UNIVERSITY OF CALIFORNIA

Los Angeles

Determinants of the epigenetic clock

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Human Genetics

by

Austin Quach

2018

ABSTRACT OF THE DISSERTATION

Determinants of the epigenetic clock

by

Austin Quach

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2018

Professor Stefan Horvath, Co-Chair

Professor Kym F Faull, Co-Chair

It has been observed that the epigenome exhibits significant changes with increased age. These changes are so consistent that they have been used to develop "epigenetic clock" models which can predict chronological age with high accuracy and have been shown to be independent predictors of age-related disease outcomes. In this dissertation, I investigate the relationship between epigenetic aging and lifestyle and transcriptomic factors in order to elucidate the underlying biology of this phenomenon. I find that epigenetic aging in blood is multifactorial and is consistent with modern notions of health. My experiences with poor sample annotations associated with high dimensional genomics data led me to develop a new assay to simultaneously measure proteins, lipids, metabolites, and other molecules.

The dissertation of Austin Quach is approved.

Roel A. Ophoff

Janet S. Sinsheimer

Kym F. Faull, Committee Co-Chair

Steve Horvath, Committee Co-Chair

University of California, Los Angeles

2018

To my loving wife who has supported my endeavors throughout and without whom this would not be possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my co-mentors Steve Horvath and Kym Faull. Under the mentorship of Dr. Horvath, I was able to gain a firm grasp on statistical analysis, machine learning, and systems biology. Under the mentorship of Dr. Faull, I was able to learn about analytical chemistry, mass spectrometry, and biomarker development. Besides their intellectual and academic support, both of them have offered me invaluable guidance, while at the same time providing an ideal amount freedom to develop my own independent sense of curiosity.

I would also like to thank my doctoral committee members Professors Janet Sinsheimer and Roel Ophoff for their flexibility and guidance throughout my doctoral training. Their critical assessment of my work has driven my analytical standards to new levels. I am grateful for the acceptance and support of the UCLA ACCESS and Human Genetics programs, the Burroughs Wellcome Fund Inter-school Training Program in Chronic Diseases, and the UCLA Graduate Division Dissertation Year Fellowship. I would like to thank my friends and family for their continuous encouragement and support of my continued education. Finally, I would like to thank God for making all of this possible.

I started my doctoral training with the idea that computational biology would be important for my long-term research interest in personalized medicine. With this in mind, I sought the mentorship of Professors Steve Horvath and Kym Faull in order to gain an understanding of systems biology and analytical chemistry, respectively. Through my experiences and explorations in these research groups, I was able to continually refine my understanding of the biomedical field and to begin to formulate ideas my own ideas. As a result this dissertation does not examine a single scientific topic but rather reflects my evolving curiosities over the course of my doctoral training.

In Chapter 0, I briefly introduce biological aging, DNA methylation, and the epigenetic clock. This includes technical background information and the motivation for studying these topics in order to communicate the context of the dissertation and my research contributions.

Chapter 1 is adapted from a published paper: Quach, A., Levine, M. E., Tanaka, T., Lu, A. T., Chen, B. H., Ferrucci, L., … Horvath, S. (2017). Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging (Albany NY), 9*(2), 419–437. http://doi.org/10.18632/aging.101168. The authors would like to thank the Women's Health Initiative and the Invecchaire Chianti cohort studies for allowing us to contribute to their study.

Chapter 2 is adapted from a published pape: Quach, A., Horvath, S., Nemanim, N., Vatakis, D., Witt, M. D., Miller, E. N., ... & Levine, A. J. (2018). No reliable gene expression biomarkers of current or impending neurocognitive impairment in peripheral blood monocytes of persons living with HIV. *Journal of neurovirology*, 1-12. https://doi.org/10.1007/s13365-018-0625-5. The authors would like to thank the participants and staff of the Multicenter AIDS Cohort Study in Los Angeles.

Chapter 3 describes unpublished work on the associations between epigenetic age acceleration and global gene expression in peripheral blood mononuclear cells and in purified monocytes. I would like to thank the Framingham Heart Study and the Multi-Ethnic Study of Atherosclerosis cohort studies for allowing me to analyze their data.

Chapter 4 describes unpublished work on the development of a universal liquid chromatography-mass spectrometry assay. I would like to thank the Pasarow Mass Spectrometry Laboratory for allowing me to use their instrumentation and laboratory equipment and Thermo Fischer for generously supplying with liquid chromatography supplies.

