	1.09744061	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr14:24679749-24680084	2.63834051	chr14:24630954-24634267	0.75969397	GZMB	granzyme B
chr16:68282178-68285510	1.52222416	chr16:68264516-68301823	-0.4272165	SLC7A6	solute carrier family 7 member 6
chr7:72937841-72938227	4.24885684	chr7:72947581-72954790	-0.6382985	NSUN5P2	NOP2/Sun RNA methyltransferase family member 5 pseudogene 2

chr17:2028882-2029702	-1.5995383	chr17:2030110-2043430	-0.4476532	DPH1	diphthamide biosynthesis 1
chr4:73844771-73845665	3.10340386	chr4:73853189-73854155	1.24200011	PF4V1	platelet factor 4 variant 1
chr4:73848123-73848665	2.90495647	chr4:73853189-73854155	1.24200011	PF4V1	platelet factor 4 variant 1
chr12:50059844-50060195	-1.9128452	chr12:50057548-50083611	-1.6589348	ASIC1	acid sensing ion channel subunit 1
chr12:50059244-50059844	-1.9967704	chr12:50057548-50083611	-1.6589348	ASIC1	acid sensing ion channel subunit 1
chr12:50058032-50059218	-1.8512003	chr12:50057548-50083611	-1.6589348	ASIC1	acid sensing ion channel subunit 1
chr12:50050133-50051650	-1.1595527	chr12:50057548-50083611	-1.6589348	ASIC1	acid sensing ion channel subunit 1
chr6:35370130-35370911	1.12499712	chr6:35342558-35428191	-0.2299371	PPARD	peroxisome proliferator activated receptor delta
chr6:35361957-35362536	2.81413339	chr6:35342558-35428191	-0.2299371	PPARD	peroxisome proliferator activated receptor delta
chr6:35368732-35369157	3.49294484	chr6:35342558-35428191	-0.2299371	PPARD	peroxisome proliferator activated receptor delta
chr6:35370911-35371726	1.73929088	chr6:35342558-35428191	-0.2299371	PPARD	peroxisome proliferator activated receptor delta
chr6:35369157-35369956	1.85578635	chr6:35342558-35428191	-0.2299371	PPARD	peroxisome proliferator activated receptor delta
chr6:43076194-43076673	-2.1219302	chr6:43076268-43161719	-0.9564751	PTK7	protein tyrosine kinase 7 (inactive)
chr6:43075774-43076194	-1.6829499	chr6:43076268-43161719	-0.9564751	PTK7	protein tyrosine kinase 7 (inactive)
chr2:25483218-25483990	2.58620358	chr2:25227855-25342590	-0.4495961	DNMT3A	DNA methyltransferase 3 alpha
chr10:97766733-97767373	-1.6907552	chr10:97766751-97771952	-1.6153515	SFRP5	secreted frizzled related protein 5
chr3:113740545-113741456	2.0450597	chr3:113716460-113746300	0.25540833	NAA50	N(alpha)-acetyltransferase 50, NatE catalytic subunit
chr3:113742954-113744039	3.97084939	chr3:113716460-113746300	0.25540833	NAA50	N(alpha)-acetyltransferase 50, NatE catalytic subunit
chr3:113736965-113737480	3.50261704	chr3:113716460-113746300	0.25540833	NAA50	N(alpha)-acetyltransferase 50, NatE catalytic subunit
chr3:113744496-113745331	3.49965536	chr3:113716460-113746300	0.25540833	NAA50	N(alpha)-acetyltransferase 50, NatE catalytic subunit
chr3:113741535-113742306	3.33872486	chr3:113716460-113746300	0.25540833	NAA50	N(alpha)-acetyltransferase 50, NatE catalytic subunit
chr9:133611128-133611961	-2.5516226	chr9:133636360-133659344	-1.1727947	DBH	dopamine beta-hydroxylase
chr12:68612008-68612973	3.19114595	chr12:68610839-68671901	0.33328492	RAP1B	RAP1B, member of RAS oncogene family

chr12:68613797-68615385	2.53499624	chr12:68610839-68671901	0.33328492	RAP1B	RAP1B, member of RAS oncogene family
chr12:68611865-68612008	4.17158332	chr12:68610839-68671901	0.33328492	RAP1B	RAP1B, member of RAS oncogene family
chr12:68591982-68592197	3.6806602	chr12:68610839-68671901	0.33328492	RAP1B	RAP1B, member of RAS oncogene family
chr12:68591565-68591982	2.56597605	chr12:68610839-68671901	0.33328492	RAP1B	RAP1B, member of RAS oncogene family
chr12:68615385-68615947	2.09035705	chr12:68610839-68671901	0.33328492	RAP1B	RAP1B, member of RAS oncogene family
chr22:23073601-23074081	-1.4640747	chr22:23070361-23125037	0.84447425	GNAZ	G protein subunit alpha z
chr8:17800650-17800975	-2.6961684	chr8:17643795-17800917	-1.087405	MTUS1	microtubule associated tumor suppressor 1
chr8:17800975-17801719	-2.842286	chr8:17643795-17800917	-1.087405	MTUS1	microtubule associated tumor suppressor 1
chr3:15045656-15046236	2.36133037	chr3:15042460-15065335	-0.2297381	MRPS25	mitochondrial ribosomal protein S25
chr6:85449734-85450685	-1.6295576	chr6:85449584-85495791	-0.8696894	NT5E	5'-nucleotidase ecto
chr2:130837462-130837733	-3.7715943	chr2:130836916-131047263	-0.8260299	ARHGEF4	Rho guanine nucleotide exchange factor 4
chr2:130914673-130915247	-2.1177408	chr2:130836916-131047263	-0.8260299	ARHGEF4	Rho guanine nucleotide exchange factor 4
chr2:130915247-130916079	-2.5562655	chr2:130836916-131047263	-0.8260299	ARHGEF4	Rho guanine nucleotide exchange factor 4
chr2:130964649-130965564	-2.2115959	chr2:130836916-131047263	-0.8260299	ARHGEF4	Rho guanine nucleotide exchange factor 4
chr2:130836670-130837010	-3.0231721	chr2:130836916-131047263	-0.8260299	ARHGEF4	Rho guanine nucleotide exchange factor 4
chr2:130837010-130837462	-3.0682417	chr2:130836916-131047263	-0.8260299	ARHGEF4	Rho guanine nucleotide exchange factor 4
chr2:130963318-130964368	-2.8497399	chr2:130836916-131047263	-0.8260299	ARHGEF4	Rho guanine nucleotide exchange factor 4
chr13:75481318-75482672	-1.4954184	chr13:75284665-75482114	-0.60917	TBC1D4	TBC1 domain family member 4
chr13:75480938-75481318	-2.7603043	chr13:75284665-75482114	-0.60917	TBC1D4	TBC1 domain family member 4
chr1:225836649-225837201	2.45543027	chr1:225810092-225845563	-0.4419496	EPHX1	epoxide hydrolase 1
chr5:137695043-137695521	2.63003535	chr5:137617500-137736090	-0.6248682	KLHL3	kelch like family member 3
chr7:45973018-45973981	-2.0360618	chr7:45912245-45921874	1.22170459	IGFBP3	insulin like growth factor binding protein 3
chr7:45920659-45921407	-3.3753106	chr7:45912245-45921874	1.22170459	IGFBP3	insulin like growth factor binding protein 3
chr9:121365899-121366676	3.67346585	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121352766-121354500	1.79211835	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121363749-121364742	2.78920554	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121367630-121368618	3.94239526	chr9:121338988-121370304	0.57428455	STOM	stomatin

chr9:121350173-121351310	2.83022659	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121359282-121359657	4.81677049	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121395466-121395864	2.7365264	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121355668-121356912	2.05464045	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121360629-121361434	2.81129997	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121359657-121360003	4.24858182	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121344979-121345672	3.1210978	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121354501-121355489	1.69264911	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121357919-121359255	1.70059752	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121365104-121365370	4.66374831	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121397878-121398569	2.10982566	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121399452-121400179	2.66468118	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr9:121363046-121363506	3.53886332	chr9:121338988-121370304	0.57428455	STOM	stomatin
chr10:126388017-126388571	-2.9112892	chr10:126012381-126388455	-1.1260976	ADAM12	ADAM metallopeptidase domain 12
chr4:22515466-22516247	-1.8090522	chr4:22345071-22516054	-0.9186261	ADGRA3	adhesion G protein-coupled receptor A3
chr3:133895343-133895790	-3.1017505	chr3:133824239-133895836	0.79673236	RAB6B	RAB6B, member RAS oncogene family
chr3:133895790-133896076	-3.6342333	chr3:133824239-133895836	0.79673236	RAB6B	RAB6B, member RAS oncogene family
chr1:64469868-64471230	-1.6768561	chr1:64470792-64693058	-1.9386234	CACHD1	cache domain containing 1
chr1:64472504-64473826	-2.2295504	chr1:64470792-64693058	-1.9386234	CACHD1	cache domain containing 1
chr1:64513370-64513696	3.27680276	chr1:64470792-64693058	-1.9386234	CACHD1	cache domain containing 1
chr1:64511590-64512009	3.06362215	chr1:64470792-64693058	-1.9386234	CACHD1	cache domain containing 1
chr1:64471230-64471515	-1.8002106	chr1:64470792-64693058	-1.9386234	CACHD1	cache domain containing 1
chr1:1359276-1360220	-1.5403082	chr1:1352689-1361777	-0.571779	MXRA8	matrix remodeling associated 8
chr1:1354121-1355661	-2.9383052	chr1:1352689-1361777	-0.571779	MXRA8	matrix remodeling associated 8
chr1:1357371-1357880	-2.536794	chr1:1352689-1361777	-0.571779	MXRA8	matrix remodeling associated 8
chr1:247347931-247348274	-1.4560862	chr1:247297412-247331846	-0.8950014	ZNF496	zinc finger protein 496

chr3:13550436-13551344	-1.775146	chr3:13549131-13638422	-1.1699153	FBLN2	fibulin 2
chr3:13545986-13546584	-1.7440301	chr3:13549131-13638422	-1.1699153	FBLN2	fibulin 2
chr3:13548813-13549821	-2.297947	chr3:13549131-13638422	-1.1699153	FBLN2	fibulin 2
chr1:155126597-155130604	-0.7781952	chr1:155127460-155134857	-1.3373672	EFNA1	ephrin A1
chr1:155125891-155126584	-2.9026586	chr1:155127460-155134857	-1.3373672	EFNA1	ephrin A1
chr16:31538081-31539381	-2.7580402	chr16:31527864-31528803	2.48352136	AHSP	alpha hemoglobin stabilizing protein
chr16:31537445-31537759	-2.6323905	chr16:31527864-31528803	2.48352136	AHSP	alpha hemoglobin stabilizing protein
chr2:100820259-100820737	-2.9382575	chr2:100820152-100996829	-1.254777	NPAS2	neuronal PAS domain protein 2
chr2:100818841-100819363	-2.2663392	chr2:100820152-100996829	-1.254777	NPAS2	neuronal PAS domain protein 2
chr2:100819363-100820161	-2.7228571	chr2:100820152-100996829	-1.254777	NPAS2	neuronal PAS domain protein 2
chr20:31723990-31724840	1.71694738	chr20:31664452-31723989	0.50293874	BCL2L1	BCL2 like 1
chr7:134645881-134648252	1.56964879	chr7:134646808-134679813	0.30798484	BPGM	bisphosphoglycerate mutase
chr7:134677375-134678364	3.38223135	chr7:134646808-134679813	0.30798484	BPGM	bisphosphoglycerate mutase
chr7:134641735-134642065	3.66204242	chr7:134646808-134679813	0.30798484	BPGM	bisphosphoglycerate mutase
chr7:134670423-134671412	4.31228407	chr7:134646808-134679813	0.30798484	BPGM	bisphosphoglycerate mutase
chr7:134639999-134640703	2.72907898	chr7:134646808-134679813	0.30798484	BPGM	bisphosphoglycerate mutase
chr7:134669622-134670165	2.00700203	chr7:134646808-134679813	0.30798484	BPGM	bisphosphoglycerate mutase
chr7:134635830-134636445	2.2914387	chr7:134646808-134679813	0.30798484	BPGM	bisphosphoglycerate mutase
chr7:134641155-134641735	3.40185187	chr7:134646808-134679813	0.30798484	BPGM	bisphosphoglycerate mutase
chr7:111026758-111027318	2.83922563	chr7:111091006-111125454	-1.4113622	LRRN3	leucine rich repeat neuronal 3
chr5:179549151-179549951	1.42294097	chr5:179550558-179610026	0.39982442	RUFY1	RUN and FYVE domain containing 1
chr3:18679089-18679533	3.29512077	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18760285-18761396	3.10772718	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18677555-18678752	3.04714562	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18402006-18402707	3.11969449	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18738621-18739093	3.86621952	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18739093-18739952	3.17592507	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1

chr3:18725361-18725953	2.44660842	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18439114-18439567	4.14507519	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18727913-18728298	3.64017104	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18437341-18438134	3.25216461	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18747073-18748002	2.42042872	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18429662-18430931	2.43603791	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18744073-18744691	2.03792288	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18727648-18727913	2.52493021	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18593087-18594010	3.95460939	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18723696-18725361	2.6112733	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18739952-18740380	3.00627704	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18419540-18420482	4.06516312	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18744695-18745420	3.99234298	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18740917-18741383	2.42321123	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18726036-18726728	2.84292565	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18741815-18742428	3.04198302	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18745902-18746726	3.16940276	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18740539-18740917	2.60900762	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18743377-18743792	3.01100861	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18403048-18404762	2.63145232	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr3:18741383-18741815	4.32893241	chr3:18345387-18445588	-0.500976	SATB1	SATB homeobox 1
chr12:55824289-55824838	2.78200853	chr12:55829745-55836246	-0.504748	TMEM198B	transmembrane protein 198B (pseudogene)
chr3:42906071-42906765	-1.7572802	chr3:42905731-42917641	-1.2282273	ZNF662	zinc finger protein 662
chr3:42906819-42907383	-1.8598279	chr3:42905731-42917641	-1.2282273	ZNF662	zinc finger protein 662
chr8:4992052-4992631	-3.0748089	chr8:2935353-4994972	2.4187487	CSMD1	CUB and Sushi multiple domains 1
chr8:4993793-4994507	-2.4440007	chr8:2935353-4994972	2.4187487	CSMD1	CUB and Sushi multiple domains 1
chr20:43507186-43507472	-1.7195335	chr20:43507680-43550950	-0.3822358	L3MBTL1	l(3)mbt-like 1 (Drosophila)

chr20:43507472-43508054	-2.0111582	chr20:43507680-43550950	-0.3822358	L3MBTL1	l(3)mbt-like 1 (Drosophila)
chr16:170547-171477	-2.8550701	chr16:172847-173710	1.81467013	HBA2	hemoglobin subunit alpha 2
chr22:43997272-43997912	2.10958344	chr22:43999211-44172949	0.65361968	PARVB	parvin beta
chr22:44090975-44091875	2.16546814	chr22:43999211-44172949	0.65361968	PARVB	parvin beta
chr7:100170276-100171992	-1.0901479	chr7:100159244-100168750	-0.714478	GAL3ST4	galactose-3-O-sulfotransferase 4
chr7:100164376-100165473	-1.617404	chr7:100159244-100168750	-0.714478	GAL3ST4	galactose-3-O-sulfotransferase 4
chr17:47204375-47204861	2.04900177	chr17:47200446-47223679	4.34518242	MYL4	myosin light chain 4
chr16:177401-178174	-2.4472986	chr16:176680-177522	1.79859024	HBA1	hemoglobin subunit alpha 1
chr16:165322-165819	-2.6381196	chr16:153892-166768	2.18056074	HBM	hemoglobin subunit mu
chr16:165819-166207	-2.7495031	chr16:153892-166768	2.18056074	HBM	hemoglobin subunit mu
chr16:168504-168837	-2.5001118	chr16:153892-166768	2.18056074	HBM	hemoglobin subunit mu
chr16:166857-167979	-3.5649863	chr16:153892-166768	2.18056074	HBM	hemoglobin subunit mu
chr2:28756812-28757514	3.2396292	chr2:28751640-28802940	0.45716745	PPP1CB	protein phosphatase 1 catalytic subunit beta
chr2:28675819-28676433	2.96378784	chr2:28751640-28802940	0.45716745	PPP1CB	protein phosphatase 1 catalytic subunit beta
chr2:28749569-28750546	2.68906242	chr2:28751640-28802940	0.45716745	PPP1CB	protein phosphatase 1 catalytic subunit beta
chr2:28748484-28749078	4.51618545	chr2:28751640-28802940	0.45716745	PPP1CB	protein phosphatase 1 catalytic subunit beta
chr2:28716161-28716988	3.19142488	chr2:28751640-28802940	0.45716745	PPP1CB	protein phosphatase 1 catalytic subunit beta
chr2:28712731-28713392	1.95948505	chr2:28751640-28802940	0.45716745	PPP1CB	protein phosphatase 1 catalytic subunit beta
chr2:28771316-28772071	3.15616784	chr2:28751640-28802940	0.45716745	PPP1CB	protein phosphatase 1 catalytic subunit beta
chr2:28716988-28717421	2.61049018	chr2:28751640-28802940	0.45716745	PPP1CB	protein phosphatase 1 catalytic subunit beta
chr3:73038019-73038254	2.97634455	chr3:73061659-73063337	-1.062678	EBLN2	endogenous Bornavirus-like nucleoprotein 2
chr3:73040854-73042381	2.95713225	chr3:73061659-73063337	-1.062678	EBLN2	endogenous Bornavirus-like nucleoprotein 2
chr3:73029272-73030712	3.09026767	chr3:73061659-73063337	-1.062678	EBLN2	endogenous Bornavirus-like nucleoprotein 2
chr3:73051960-73052999	4.47731	chr3:73061659-73063337	-1.062678	EBLN2	endogenous Bornavirus-like nucleoprotein 2
chr3:73037458-73037980	2.45535371	chr3:73061659-73063337	-1.062678	EBLN2	endogenous Bornavirus-like nucleoprotein 2
chr17:35795102-35796076	-1.8057306	chr17:35756249-35795707	-1.3531576	MMP28	matrix metallopeptidase 28