VITA

Austin Quach

Education

Bachelor of Science, Biochemistry, University of California, Los Angeles, 2012

Fellowships and Grants

UCLA Dissertation Year Fellowship, Summer 2017 to Spring 2018

Burroughs Wellcome Fund Inter-school Training Program in Chronic Diseases Fellowship, Fall 2015 to Spring 2017

Publications

Levine, M. E., Lu, A. T., **Quach, A.**, Chen, B., Assimes, T. L., Bandinelli, S., ... & Whitsel, E. A. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging*. In press.

Lu, A. T., Xue, L., Salfati, E. L., Chen, B. H., Ferrucci, L., Levy, D., ... **Quach, A.**, ... & Horvath, S. (2018). GWAS of epigenetic aging rates in blood reveals a critical role for TERT. *Nature communications, 9*(1), 387.

Quach, A., Horvath, S., Nemanim, N., Vatakis, D., Witt, M. D., Miller, E. N., ... & Levine, A. J. (2018). No reliable gene expression biomarkers of current or impending neurocognitive impairment in peripheral blood monocytes of persons living with HIV. *Journal of neurovirology*, 1-12.

Quach, A., Levine, M. E., Tanaka, T., Lu, A. T., Chen, B. H., Ferrucci, L., ... Horvath, S. (2017). Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging*. doi:10.18632/aging.101168

Chuang, Y.-H., **Quach, A.**, Absher, D., Assimes, T., Horvath, S., & Ritz, B. (2017). Coffee consumption is associated with DNA methylation levels of human blood. *Eur J Hum Genet*. doi:10.1038/ejhg.2016.175

Levine, M. E., Lu, A. T., Chen, B. H., Hernandez, D. G., Singleton, A. B., Ferrucci, L., ... **Quach, A.**, ... Horvath, S. (2016). Menopause accelerates biological aging. *Proceedings of the National Academy of Sciences*, 113(33), 9327-9332. doi:10.1073/pnas.1604558113

Levine, A. J., **Quach, A.**, Moore, D. J., Achim, C. L., Soontornniyomkij, V., Masliah, E., ... Horvath, S. (2016). Accelerated epigenetic aging in brain is associated with pre-mortem HIV-associated neurocognitive disorders. *Journal of neurovirology*, 22(3), 366-375. doi:10.1007/s13365-015-0406-3

Rickabaugh, T. M., Baxter, R. M., Sehl, M., Sinsheimer, J. S., Hultin, P. M., Hultin, L. E., ... **Quach, A.**, ... Jamieson, B. D. (2015). Acceleration of Age-Associated Methylation Patterns in HIV-1-Infected Adults. *PLoS one*, 10(3), e0119201. doi:10.1371/journal.pone.0119201

Thomer, L., Becker, S., Emolo, C., **Quach, A.**, Kim, H. K., Rauch, S., ... Missiakas, D. (2014). N-Acetylglucosaminylation of Serine-Aspartate Repeat Proteins Promotes Staphylococcus aureus Bloodstream Infection. *Journal of Biological Chemistry*, 289(6), 3478-3486. doi:10.1074/jbc.M113.532655

Chin, R. M., Fu, X., Pai, M. Y., Vergnes, L., Hwang, H., Deng, G., ... **Quach, A.**, ... Huang, J. (2014). The metabolite [alpha]-ketoglutarate extends lifespan by inhibiting ATP synthase and TOR. *Nature*, 510(7505), 397-401. doi:10.1038/nature13264

Oral presentations

Quach, A., Horvath S. (2017 October 25) Diet and epigenetic ageing. Oral presentation at: *Food for Healthy Aging*. Amsterdam, Netherlands.

Quach, A., Faull K.F. (2016 September 27). Approaches to analyzing global GC/MS metabolomics data. Oral presentation at: *Metabolomic and Lipidomic Symposium hosted by Thermo Fisher Scientific at California Institute of Technology*. Pasadena, CA.

Quach, A., Levine, M. E., Tanaka, T., Lu, A. T., Chen, B. H., Ferrucci, L., ... Horvath, S. (2016 April 28). Epigenetic clock analysis of diet and metabolic health. Selected abstract oral presentation at: *UCLA Annual Joint Research Symposium*. Los Angeles, CA.

Quach, A. (2016 April 28) Diet and epigenetic aging. The Systems and Integrative Biology Training Grant, The Biomedical Big Data Training Grant, The Burroughs Wellcome Fund Inter-school Training Program in Metabolic Diseases, The Genomic Analysis & Interpretation Training Program Joint Research Symposium. Los Angeles, California.