Chapter V: Discussion

Epigenetic Biomarkers of Aging Show Promise

This work has contributed to the base of scientific knowledge by further exploring the utility of epigenetic Biomarkers of Aging (BoA) both by characterizing existing markers and by exploring epigenetic modifications and age-related alterations that could be used to develop novel markers. In Chapter II, an existing DNAm biomarker is examined in longitudinal DNAm data, noting the biomarker's stability over a 16-year study, as well as its contribution to modeling BMI and possibly fasting glucose levels. In Chapter III, a novel biomarker based on the degree to which DNAm shows variation across individuals is developed; however, this simple approach to developing a biomarker does not correlate with age-related phenotypes or mortality. Lastly, in Chapter IV, changes in chromatin accessibility and histone modifications with age are characterized, shedding light on which epigenetic modifications are informative in modeling age in biomarkers.

A recent review of BoA characterized the 6 current types of biomarkers, which include: epigenetic clocks, telomere length, transcriptomic predictors, proteomic predictors, metabolomics-based predictors, and composite biomarkers. This review concludes that the most promising current biomarkers are the epigenetic clocks[109]—motivating the importance of their study – though this review also suggests that the most useful future biomarkers will likely be more similar to the composite biomarkers that combine information across the six types.

It is known that epigenetic modifications interact but contribute overlapping information.[142] An example of this is present in the structure of heterochromatin, which is marked by condensed chromatin, the absence of histone modifications associated with active transcription, an enrichment of histone modifications associated with repressed transcription, as well as the presence of DNAm.[147] While assaying just one of these mechanisms doesn't allow for precise prediction of whether a region is heterochromatic, assaying all mechanisms leads to diminishing returns. Thus, the epigenetic modifications most informative in aging should be characterized in a model that is both accurate and parsimonious. Many epigenetic aging clocks are informed by patterns in DNAm, but do not take into account other informative levels of the epigenome; this information may be able to inform age-related changes and may benefit future aging clocks.

In addition to which levels of the epigenome are assayed in aging clocks, the types of changes that the epigenome undergoes can reflect different biological phenomenon, with age-related changes being linear, non-linear, or occurring randomly, reflecting epigenetic drift. Understanding the driving forces behind the different types of changes occurring, whether they are a result of programmed changes with age or a result of drift, will be vital to the improvement of models of aging.

Considerations for Populations Included in Future Epigenetics Studies

In the United States, health disparities disproportionately affect historically disadvantaged groups of racial and ethnic minorities, leading to inequities in rates of morbidity and mortality,[252] however, in many scientific studies, these groups are underrepresented.[253] Black Americans in particular have consistently worse health outcomes when compared to white Americans—despite recent improvements in access to health care.[254] Part of the driving force for this difference may lay in epigenetic differences between these groups driven both by underlying genetic differences as well as differing interactions with the environment. These differing interactions leading to epigenetic changes are driven by the effects of: stress, racism, diet, socioeconomic status (SES), and residing in areas of the US disproportionately plagued by harmful environmental exposures.[252]

A recent study used the 450K array to assay DNAm in B-lymphocytes among white, black, and Han-Chinese Americans and found that methylation at just 439 sites allowed them to accurately separate participants by population.[255] In addition to these sites providing insight into which population the participant came from, they were also associated with distinct phenotypic characteristics, drug metabolism, response to external stimuli, and, most relevant to this work, disease susceptibility. Two thirds of these sites were driven by the underlying genetic background, leaving one third of these sites reflecting DNAm differences that do not appear to be driven by genetic variance yet remain epigenetic markers indicating population of origin.

Applying DNAm aging clocks to populations of black, white, and Hispanic Americans, it was revealed that epigenetic aging rates also appear linked to the racial and ethnic populations to which a participant belongs; even finding that these different populations have different mortality rates after accounting for their differing SES.[121] Of note, the study found that Hispanic Americans, despite their disadvantaged status in the U.S., actually have lower levels of age acceleration when compared to white Americans. This finding from an aging clock echoes health data that also shows Hispanic Americans have a lower overall risk of mortality compared to white Americans—a phenomenon known as the 'Hispanic mortality paradox.'[256] This study also identified differences between black and white Americans in proportions of CD8+ T cells, and between Hispanic and white Americans in CD4+ T cells. Not accounting for these age-dependent proportions is a known confounder in aging studies conducted in blood because different cell types in blood feature different DNAm profiles. The finding that these differences exist across racial lines reaffirms the importance of using models that allow for estimation of blood cell proportions that are either reference free or matched to the population to which they will be applied. Additionally, because some aging clocks are influenced by cell proportions in blood, not accounting for racial differences in cell proportions in blood could lead different clocks to reach different accuracy in different populations.

These findings, in addition to the discovery that women consistently reflect lower DNAm-based age acceleration despite higher rates of morbidity,[121] and that racial differences exist in telomere lengths,[257] complicate the notion that a single, sex and race/ethnicity agnostic BoA can be developed. In order to develop models that are both accurate and applicable to a wide range of Americans, more underrepresented populations must be included in epigenetic studies. Yet, many genetic studies (genome-wide association studies, GWASs) focus on and feature majority a European population. A 2009 analysis found that 96% of participants in GWASs were of European descent, despite this population comprising only 16% of the global population; an updated 2016 analysis found this number decreased—but only to 80%.[253]

This lack of diversity among those included in genetics and epigenetics research has been well documented and lamented in several publications, with many researchers touting the many clinical benefits and scientific discoveries that more inclusion among the participants in genetics studies would engender.[253,258-261] Arguably, this lack of diversity is already worsening health outcomes for minorities. A study found that, because the algorithm used to determine the appropriate prescription of an anticoagulant was trained on a homogenous, European population it did not account for genotypes more common in black Americans, leading many black Americans to be prescribed the incorrect dosage.[262] Moreover, a recent study of BoA developed using machine learning methods (specifically, deep neural networks) trained on 41 clinical biomarkers found the resulting model to be highly accurate in predicting age.[263] However, Cohen *et al*.[264] found this model to be highly population dependent, not performing nearly as well as it did in the homogenous population in which it was developed (which was

90% Eastern European). This suggests that using newer, and more complex deep learning techniques to develop algorithms to predict age and age-related phenotypes may create models that are accurate but lack generalizability to wider, more diverse populations.

In addition to considerations into the composition of participants in future studies of epigenetics and aging, the study design must also be taken into account. While some aging studies rely on cross-sectional data, in which subjects of different ages are compared, others leverage longitudinal data's ability to allow for the observation of age-related changes in one individual as they occur. Both types of studies are associated with potential sources of bias,[265] though they can be less expensive and carried out faster, cross-sectional studies can be affected by selection and cohort bias, in which differences are due to specific members of a cohort and factors specific to the time at which they lived. Conversely, longer and more expensive longitudinal studies are susceptible to survivor bias or loss-to-follow-up, where participants leave the study in a non-random way or those informative to the phenotype of interest do not survive long enough to be studied.