Chapter 0: A brief introduction to epigenetic aging

Biological aging

Biological aging occurs across nearly all known animal species with rare exceptions such as hydra [1] and jellyfish [2]. Many different lenses have been used to study the influence of genetic and environmental factors on biological aging including laboratory research on model organisms and epidemiological studies in human populations. This body of work has identified a number of molecular changes that coincide with aging including the accumulation of genomic damage, telomere attrition, epigenomic alterations, and loss of proteo-homeostasis [3]. Among all of the proposed interventions in the field perhaps the most well-established is dietary restriction which appears to act in large part through the inhibition of the cellular kinase mTOR to promote protein turnover through autophagy. Various forms of dietary restriction, such as reduction of ad libitum caloric intake by ~30%, have been demonstrated to extend healthy lifespan in a range of animals including worms, fruit flies, rodents, and (arguably) rhesus monkeys [4].

Studying the effects of anti-aging interventions such as dietary restriction in humans presents a number of logistical challenges. Difficulties in observational studies include the costs of measuring long-term outcomes such as disease incidence or mortality, measurement error due to self-reporting, and confounding due to unmeasured factors. Though some of these issues can be addressed through randomized clinical trials, conventional aging-related outcomes still require long-term follow-ups in order to observe significant inhibition or reversal. Surrogate biomarkers (e.g. cholesterol for cardiovascular disease risk) can be used to track intermediate stages of phenotypic progression and thus avoid the costs of conducting full-term aging studies. The advent of DNA methylation based predictors of age has elicited much interest as these

biomarkers may act as direct surrogates for biological aging. In 2013 the epigenetic clock was developed, enabling the accurate prediction of chronological age across nearly all human tissues using the methylation levels of only a few hundred CpG sites [5]. Intriguingly this model is able to perform well even in non-dividing cells such as neurons. This and other DNA methylation based models promise to catalyze progress in the field by providing surrogate measures which will enable the study of aging with a reduced need to track traditional age-related outcomes.

Epigenetics and DNA methylation

Within a human individual, multitudes of cells varying drastically in form and function are all derived from a single diploid genome which is passed down from a single embryo to each daughter cell with high-fidelity. Despite sharing nearly identical genomic sequences, these cells distinguish themselves by committing to different lineages and adopting new cellular roles through differential gene expression. This complex process that controls which genes are expressed under what conditions is known as epigenetics.

DNA methylation is one of many molecular marks that contribute to this regulatory system which also includes the post-translational modification of histones and regulatory RNA species. In mammalian genomes, DNA methylation occurs almost exclusively on cytosine residues at CpG dinucleotide sites. Classically, hypermethylation is associated with inaccessible heterochromatin and silencing of local gene expression as is observed female X chromosome inactivation. More recent studies have shown that the autosomal regulatory relationships between DNA methylation and gene expression are much more complex. The distribution of CpGs is non-random across mammalian genomes with genome-wide depletion compared to other dinucleotide motifs, clustering into high density CpG islands, and enrichment the near the majority of promoters [6]. DNA methylation is now thought to play context-dependent roles in

promoter usage, transcriptional elongation, alternative splicing, and long-range regulation by enhancers and insulators [7].

The epigenetic clock

Chronological age has been shown to have a profound effect on DNA methylation levels [8-16]. As a result, several highly accurate epigenetic biomarkers of chronological age have been proposed [17-21]. These biomarkers use weighted averages of methylation levels at specific CpG sites to produce estimates of age (in units of years), referred to as "DNA methylation age" (DNAm age) or "epigenetic age". To facilitate the study of the age-independent aspects of this phenomenon, measures of "epigenetic age acceleration" or "epigenetic aging" have also been developed. These can be thought of as the difference between the chronological age of a sample and the measured age of a sample based on DNA methylation. Positive epigenetic age acceleration indicates that a sample appears older than it should epigenetically and likewise negative epigenetic age acceleration indicates that a sample appears is younger than expected.

Recent studies support the idea that these measures are at least passive biomarkers of biological age. For instance, the epigenetic age of blood has been found to be predictive of all-cause mortality [22-25], frailty [26], lung cancer [27], and cognitive and physical functioning [28], while the blood of the offspring of Italian semi-supercentenarians (i.e. participants aged 105 or older) was shown to have a lower epigenetic age than that of age-matched controls [29]. Further, the utility of the epigenetic clock method using various tissues and organs has been demonstrated in applications surrounding Alzheimer's disease [30], centenarian status [29, 31], development [32], Down syndrome [33], frailty [26], HIV infection [34], Huntington's disease [35], obesity [36], lifetime stress [37], menopause [38], osteoarthritis [39], and Parkinson's disease [40]. Though there has been much progress in discerning the biological significance of

epigenetic aging, it is still unclear what factors control this process and if intervening on epigenetic aging will inhibit and/or reverse biological aging.

Chapter 1: Associations between epigenetic aging and diet, lifestyle, and sociodemographic factors

ABSTRACT

Behavioral and lifestyle factors have been shown to relate to a number of health-related outcomes, yet there is a need for studies that examine their relationship to molecular aging rates. Toward this end, we use recent epigenetic biomarkers of age that have previously been shown to predict all-cause mortality, chronic conditions, and age-related functional decline. We analyze cross-sectional data from 4,173 postmenopausal female participants from the Women's Health Initiative, as well as 402 male and female participants from the Italian cohort study, Invecchiare nel Chianti.

Extrinsic epigenetic age acceleration (EEAA) exhibits significant associations with fish intake ($p=0.02$), moderate alcohol consumption ($p=0.01$), education ($p=3x10^{-5}$), BMI ($p=0.01$), and blood carotenoid levels ($p=1x10^{-5}$)—an indicator of fruit and vegetable consumption, whereas intrinsic epigenetic age acceleration (IEAA) is associated with poultry intake ($p=0.03$) and BMI ($p=0.05$). Both EEAA and IEAA were also found to relate to indicators of metabolic syndrome, which appear to mediate their associations with BMI. Finally, longitudinal data suggests that an increase in BMI is associated with increase in both EEAA and IEAA.

Overall, the epigenetic age analysis of blood confirms the conventional wisdom regarding the benefits of eating a plant-based diet with lean meats, moderate alcohol consumption, physical activity, and education, as well as the health risks of obesity and metabolic syndrome.

INTRODUCTION

A number of behavioral lifestyle factors have been shown to relate to health, including diet, physical activity, moderate alcohol consumption, and educational attainment. For instance,

diet is a modifiable behavior with the potential to mitigate chronic disease risk. Various dietary components have been reported to influence intermediate risk factors and the prevalence of age-related disease outcomes; thus there is a growing consensus regarding nutritional recommendations for maintaining optimal health. These dietary factors include whole grain & dietary fiber [41], fish & omega-3 fatty acids [42], and fruits & vegetables [43], all of which may be involved in reducing systemic inflammation [44]. Further, metabolic health has been established as one of the primary mechanisms through which diet affects health and disease [45]. Conditions such as, insulin resistance, hypercholesterolemia, hypertension, hypertriglyceridemia, and systemic inflammation can be promoted by poor dietary habits and often coalesce, influencing a person's risk of atherosclerosis, diabetes mellitus, and stroke [46-48].

In addition to diet, other behaviors such as moderate alcohol consumption, increased physical activity, and higher educational attainment have all been linked to reductions in morbidity and mortality risk [49-56]. Yet, despite the strong evidence connecting lifestyle factors to health outcomes, it is still unclear whether these factors directly influence aging on a molecular level. In previous work, leukocyte telomere length (LTL) has been used to investigate the influence of lifestyle factors on replicative aging in blood [57-61]. A cross-sectional study of 2,284 participants from the Nurses' Health Study reported that LTL was associated with BMI, waist circumference, and dietary intake of total fat, polyunsaturated fatty acids, and fiber [62]. LTL was also found to be longer among individuals who were more physically active [63, 64], as well has those with higher levels of education [65].

However, relatively little is known about the relationship between epigenetic aging rates and lifestyle factors, such as diet, alcohol consumption, physical activity, and educational

attainment. Here, we investigate these relationships by leveraging blood DNA methylation data from two large epidemiological cohorts. In our primary analysis, we use data from older women within the Women's Health Initiative (**WHI**) to examine the relationships between epigenetic age acceleration in blood and dietary variables, education, alcohol, and exercise. In our secondary analysis, we sought to validate the results in the Invecchiare nel Chianti (**InCHIANTI**) Study, which is a population-based prospective cohort study of residents ages 21 or older from two areas in the Chianti region of Tuscany, Italy.

RESULTS

Sample characteristics

The WHI sample consisted of 4,173 postmenopausal women including 2,045 Caucasians, 1,192 African Americans, and 717 Hispanics. Chronological age ranged from 50-82 years (mean=64, s.d.=7.1). The InCHIANTI sample was composed of 402 participants from a European (Italian) population, including 178 men (44%) and 229 women (56%). We used the most current cross-sectional wave for this cohort, and at that time-point participants ranged in age from 30 to 100 years (mean=71, s.d.=16). Additional details on participant characteristics can be found in the **Methods** and in **Table 1-1**.