Though the two types of studies offer their own advantages and drawbacks, some researchers have found that, by analyzing a subset their longitudinal data as cross-sectional, they are not able to detect significant associations using cross-sectional data that are evident when the data is analyzed longitudinally.[266,267] This suggests that, in order to best capture changes in the trajectory of aging, longitudinal studies must be utilized, and would ideally follow participants through different stages of aging, not just in old age after survivor bias may already have been at work. Additionally, the very real effect of cohort biases suggest that all participants, ideally, be the same age at that start of the study.[267] Studies applying epigenetic clocks to babies at birth[268] or to teens[269] propose that aging trajectories are already being established early in life. This suggests that studies of aging would improve by incorporating epigenetic information from participants in early life when their aging trajectories are first being established.

Future Directions for Biomarkers of Aging

Biomarkers of the future will likely leverage different information than those currently in use, incorporating changes across the different hallmarks of aging[13] as well as phenotypic and clinical data to create more comprehensive indicators of health.[17] The original DNAm clocks included little to no clinical data, while the newly developed clocks incorporate several clinical measures.[123] Additionally, many of these clocks depend on DNAm data from the 27K and 450K DNAm arrays, which feature CpGs enriched at promotors and CpG islands.[270] The newly developed 850K EPIC array assays many of the same sites as well as additional sites in regions identified as enhancers by the ENCODE project.[271] The results of Chapter IV suggest widespread age-related changes at regions marked by a histone modification common to enhancers, suggesting that these sites may be involved in phenotypes accompanying the aging process. The development of the CRISPR-Cas9 system[272,273] offers promise for the editing of both the genome and the epigenome. By creating fusion proteins between the CRISPR-Cas9 complex and histone modifiers (like histone acetyltransferases, histone acetyltransferase inhibitors, etc), the Cas9 can guide the histone modifier to specific genomic locations, thereby directly targeting the actions of enhancers found to play a role in aging and disease.[274,275]

Additionally, many DNAm clocks are informed by individual CpGs across the genome, each with very little predictive ability of an individual's age. DNAm, however, can also be characterized at the regional level by identifying differentially methylated regions (DMRs). It has been suggested that these DMRs are more functionally important than methylation changes at individual CpGs in informing gene expression.[133] Assaying DNAm changes at the regional level can uncover previously uncharacterized CpGs associated with age and age-related disease, allowing for the discovery of informative regions outside of those assayed by arrays.[276]

As useful as these DNAm arrays are, they have their own biases and limitations, with the biggest being the limitation to assaying a set number of sites in the genome that are usually in close proximity to each other.[277] In addition, the integration of array-based and capture-based epigenetic assays can prove difficult, impeding the understanding of the interplay between sequence-level changes with broader chromatin and histone modification changes.[278] Because the cost of sequencing the genome is decreasing, future epigenetic studies may increasingly rely on whole genome sequencing techniques, favoring whole genome bisulfite sequencing for capturing DNAm patterns over array-based methods.

In addition to being able to assay more CpGs across the genome, sequencing techniques will allow for the distinct types of modifications to DNA besides 5-methylcytosine (5-mC) to be assayed. Employing such a sequencing technique,[279] genome-wide patterns of 5-hydroymethylcytosine (5-hmC) were recently assayed in whole blood among the Old and Young subjects also assayed in Chapter IV.[225] Johnson *et al.* found thousands of sites undergoing age-related changes in DNA hydroxymethylation that were also linked to gene expression, suggesting this mark too may be useful in future aging studies in blood.

Future markers, in addition to relying on different assay techniques of the epigenome, may also rely on different types of epigenetic changes observed in age. In Chapter III, CpGs reflecting changes in their variability with age in a cross-sectional cohort are identified, with the assumption that these regions reflect stochastic processes and a lack of maintenance at the level of the epigenome.[67] At the genomic level, this also occurs in the form of DNA damage and repair. The efficiency with which these repairs occur varies at genomic locations based on their transcriptional activity, this leads some regions to maintain their integrity while others accumulate mutations.[280] In theory, the same could be true of the epigenome—that some regions are protected from drift because of their functional importance and genomic locations. A better understanding of the mechanisms that promote epigenomic integrity could lead to interventions that maintain that integrity epigenome-wide. The development of recent techniques in single-cell profiling techniques may be vital to understanding this variability between cells, informing which mechanisms drive this variability on a molecular scale.[281,282] Attenuating epigenetic dysregulation could have protective effects against the aberrant changes in chromatin configuration and gene expression with age, and thus, a protective effect against age-related disease development.

Conclusions

Cellular changes seen in aging are similar to those observed in disease, suggesting that aging and disease share a common pathway.[4] Because it is established by enzymes, and thus reversible, the epigenome is a promising target for therapeutic interventions and personalized medicine.[80] Of interest, these concepts of interventions were recently applied in model organism studies, showing that interventions known to prolong lifespan in mice actually slow age-related changes in DNA methylation.[283] This finding supports the hypothesis that a better understanding of the contributors to biological aging may allow for interventions that can delay the onset and progression of age-related disease, ultimately uncoupling the normal process of aging from the development age-related disease.

With an increasing proportion of the population surviving to old age, it is imperative that biomarkers are developed that can accurately predict not just lifespan, but healthspan. These biomarkers will form the basis of accurate personalized medicine and be integrated into geriatric care. These biomarkers will benefit from a development that takes into account the makeup of the participants included as well as the structure of sample collection. Longitudinal studies of aging are advantageous over cross-sectional studies because the allow for the characterization of aging in an individual as it occurs. Lastly, more diversity in genetics and epigenetics studies will be vital to ensure that the advances made from the aging biomarkers developed are beneficial to all.

References

1. Kirkwood TB. Understanding the odd science of aging. *Cell.* 2005;120(4):437-447.

2. Kaeberlein M, Rabinovitch PS, Martin GM. Healthy aging: The ultimate preventative medicine. *Science.* 2015;350(6265):1191-1193.

3. Kennedy BK, Berger SL, Brunet A, et al. Geroscience: linking aging to chronic disease. *Cell.* 2014;159(4):709-713.

4. Niccoli T, Partridge L. Ageing as a risk factor for disease. *Curr Biol.* 2012;22(17):R741-752.

5. Gavrilov LA, Gavrilova NS. Evolutionary theories of aging and longevity. *ScientificWorldJournal.* 2002;2:339-356.

6. He W, D. Goodkind, and P. Kowal. An aging world: 2015. In. *International Population Reports*2016.

7. Bell CG, Lowe R, Adams PD, et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* 2019;20(1):249.

8. Vaupel JW. Biodemography of human ageing. *Nature.* 2010;464(7288):536-542.

9. Burch JB, Augustine AD, Frieden LA, et al. Advances in geroscience: impact on healthspan and chronic disease. *J Gerontol A Biol Sci Med Sci.* 2014;69 Suppl 1:S1-3.

10. Lara J, Cooper R, Nissan J, et al. A proposed panel of biomarkers of healthy ageing. *BMC Med.* 2015;13:222.

11. Ferrucci L, Giallauria F, Guralnik JM. Epidemiology of aging. *Radiol Clin North Am.* 2008;46(4):643-652, v.

12. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature.* 2007;447(7143):433-440.

13. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell.* 2013;153(6):1194-1217.

14. Bratic A, Larsson NG. The role of mitochondria in aging. *J Clin Invest.* 2013;123(3):951-957.

15. Tissenbaum HA, Guarente L. Model organisms as a guide to mammalian aging. *Dev Cell.* 2002;2(1):9-19.

16. Guarente L, Kenyon C. Genetic pathways that regulate ageing in model organisms. *Nature.* 2000;408(6809):255-262.

17. Wagner KH, Cameron-Smith D, Wessner B, Franzke B. Biomarkers of Aging: From Function to Molecular Biology. *Nutrients.* 2016;8(6).

18. Strimbu K, Tavel JA. What are biomarkers? *Curr Opin HIV AIDS.* 2010;5(6):463-466.

19. Vainio H. Use of biomarkers in risk assessment. *Int J Hyg Environ Health.* 2001;204(2-3):91-102.

20. Di Sanzo M, Cipolloni L, Borro M, et al. Clinical Applications of Personalized Medicine: A New Paradigm and Challenge. *Curr Pharm Biotechnol.* 2017;18(3):194-203.

21. Vogenberg FR, Isaacson Barash C, Pursel M. Personalized medicine: part 1: evolution and development into theranostics. *P T.* 2010;35(10):560-576.

22. George T. Baker RLS. Biomarkers of aging. *Experimental Gerontology.* 1988;23(4-5):223-239.

23. Johnson TE. Recent results: biomarkers of aging. *Exp Gerontol.* 2006;41(12):1243-1246.

24. Seals DR, Melov S. Translational geroscience: emphasizing function to achieve optimal longevity. *Aging (Albany NY).* 2014;6(9):718-730.