Table 1-1. Characteristics of the WHI and InCHIANTI samples. The cohort samples are listed for each column and variables of interest are listed for each row. The upper portion of the table correspond to categorical variables and are described using counts and percentages; the lower portion of the table displays continuous variables which are described using means and standard deviations (SD).

			WHI		InCHIANTI	
			Count	Percent	Count	Percent
Ethnic	American Indian or Alaskan Native		56	1%		
	Asian or Pacific Islander		140	3%		
	Black or African-American		1277	28%		
	Hispanic/Latino		784	17%		
	White (not of Hispanic origin)		2196	49%		
	Other		37	1%		
WHI data set	BA23		2098	47%		
	AS315		2392	53%		
Sex	Male				178	44%
	Female				229	56%
Current smoker	Smoker		4027	90%	367	90%
	Nonsmoker		439	10%	40	10%
Education	< Primary		43	1%	80	20%
	> Primary		154	3%	154	38%
	> Lower secondary		293	7%	91	22%
	> Upper secondary		2588	58%	62	15%
	> Higher		1393	31%	20	5%
Physical activity	active		894	20%	329	81%
	inactive		3572	80%	78	19%
			Mean	**SD**	**Mean**	**SD**
Diet	Total energy, kcal	kcal/day	1641	777	2069	573
	Carbohydrate	% kcal	49	9.1	52.4	6.9
	Protein	% kcal	16.5	3.3	15.8	2
	Fat	% kcal	34.6	8.1	30.9	5.5
	Red meat	serv/day	0.8	0.7	1.1	0.5
	Poultry	serv/day	0.4	0.3	0.2	0.2
	Fish	serv/day	0.3	0.3	0.2	0.2
	Dairy	serv/day	1.6	1.3	2.8	1.8
	Whole grains	serv/day	1.2	0.9		
	Nuts	serv/day	0.2	0.3	0	0.1
	Fruits	serv/day	1.7	1.3	1.9	0.9
	Vegetables	serv/day	1.9	1.3	1.6	0.8
	Alcohol	g/day	3.6	9.6	12.7	14.9
Measurements	C-reactive protein	mg/L	5.2	6.6	3.9	7.4
	Insulin	mg/dL	57.1	115.3		
	Glucose	mg/dL	106.3	38	93	21.3
	Triglycerides	mg/dL	146.4	85.6	122.7	81.5
	Total cholesterol	mg/dL	228.4	42.7	207.2	36.6
	LDL cholesterol	mg/dL	144.9	39.7	125.5	32.1
	HDL cholesterol	mg/dL	54	14.3	57.6	15.7
	Creatinine	mg/dL	0.8	0.2	0.9	0.4
	Systolic blood pressure	mmHg	130	18	129.3	19.8
	Diastolic blood pressure	mmHg	75.8	9.4	77.2	10.3
	Waist / hip ratio	cm/cm	0.8	0.1	0.9	0.1
	BMI	cm/m2	29.7	6	27	4.3

Dietary and metabolic associations with measures of age acceleration

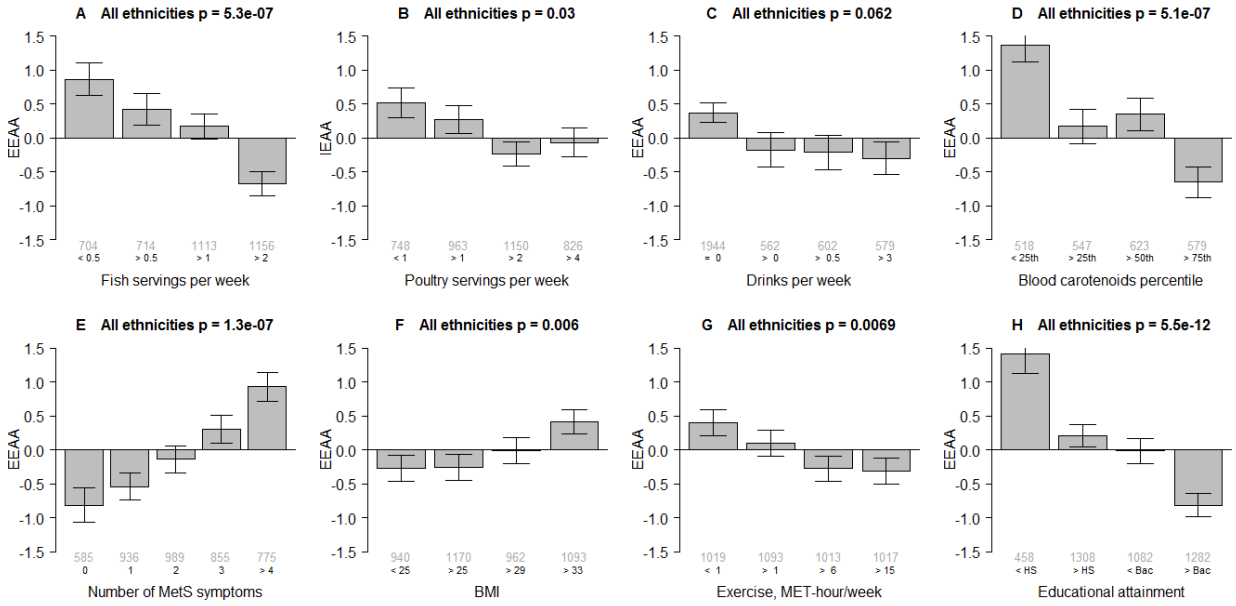
Here we leverage two distinct measures of epigenetic age acceleration which are based on different sets of CpGs: *intrinsic* epigenetic age acceleration (IEAA), and *extrinsic* epigenetic age acceleration (EEAA) (**Methods**). Epigenetic age acceleration is broadly defined as the epigenetic age left unexplained by chronological age, where intrinsic and extrinsic denote additional modifications to this concept. In addition to adjusting for chronologic age, IEAA also adjusts the epigenetic clock for blood cell count estimates, arriving at a measure that is unaffected by both variation in chronologic age and blood cell composition. EEAA, on the other hand, integrates known age-related changes in blood cell counts with a blood-based measure of epigenetic age [19] before adjusting for chronologic age, making EEAA dependent on age-related changes in blood cell composition. In essence, IEAA can be interpreted as a measure of cell-intrinsic aging and EEAA as a measure of immune system aging, where for both, a positive value indicates that the epigenetic age of an individual (organ or tissue) is higher than expected based on their chronological age—or that the individual is exhibiting accelerated epigenetic aging. We find that IEAA is only moderately correlated with EEAA (r=0.37), and that measurements on the same individuals at different time points (mean difference 3.0 years between visit dates) showed moderately strong correlations (IEAA r=0.70, EEAA r=0.66).

We first used a robust correlation test to relate our two measures of epigenetic aging (IEAA and EEAA) to select reported dietary exposures, blood nutrient levels, cardiometabolic plasma biomarkers, and lifestyle factors, designating a Bonferroni-corrected significance threshold of $\alpha=7x10^{-4}$ (**Figure 1-1**); these correlations were found to be consistent within racial strata are presented (results not shown). Select marginal associations are shown as bar plots in **Figure 1-2**.

Figure 1-1. Marginal correlations with epigenetic age acceleration. Correlations (bicor, biweight midcorrelation) between select variables and the two measures of epigenetic age acceleration are colored according to their magnitude with positive correlations in red, negative correlations in blue, and statistical significance (p-values) in green. Blood biomarkers were measured from fasting plasma collected at baseline. Food groups and nutrients are inclusive, including all types and all preparation methods, e.g. folic acid includes synthetic and natural, dairy includes cheese and all types of milk, etc.