25. Seals DR, Justice JN, LaRocca TJ. Physiological geroscience: targeting function to increase healthspan and achieve optimal longevity. *J Physiol.* 2016;594(8):2001-2024.

26. Belsky DW, Caspi A, Houts R, et al. Quantification of biological aging in young adults. *Proc Natl Acad Sci U S A.* 2015;112(30):E4104-4110.

27. Christensen K, Thinggaard M, McGue M, et al. Perceived age as clinically useful biomarker of ageing: cohort study. *BMJ.* 2009;339:b5262.

28. Burkle A, Moreno-Villanueva M, Bernhard J, et al. MARK-AGE biomarkers of ageing. *Mech Ageing Dev.* 2015;151:2-12.

29. Waddington CH. The Epigenotpye. In. Endeavour1942 18–20.

30. Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. *Cell.* 2007;128(4):635-638.

31. Schuebel K, Gitik M, Domschke K, Goldman D. Making Sense of Epigenetics. *Int J Neuropsychopharmacol.* 2016;19(11).

32. Kanherkar RR, Bhatia-Dey N, Csoka AB. Epigenetics across the human lifespan. *Front Cell Dev Biol.* 2014;2:49.

33. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317-330.

34. Li E, Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol.* 2014;6(5):a019133.

35. Bird A. Perceptions of epigenetics. *Nature.* 2007;447(7143):396-398.

36. Geiman TM, Robertson KD. Chromatin remodeling, histone modifications, and DNA methylation-how does it all fit together? *J Cell Biochem.* 2002;87(2):117-125.

37. Kouzarides T. Chromatin modifications and their function. *Cell.* 2007;128(4):693-705.

38. Ho L, Crabtree GR. Chromatin remodelling during development. *Nature.* 2010;463(7280):474-484.

39. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin.* 2014;7(1):33.

40. Zentner GE, Henikoff S. High-resolution digital profiling of the epigenome. *Nat Rev Genet.* 2014;15(12):814-827.

41. Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. Epigenome characterization at single base-pair resolution. *Proc Natl Acad Sci U S A.* 2011;108(45):18318-18323.

42. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc.* 2010;2010(2):pdb.prot5384.

43. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin,

DNA-binding proteins and nucleosome position. *Nat Methods.* 2013;10(12):1213-1218.

44. Noll M. Subunit structure of chromatin. *Nature.* 1974;251(5472):249-251.

45. Cheung P, Allis CD, Sassone-Corsi P. Signaling to chromatin through histone modifications. *Cell.* 2000;103(2):263-271.

46. Nathan D, Sterner DE, Berger SL. Histone modifications: Now summoning sumoylation. *Proc Natl Acad Sci U S A.* 2003;100(23):13118-13120.

47. Robzyk K, Recht J, Osley MA. Rad6-dependent ubiquitination of histone H2B in yeast. *Science.* 2000;287(5452):501-504.

48. Kuo MH, Allis CD. Roles of histone acetyltransferases and deacetylases in gene regulation. *Bioessays.* 1998;20(8):615-626.

49. Kouzarides T. Histone methylation in transcriptional control. *Curr Opin Genet Dev.* 2002;12(2):198-209.

50. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res.* 2011;21(3):381-395.

51. Henikoff S, Shilatifard A. Histone modification: cause or cog? *Trends Genet.* 2011;27(10):389-396.

52. Mardis ER. ChIP-seq: welcome to the new frontier. *Nat Methods.* 2007;4(8):613-614.

53. Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. *Science.* 1975;187(4173):226-232.

54. Riggs AD. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet.* 1975;14(1):9-25.

55. Zampieri M, Ciccarone F, Calabrese R, Franceschi C, Burkle A, Caiafa P. Reconfiguration of DNA methylation in aging. *Mech Ageing Dev.* 2015;151:60-70.

56. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature.* 2009;462(7271):315-322.

57. Schultz MD, He Y, Whitaker JW, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature.* 2015;523(7559):212-216.

58. Eckhardt F, Lewin J, Cortese R, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet.* 2006;38(12):1378-1385.

59. Frommer M, McDonald LE, Millar DS, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A.* 1992;89(5):1827-1831.

60. Gao F, Xia Y, Wang J, et al. Integrated detection of both 5-mC and 5-hmC by high-throughput tag sequencing technology highlights methylation reprogramming of bivalent genes during cellular differentiation. *Epigenetics.* 2013;8(4):421-430.

61. Wu H, Zhang Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell.* 2014;156(1-2):45-68.

62. Shen L, Zhang Y. 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. *Curr Opin Cell Biol.* 2013;25(3):289-296.

63. Colquitt BM, Allen WE, Barnea G, Lomvardas S. Alteration of genic 5-hydroxymethylcytosine patterning in olfactory neurons correlates with changes in

gene expression and cell identity. *Proc Natl Acad Sci U S A.* 2013;110(36):14682-14687.

64. Bibikova M, Le J, Barnes B, et al. Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics.* 2009;1(1):177-200.

65. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet.* 2012;13(7):484-492.

66. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57-74.

67. Issa JP. Aging and epigenetic drift: a vicious cycle. *J Clin Invest.* 2014;124(1):24-29.

68. van Otterdijk SD, Mathers JC, Strathdee G. Do age-related changes in DNA methylation play a role in the development of age-related diseases? *Biochem Soc Trans.* 2013;41(3):803-807.

69. Wahl S, Drong A, Lehne B, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature.* 2017;541(7635):81-86.

70. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol.* 2012;41(1):161-176.

71. Mendelson MM, Marioni RE, Joehanes R, et al. Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Med.* 2017;14(1):e1002215.

72. Bonder MJ, Luijk R, Zhernakova DV, et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet.* 2017;49(1):131-138.

73. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet.* 2011;88(4):450-457.

74. Lindholm ME, Marabita F, Gomez-Cabrero D, et al. An integrative analysis reveals coordinated reprogramming of the epigenome and the transcriptome in human skeletal muscle after training. *Epigenetics.* 2014;9(12):1557-1569.

75. Lim U, Song MA. Dietary and lifestyle factors of DNA methylation. *Methods Mol Biol.* 2012;863:359-376.

76. Mitchell C, Schneper LM, Notterman DA. DNA methylation, early life environment, and health outcomes. *Pediatr Res.* 2016;79(1-2):212-219.

77. Yang BZ, Zhang H, Ge W, et al. Child abuse and epigenetic mechanisms of disease risk. *Am J Prev Med.* 2013;44(2):101-107.

78. Fraga MF, Ballestar E, Paz MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A.* 2005;102(30):10604-10609.

79. Fraga MF. Genetic and epigenetic regulation of aging. *Curr Opin Immunol.* 2009;21(4):446-453.

80. Pal S, Tyler JK. Epigenetics and aging. *Sci Adv.* 2016;2(7):e1600584.

81. Fraga MF, Esteller M. Epigenetics and aging: the targets and the marks. *Trends Genet.* 2007;23(8):413-418.

82. Lopatina N, Haskell JF, Andrews LG, Poole JC, Saldanha S, Tollefsbol T. Differential maintenance and de novo methylating activity by three DNA methyltransferases in aging and immortalized fibroblasts. *J Cell Biochem.* 2002;84(2):324-334.

83. Casillas MA, Lopatina N, Andrews LG, Tollefsbol TO. Transcriptional control of the DNA methyltransferases is altered in aging and neoplastically-transformed human fibroblasts. *Mol Cell Biochem.* 2003;252(1-2):33-43.

84. Marttila S, Kananen L, Häyrynen S, et al. Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression. *BMC Genomics.* 2015;16:179.

85. Issa JP, Ottaviano YL, Celano P, Hamilton SR, Davidson NE, Baylin SB. Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nat Genet.* 1994;7(4):536-540.

86. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature.* 1983;301(5895):89-92.

87. Teschendorff AE, Menon U, Gentry-Maharaj A, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* 2010;20(4):440-446.

88. Hernandez DG, Nalls MA, Gibbs JR, et al. Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Hum Mol Genet.* 2011;20(6):1164-1172.

89. Rakyan VK, Down TA, Maslau S, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* 2010;20(4):434-439.

90. Bocklandt S, Lin W, Sehl ME, et al. Epigenetic predictor of age. *PLoS One.* 2011;6(6):e14821.

91. Bell JT, Tsai PC, Yang TP, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.* 2012;8(4):e1002629.

92. Slieker RC, Relton CL, Gaunt TR, Slagboom PE, Heijmans BT. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics Chromatin.* 2018;11(1):25.

93. Fagnoni FF, Vescovini R, Passeri G, et al. Shortage of circulating naive CD8(+) T cells provides new insights on immunodeficiency in aging. *Blood.* 2000;95(9):2860-2868.

94. Jergović M, Smithey MJ, Nikolich-Žugich J. Intrinsic and extrinsic contributors to defective CD8+ T cell responses with aging. *Exp Gerontol.* 2018;105:140-145.

95. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One.* 2012;7(7):e41361.

96. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012;13:86.

97. Koestler DC, Jones MJ, Usset J, et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics.* 2016;17:120.

98. Houseman EA, Kile ML, Christiani DC, Ince TA, Kelsey KT, Marsit CJ. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics.* 2016;17:259.

99. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell.* 2013;49(2):359-367.

100. Garagnani P, Bacalini MG, Pirazzini C, et al. Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell.* 2012;11(6):1132-1134.

101. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013;14(10):R115.

102. Alisch RS, Barwick BG, Chopra P, et al. Age-associated DNA methylation in pediatric populations. *Genome Res.* 2012;22(4):623-632.

103. Bollati V, Schwartz J, Wright R, et al. Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech Ageing Dev.* 2009;130(4):234-239.

104. Chen BH, Marioni RE, Colicino E, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY).* 2016;8(9):1844-1865.

105. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Mol Cell.* 2018;71(6):882-895.

106. Slieker RC, van Iterson M, Luijk R, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.* 2016;17(1):191.

107. Weidner CI, Lin Q, Koch CM, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* 2014;15(2):R24.

108. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet.* 2018;19(6):371-384.

109. Jylhävä J, Pedersen NL, Hägg S. Biological Age Predictors. *EBioMedicine.* 2017;21:29-36.

110. Marioni RE, Shah S, McRae AF, et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* 2015;16:25.

111. Quach A, Levine ME, Tanaka T, et al. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging (Albany NY).* 2017.

112. Grant CD, Jafari N, Hou L, et al. A longitudinal study of DNA methylation as a potential mediator of age-related diabetes risk. *Geroscience.* 2017;39(5-6):475-489.

113. Levine ME, Lu AT, Chen BH, et al. Menopause accelerates biological aging. *Proc Natl Acad Sci U S A.* 2016;113(33):9327-9332.

114. Zheng Y, Joyce BT, Colicino E, et al. Blood Epigenetic Age may Predict Cancer Incidence and Mortality. *EBioMedicine.* 2016;5:68-73.

115. Breitling, Philipp L, Saum K-U, et al. Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. In. Vol 82016:1.

116. Horvath S, Garagnani P, Bacalini MG, et al. Accelerated epigenetic aging in Down syndrome. *Aging Cell.* 2015;14(3):491-495.

117. Zannas AS, Arloth J, Carrillo-Roa T, et al. Lifetime stress accelerates epigenetic aging in an urban, African American cohort: relevance of glucocorticoid signaling. *Genome Biol.* 2015;16:266.

118. Nevalainen T, Kananen L, Marttila S, et al. Obesity accelerates epigenetic aging in middle-aged but not in elderly individuals. *Clin Epigenetics.* 2017;9:20.

119. Levine ME, Lu AT, Bennett DA, Horvath S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging (Albany NY).* 2015;7(12):1198-1211.

120. Horvath S, Pirazzini C, Bacalini MG, et al. Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. *Aging (Albany NY).* 2015;7(12):1159-1170.

121. Horvath S, Gurven M, Levine ME, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol.* 2016;17(1):171.

122. Levine ME, Lu AT, Quach A, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY).* 2018;10(4):573-591.

123. Lu AT, Quach A, Wilson JG, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY).* 2019;11(2):303-327.

124. Levine M, E, Hosgood H, et al. DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. In. Vol 7: Aging; 2015:690.

125. Perna L, Zhang Y, Mons U, Holleczek B, Saum KU, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin Epigenetics.* 2016;8:64.

126. Kananen L, Marttila S, Nevalainen T, et al. The trajectory of the blood DNA methylome ageing rate is largely set before adulthood: evidence from two longitudinal studies. *Age (Dordr).* 2016;38(3):65.

127. Peters MJ, Joehanes R, Pilling LC, et al. The transcriptional landscape of age in human peripheral blood. *Nat Commun.* 2015;6:8570.

128. Zhang Q, Vallerga CL, Walker RM, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* 2019;11(1):54.

129. Talens RP, Christensen K, Putter H, et al. Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell.* 2012;11(4):694-703.

130. Christiansen L, Lenart A, Tan Q, et al. DNA methylation age is associated with mortality in a longitudinal Danish twin study. *Aging Cell.* 2016;15(1):149-154.

131. Wang Y, Pedersen NL, Hägg S. Implementing a method for studying longitudinal DNA methylation variability in association with age. *Epigenetics.* 2018;13(8):866-874.

132. Zhang Q, Marioni RE, Robinson MR, et al. Genotype effects contribute to variation in longitudinal methylome patterns in older people. *Genome Med.* 2018;10(1):75.

133. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet.* 2018;19(3):129-147.

134. Wang Y, Karlsson R, Lampa E, et al. Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins. *Epigenetics.* 2018;13(9):975-987.

135. Garg P, Joshi RS, Watson C, Sharp AJ. A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PLoS Genet.* 2018;14(10):e1007707.

136. Langevin SM, Pinney SM, Leung YK, Ho SM. Does epigenetic drift contribute to age-related increases in breast cancer risk? *Epigenomics.* 2014;6(4):367-369.

137. Franceschi C, Campisi J. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *J Gerontol A Biol Sci Med Sci.* 2014;69 Suppl 1:S4-9.

138. Webster AP, Plant D, Ecker S, et al. Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. *Genome Med.* 2018;10(1):64.

139. Cheung P, Vallania F, Warsinske HC, et al. Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. *Cell.* 2018;173(6):1385-1397.e1314.

140. Karlić R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A.* 2010;107(7):2926-2931.

141. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.* 2011;473(7345):43-49.

142. Rose NR, Klose RJ. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta.* 2014;1839(12):1362-1372.

143. Tserel L, Kolde R, Limbach M, et al. Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. *Sci Rep.* 2015;5:13107.

144. Giaimo BD, Ferrante F, Herchenröther A, Hake SB, Borggrefe T. The histone variant H2A.Z in gene regulation. *Epigenetics Chromatin.* 2019;12(1):37.

145. Yin Y, Morgunova E, Jolma A, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science.* 2017;356(6337).

146. Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet.* 2009;10(5):295-304.

147. Tsurumi A, Li WX. Global heterochromatin loss: a unifying theory of aging? *Epigenetics.* 2012;7(7):680-688.

148. Nandakumar J, Cech TR. Finding the end: recruitment of telomerase to telomeres. *Nat Rev Mol Cell Biol.* 2013;14(2):69-82.

149. Feser J, Tyler J. Chromatin structure as a mediator of aging. *FEBS Lett.* 2011;585(13):2041-2048.

150. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860-921.

151. Miki Y, Nishisho I, Horii A, et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.* 1992;52(3):643-645.

152. De Cecco M, Criscione SW, Peterson AL, Neretti N, Sedivy JM, Kreiling JA. Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging (Albany NY).* 2013;5(12):867-883.

153. Song S, Johnson FB. Epigenetic Mechanisms Impacting Aging: A Focus on Histone Levels and Telomeres. *Genes (Basel).* 2018;9(4).

154. O'Sullivan RJ, Kubicek S, Schreiber SL, Karlseder J. Reduced histone biosynthesis and chromatin changes arising from a damage signal at telomeres. *Nat Struct Mol Biol.* 2010;17(10):1218-1225.

155. Kreiling JA, Tamamori-Adachi M, Sexton AN, et al. Age-associated increase in heterochromatic marks in murine and primate tissues. *Aging Cell.* 2011;10(2):292-304.

156. Sarg B, Koutzamani E, Helliger W, Rundquist I, Lindner HH. Postsynthetic trimethylation of histone H4 at lysine 20 in mammalian tissues is associated with aging. *J Biol Chem.* 2002;277(42):39195-39201.

157. Liu L, Cheung TH, Charville GW, et al. Chromatin modifications as determinants of muscle stem cell quiescence and chronological aging. *Cell Rep.* 2013;4(1):189-204.

158. McCauley BS, Dang W. Histone methylation and aging: lessons learned from model systems. *Biochim Biophys Acta.* 2014;1839(12):1454-1462.