		Pooled WHI samples Adjusted for ethnicity and dataset					
		n	μ	IEAA		EEAA	
				bicor	p	bicor	p
Diet	log2(Total energy)	3687	10.53	0.00	0.96	-0.02	0.19
	Carbohydrate	3687	49.01	0.02	0.29	0.00	0.96
	Protein	3687	16.50	-0.02	0.15	-0.03	0.10
	Fat	3687	34.66	0.00	0.97	0.02	0.15
	log2(1+Red meat)	3687	0.75	0.03	0.10	0.02	0.28
	log2(1+Poultry)	3687	0.45	-0.05	4E-3	-0.03	0.05
	log2(1+Fish)	3687	0.31	-0.02	0.30	-0.07	**2E-5**
	log2(1+Dairy)	3687	1.25	0.00	0.99	-0.02	0.29
	log2(1+Whole grains)	3687	1.03	0.00	0.85	-0.02	0.19
	log2(1+Nuts)	3687	0.19	0.01	0.51	-0.02	0.36
	log2(Fruits)	3687	0.32	0.00	0.81	-0.03	0.04
	log2(Vegetables)	3687	0.62	0.00	0.98	-0.04	0.01
Blood nutrients	Retinol	2268	0.59	0.02	0.46	-0.01	0.69
	Mean carotenoids	2267	0.01	-0.06	4E-3	-0.13	**2E-9**
	Lycopene	2268	0.40	-0.02	0.44	-0.03	0.17
	log2(alpha-Carotene)	2268	-4.22	-0.04	0.04	-0.11	**9E-8**
	log2(beta-Carotene)	2267	-2.18	-0.06	0.01	-0.11	**3E-7**
	log2(Lutein+Zeaxanthin)	2268	-2.38	-0.04	0.09	-0.09	**1E-5**
	log2(beta-Cryptoxanthin)	2268	-3.74	-0.06	2E-3	-0.11	**3E-7**
	log2(alpha-Tocopherol)	2268	3.94	-0.04	0.07	-0.06	0.01
	log2(gamma-Tocopherol)	2268	0.68	0.08	**2E-4**	0.09	**9E-6**
Measurements	log2(C-reactive protein)	2809	1.54	0.08	**6E-5**	0.12	**2E-10**
	log2(Insulin)	4043	5.81	0.07	**2E-5**	0.11	**3E-12**
	log2(Glucose)	4145	6.66	0.06	**8E-5**	0.06	**2E-4**
	log2(Triglyceride)	4149	7.05	0.05	**5E-4**	0.07	**6E-6**
	Total cholesterol	4149	227.31	0.03	0.04	0.01	0.62
	LDL cholesterol	4085	142.85	0.03	0.06	0.01	0.41
	HDL cholesterol	4146	54.86	-0.04	0.01	-0.09	**1E-8**
	log2(Creatinine)	2748	-0.42	0.01	0.74	0.02	0.26
	Systolic blood pressure	4165	130.17	0.04	5E-3	0.07	**4E-6**
	Diastolic blood pressure	4165	75.86	0.05	3E-3	0.04	0.01
	log2(Waist / hip ratio)	4165	-0.28	0.05	3E-3	0.09	**2E-8**
	BMI	4165	29.69	0.08	**1E-6**	0.09	**2E-8**
Socio-behavioral	Education	4130	6.80	-0.02	0.14	-0.10	**3E-10**
	Income	4041	3.73	0.00	0.79	-0.06	**1E-4**
	log2(1+Exercise)	4142	2.53	-0.04	0.01	-0.07	**2E-5**
	Current smoker	4142	0.12	0.00	0.78	-0.01	0.66
	log2(1+Alcohol)	3687	1.10	-0.02	0.21	-0.07	**3E-5**

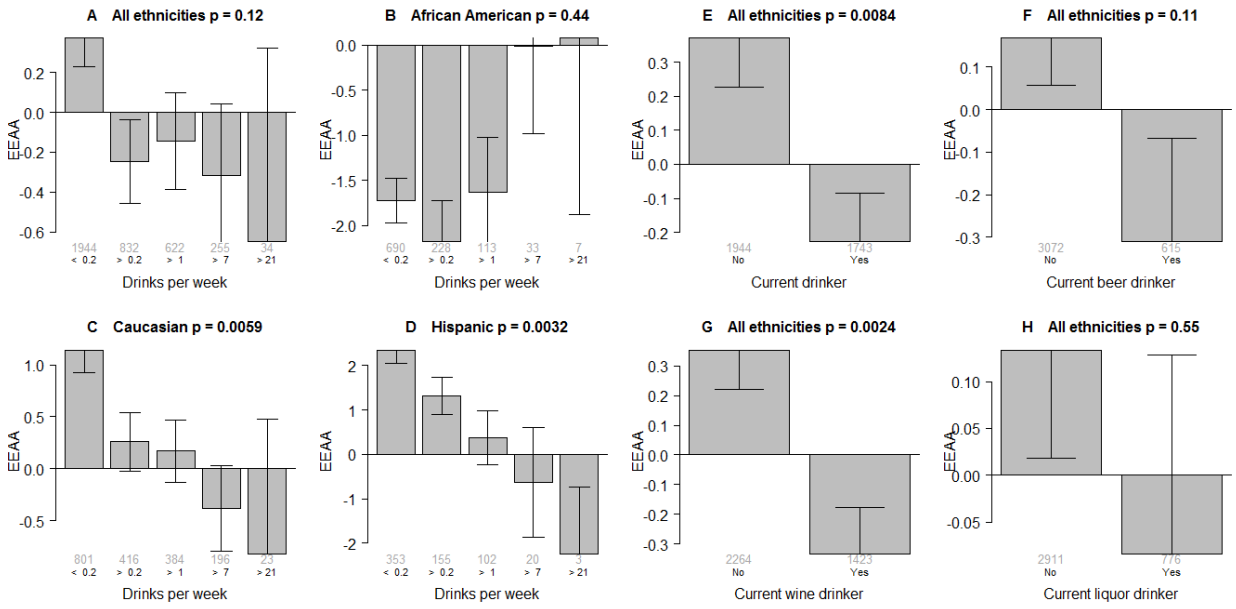
Figure 1-2. EEAA among different levels of select dietary & lifestyle habits. Panels A-H show barplots visualizing the EEAA among stratified levels of fish, poultry (IEAA in this case), alcohol intake, blood carotenoid levels, number of metabolic syndrome symptoms, BMI, exercise, and education. Cut points roughly correspond roughly to quartiles except with number of MetS symptoms and alcohol intake which were selected for evenly-sized strata as much as possible. The sample sizes for each stratum are shown in grey beneath each bar. P-values for differences between strata are listed above each bar plot. Exercise is in units of metabolic equivalent hours per week and education uses high school diploma and bachelor's degree as cut points.