159. Xu Z, Taylor JA. Genome-wide age-related DNA methylation changes in blood and other tissues relate to histone modification, expression and cancer. *Carcinogenesis.* 2014;35(2):356-364.

160. Edwards JR, O'Donnell AH, Rollins RA, et al. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res.* 2010;20(7):972-980.

161. Federation ID. IDF Diabetes Atlas, 7 ed. . In. Brussels, Belgium: International Diabetes Federation; 2015.

162. Hu FB. Globalization of diabetes: the role of diet, lifestyle, and genes. *Diabetes Care.* 2011;34(6):1249-1257.

163. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract.* 2010;87(1):4-14.

164. Ford ES, Williamson DF, Liu S. Weight change and diabetes incidence: findings from a national cohort of US adults. *Am J Epidemiol.* 1997;146(3):214-222.

165. Chan JM, Rimm EB, Colditz GA, Stampfer MJ, Willett WC. Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. *Diabetes Care.* 1994;17(9):961-969.

166. Koh-Banerjee P, Wang Y, Hu FB, Spiegelman D, Willett WC, Rimm EB. Changes in body weight and body fat distribution as risk factors for clinical diabetes in US men. *Am J Epidemiol.* 2004;159(12):1150-1159.

167. Shai I, Jiang R, Manson JE, et al. Ethnicity, obesity, and risk of type 2 diabetes in women: a 20-year follow-up study. *Diabetes Care.* 2006;29(7):1585-1590.

168. Hu FB, Manson JE, Stampfer MJ, et al. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N Engl J Med.* 2001;345(11):790-797.

169. Nathan DM, Davidson MB, DeFronzo RA, et al. Impaired fasting glucose and impaired glucose tolerance: implications for care. *Diabetes Care.* 2007;30(3):753-759.

170. Weyer C, Hanson RL, Tataranni PA, Bogardus C, Pratley RE. A high fasting plasma insulin concentration predicts type 2 diabetes independent of insulin resistance: evidence for a pathogenic role of relative hyperinsulinemia. *Diabetes.* 2000;49(12):2094-2101.

171. Stamler J, Vaccaro O, Neaton JD, Wentworth D. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. *Diabetes Care.* 1993;16(2):434-444.

172. Mokdad AH, Ford ES, Bowman BA, et al. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA.* 2003;289(1):76-79.

173. Pan XR, Li GW, Hu YH, et al. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. *Diabetes Care.* 1997;20(4):537-544.

174. Christensen BC, Houseman EA, Marsit CJ, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.* 2009;5(8):e1000602.

175. Marioni RE, Shah S, McRae AF, et al. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *Int J Epidemiol.* 2015;44(4):1388-1396.

176. Rönn T, Ling C. DNA methylation as a diagnostic and therapeutic target in the battle against Type 2 diabetes. *Epigenomics.* 2015;7(3):451-460.

177. Hidalgo B, Irvin MR, Sha J, et al. Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network study. *Diabetes.* 2014;63(2):801-807.

178. Nilsson E, Jansson PA, Perfilyev A, et al. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes.* 2014;63(9):2962-2976.

179. Group TWsHIS. Design of the women's health initiative clinical trial and observational study. In. Vol 19: *Controlled clinical trials*; 1998:61-109.

180. Hays J, Hunt JR, Hubbell F, et al. The Women's Health Initiative Recruitment Methods and Results. In. Vol 13: Annals of epidemiology; 2003:S18-S77.

181. Yokoyama H, Emoto M, Fujiwara S, et al. Quantitative insulin sensitivity check index and the reciprocal index of homeostasis model assessment in normal range weight and moderately obese type 2 diabetic patients. *Diabetes Care.* 2003;26(8):2426-2432.

182. McLaughlin T, Abbasi F, Cheal K, Chu J, Lamendola C, Reaven G. Use of metabolic markers to identify overweight individuals who are insulin resistant. *Ann Intern Med.* 2003;139(10):802-809.

183. Simental-Mendía LE, Rodríguez-Morán M, Guerrero-Romero F. The product of fasting glucose and triglycerides as surrogate for identifying insulin resistance in apparently healthy subjects. *Metab Syndr Relat Disord.* 2008;6(4):299-304.

184. Heinemann L. Insulin Assay Standardization: Leading to Measures of Insulin Sensitivity and Secretion for Practical Clinical Care Response to Staten et al. In. Vol 33: *Diabetes care*; 2010:e83-e83.

185. Patterson RE, Kristal AR, Tinker LF, Carter RA, Bolton MP, Agurs-Collins T. Measurement characteristics of the Women's Health Initiative food frequency questionnaire. *Ann Epidemiol.* 1999;9(3):178-187.

186. Association AD. Erratum. Classification and diagnosis of diabetes. Sec. 2. In Standards of Medical Care in Diabetes-2016. Diabetes Care 2016;39(Suppl. 1):S13-S22. *Diabetes Care.* 2016;39(9):1653.

187. Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics.* 2013;29(2):189-196.

188. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012;28(6):882-883.

189. Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics.* 2012;28(9):1280-1281.

190. Pinheiro J, DouglasDebRoy, Saikat Sarkar, Deepayan, Team RC. *nlme: Linear and Nonlinear Mixed Effects Models*. In. Vol 3.1-128: R package version; 2016.

191. Adalsteinsson BT, Gudnason H, Aspelund T, et al. Heterogeneity in white blood cells has potential to confound DNA methylation measurements. *PLoS One.* 2012;7(10):e46705.

192. Colberg SR, Albright AL, Blissmer BJ, et al. Exercise and type 2 diabetes: American College of Sports Medicine and the American Diabetes Association: joint position statement. Exercise and type 2 diabetes. *Med Sci Sports Exerc.* 2010;42(12):2282-2303.

193. Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient variation. In. Vol 47: *Econometrica: Journal of the Econometric Society* 1979:1287-1294.

194. Smyth G, Verbyla A. Double generalized linear models: approximate REML and diagnostics. *Statistical Modelling: Proceedings of the 14th International Workshop on Statistical Modelling.* 1999.

195. van Greevenbroek MM, Jacobs M, van der Kallen CJ, et al. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur J Clin Invest.* 2011;41(4):372-379.

196. Tigchelaar EF, Zhernakova A, Dekens JA, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open.* 2015;5(8):e006772.

197. Westendorp RG, van Heemst D, Rozing MP, et al. Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study. *J Am Geriatr Soc.* 2009;57(9):1634-1637.

198. Boomsma DI, Vink JM, van Beijsterveldt TC, et al. Netherlands Twin Register: a focus on longitudinal research. *Twin Res.* 2002;5(5):401-406.

199. Hofman A, Brusselle GG, Darwish Murad S, et al. The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol.* 2015;30(8):661-708.

200. Huisman MH, de Jong SW, van Doormaal PT, et al. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *J Neurol Neurosurg Psychiatry.* 2011;82(10):1165-1170.

201. van Iterson M, Tobi EW, Slieker RC, et al. MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics.* 2014;30(23):3435-3437.

202. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* 2017;45(4):e22.

203. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics.* 2014;30(10):1363-1369.

204. Sinke L, van Iterson M, Cats D, Slieker R, Heijmans B. DNAmArray: Streamlined workflow for the quality control, normalization, and analysis of Illumina methylation array data (Version 2.1). *Zenodo.* 2019.

205. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-127.

206. Westra HJ, Jansen RC, Fehrmann RS, et al. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics.* 2011;27(15):2104-2111.

207. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* 2014;15(2):R31.

208. van Iterson M. **https://github.com/mvaniterson/wbccPredictor**.

209. Du P, Zhang X, Huang CC, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics.* 2010;11:587.

210. Dunn PK, Smyth GK. dglm: Double Generalized Linear Models. *R package version 183.* 2016.

211. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* 2013;41(Database issue):D991-995.

212. Reynolds LM, Taylor JR, Ding J, et al. Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nat Commun.* 2014;5:5366.

213. Joehanes R, Just AC, Marioni RE, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet.* 2016;9(5):436-447.

214. Bollepalli S, Korhonen T, Kaprio J, Ollikainen M, Anders S. EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. *bioRxiv.* 2019;487975.

215. Rainer J, Gatto L, Weichenberger CX. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics.* 2019.

216. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43(Database issue):D447-452.

217. Ye J, Zhang Y, Cui H, et al. WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Res.* 2018;46(W1):W71-W75.

218. Deary IJ, Whiteman MC, Starr JM, Whalley LJ, Fox HC. The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. *J Pers Soc Psychol.* 2004;86(1):130-147.

219. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics.* 2008;24(13):1547-1548.

220. Tyler CDaJK. Histone exchange and histone modifications during transcription and aging. 2013.

221. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009;10(1):57-63.

222. Moskowitz DM, Zhang DW, Hu B, et al. Epigenomics of human CD8 T cell differentiation and aging. *Sci Immunol.* 2017;2(8).

223. Ucar D, Márquez EJ, Chung CH, et al. The chromatin accessibility signature of human immune aging stems from CD8. *J Exp Med.* 2017;214(10):3123-3144.

224. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods.* 2012;9(3):215-216.

225. Johnson ND, Huang L, Li R, et al. Age-related DNA hydroxymethylation is enriched for gene expression and immune system processes in human peripheral blood. *Epigenetics.* 2019:1-13.

226. Kovalenko A, Chable-Bessia C, Cantarella G, Israël A, Wallach D, Courtois G. The tumour suppressor CYLD negatively regulates NF-kappaB signalling by deubiquitination. *Nature.* 2003;424(6950):801-805.

227. Trompouki E, Hatzivassiliou E, Tsichritzis T, Farmer H, Ashworth A, Mosialos G. CYLD is a deubiquitinating enzyme that negatively regulates NF-kappaB activation by TNFR family members. *Nature.* 2003;424(6950):793-796.

228. Alameda JP, Ramírez Á, García-Fernández RA, et al. Premature aging and cancer development in transgenic mice lacking functional CYLD. *Aging (Albany NY).* 2019;11(1):127-159.

229. Pfaffenbach KT, Lee AS. The critical role of GRP78 in physiologic and pathologic stress. *Curr Opin Cell Biol.* 2011;23(2):150-156.

230. Dong D, Ni M, Li J, et al. Critical role of the stress chaperone GRP78/BiP in tumor proliferation, survival, and tumor angiogenesis in transgene-induced mammary tumor development. *Cancer Res.* 2008;68(2):498-505.

231. Casas C. GRP78 at the Centre of the Stage in Cancer and Neuroprotection. *Front Neurosci.* 2017;11:177.

232. Li W, Wang W, Dong H, et al. Cisplatin-induced senescence in ovarian cancer cells is mediated by GRP78. *Oncol Rep.* 2014;31(6):2525-2534.

233. O'Connell GC, Treadway MB, Petrone AB, et al. Peripheral blood AKAP7 expression as an early marker for lymphocyte-mediated post-stroke blood brain barrier disruption. *Sci Rep.* 2017;7(1):1172.

234. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-2120.

235. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25.

236. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137.

237. Kellis M, Wold B, Snyder MP, et al. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A.* 2014;111(17):6131-6138.

238. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21.

239. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-842.

240. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.

241. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. *Bioinformatics.* 2014;30(19):2811-2812.

242. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010;38(4):576-589.

243. Lawrence M, Huber W, Pagès H, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol.* 2013;9(8):e1003118.

244. Marini F. *ideal: Interactive Differential Expression AnaLysis*. In:2018.

245. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 2010;11(2):R14.

246. Watanabe C, Morita M, Hayata T, et al. Stability of mRNA influences osteoporotic bone mass via CNOT3. *Proc Natl Acad Sci U S A.* 2014;111(7):2692-2697.

247. Glass D, Viñuela A, Davies MN, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol.* 2013;14(7):R75.

248. Kim M, Lee KT, Jang HR, et al. Epigenetic down-regulation and suppressive role of DCBLD2 in gastric cancer cell proliferation and invasion. *Mol Cancer Res.* 2008;6(2):222-230.

249. Mooijaart SP, van Heemst D, Schreuder J, et al. Variation in the SHC1 gene and longevity in humans. *Exp Gerontol.* 2004;39(2):263-268.

250. Peng C, Zhang Z, Wu J, et al. A critical role for ZDHHC2 in metastasis and recurrence in human hepatocellular carcinoma. *Biomed Res Int.* 2014;2014:832712.

251. Yan SM, Tang JJ, Huang CY, et al. Reduced expression of ZDHHC2 is associated with lymph node metastasis and poor prognosis in gastric adenocarcinoma. *PLoS One.* 2013;8(2):e56366.

252. Vick AD, Burris HH. Epigenetics and Health Disparities. *Curr Epidemiol Rep.* 2017;4(1):31-37.

253. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature.* 2016;538(7624):161-164.

254. Obama B. United States Health Care Reform: Progress to Date and Next Steps. *JAMA.* 2016;316(5):525-532.

255. Heyn H, Moran S, Hernando-Herraez I, et al. DNA methylation contributes to natural human variation. *Genome Res.* 2013;23(9):1363-1372.

256. Ruiz JM, Steffen P, Smith TB. Hispanic mortality paradox: a systematic review and meta-analysis of the longitudinal literature. *Am J Public Health.* 2013;103(3):e52-60.

257. Hunt SC, Chen W, Gardner JP, et al. Leukocyte telomeres are longer in African Americans than in whites: the National Heart, Lung, and Blood Institute Family Heart Study and the Bogalusa Heart Study. *Aging Cell.* 2008;7(4):451-458.

258. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* 2016;17(1):157.

259. Morales J, Welter D, Bowler EH, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* 2018;19(1):21.

260. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell.* 2019;177(1):26-31.

261. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature.* 2019;570(7762):514-518.

262. Drozda K, Wong S, Patel SR, et al. Poor warfarin dose prediction with pharmacogenetic algorithms that exclude genotypes important for African Americans. *Pharmacogenet Genomics.* 2015;25(2):73-81.

263. Putin E, Mamoshina P, Aliper A, et al. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging (Albany NY).* 2016;8(5):1021-1033.

264. Cohen AA, Morissette-Thomas V, Ferrucci L, Fried LP. Deep biomarkers of aging are population-dependent. *Aging (Albany NY).* 2016;8(9):2253-2255.

265. Desrosiers J, Hébert R, Bravo G, Rochette A. Comparison of cross-sectional and longitudinal designs in the study of aging of upper extremity performance. *J Gerontol A Biol Sci Med Sci.* 1998;53(5):B362-368.

266. Knopf M, Neidhardt E. [Age differences versus aging--a cross-sectional and longitudinal analysis on the development of memory in advanced age]. *Z Gerontol Geriatr.* 1995;28(2):129-139.

267. Pfefferbaum A, Sullivan EV. Cross-sectional versus longitudinal estimates of age-related changes in the adult brain: overlaps and discrepancies. *Neurobiol Aging.* 2015;36(9):2563-2567.

268. Knight AK, Craig JM, Theda C, et al. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol.* 2016;17(1):206.

269. Suarez A, Lahti J, Czamara D, et al. The epigenetic clock and pubertal, neuroendocrine, psychiatric, and cognitive outcomes in adolescents. *Clin Epigenetics.* 2018;10(1):96.

270. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief Bioinform.* 2014;15(6):929-941.

271. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics.* 2016;8(3):389-399.

272. Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science.* 2013;339(6121):819-823.

273. Mali P, Yang L, Esvelt KM, et al. RNA-guided human genome engineering via Cas9. *Science.* 2013;339(6121):823-826.

274. Hilton IB, D'Ippolito AM, Vockley CM, et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol.* 2015;33(5):510-517.

275. Mendenhall EM, Williamson KE, Reyon D, et al. Locus-specific editing of histone modifications at endogenous enhancers. *Nat Biotechnol.* 2013;31(12):1133-1136.

276. Bell CG, Xia Y, Yuan W, et al. Novel regional age-associated DNA methylation changes within human common disease-associated loci. *Genome Biol.* 2016;17(1):193.

277. Barros-Silva D, Marques CJ, Henrique R, Jerónimo C. Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. *Genes (Basel).* 2018;9(9).

278. Cazaly E, Saad J, Wang W, Heckman C, Ollikainen M, Tang J. Making Sense of the Epigenome Using Data Integration Approaches. *Front Pharmacol.* 2019;10:126.

279. Gross JA, Pacis A, Chen GG, Barreiro LB, Ernst C, Turecki G. Characterizing 5-hydroxymethylcytosine in human prefrontal cortex at single base resolution. *BMC Genomics.* 2015;16:672.

280. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol.* 2008;9(12):958-970.

281. Pagiatakis C, Musolino E, Gornati R, Bernardini G, Papait R. Epigenetics of aging and disease: a brief overview. *Aging Clin Exp Res.* 2019.

282. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet.* 2013;14(9):618-630.

283. Wagner W. Epigenetic aging clocks in mice and men. *Genome Biol.* 2017;18(1):107.