EEAA exhibits weak but statistically significant correlations with fish intake (r=-0.07, $p=2x10^{-5}$), alcohol consumption (r=-0.07, $p=3x10^{-5}$, **Figure 1-3**), plasma levels of mean carotenoids (r=-0.13, $p=2x10^{-9}$), alpha-carotene (r=-0.11, $p=9x10^{-8}$), beta-carotene (r=-0.11, $p=3x10^{-7}$), lutein+zeaxanthin (r=-0.9, $p=1x10^{-5}$), beta-cryptoxanthin (r=-0.11, $p=3x10^{-7}$), gamma-tocopherol (r=0.09, $p=9x10^{-6}$), triglyceride (r=0.7, $p=6x10^{-6}$), C-reactive protein (CRP, r=0.12, $p=2x10^{-10}$), insulin (r=0.11, $p=3x10^{-12}$), HDL cholesterol (r=-0.09, $p=2x10^{-8}$), glucose (r=0.06, $p=2x10^{-4}$), systolic blood pressure (r=0.07, $p=4x10^{-6}$), waist-to-hip ratio (WHR, r=0.09, $p=2x10^{-}$

^{8}), BMI (r=0.09, p=2x10^{-8}), education (r=-0.10, p=3x10^{-10}), income (r=-0.06, p=1x10^{-4}), and exercise (r=-0.07, p=2x10^{-5}, **Figure 1-1**). In contrast, the intrinsic epigenetic aging rate exhibits weaker correlations with dietary variables and lifestyle factors: IEAA is only associated with BMI (r=0.08, p=1x10^{-6}), and plasma levels of gamma-tocopherol (r=0.08, p=2x10^{-4}), CRP (r=0.08, p=6x10^{-5}), insulin (r=0.07, p=2x10^{-5}), glucose (r=0.06, p=8x10^{-5}), and triglyceride levels (r=0.05, p=5x10^{-4}, **Figure 1-1**).

Figure 1-3. EEAA among different strata of ethnic groups, levels and types of alcohol intake. Panels A-D show bar plots visualizing the EEAA among stratified levels of alcohol intake (medium servings per week) for select ethnic groupings. Panels E-H show bar plots visualizing the EEAA among non- and current drinkers (at least one drink per month) of different types of alcoholic drinks: all types, beer, wine, and liquor. The sample sizes for each stratum are shown in grey beneath each bar. P-values for differences between strata are listed above each bar plot.



Meta-analysis of multivariable linear models link epigenetic age acceleration to diet

Associations with EEAA

We have recently shown that ethnicity relates to epigenetic aging rates: e.g. Hispanics have lower levels of IEAA compared to other ethnic groups [66]. Given the potential for confounding by sociodemographic and lifestyle factors, we used Stouffer's method to meta-analyze multivariate linear models, stratified by racial/ethnic group, in order to re-examine the suggestive associations from our marginal correlation analysis. After adjusting for sex and dataset (**Figure 1-4A**), we find that lower EEAA is significantly associated with greater intake of fish (t_{meta}=-2.92, p_{meta}=0.003), higher education (t_{meta}=-4.14, p_{meta}=3x10^{-5}), lower BMI (t_{meta}=4.86, p_{meta}=1x10^{-6}), and current drinker status (t_{meta}=-3.23, p_{meta}=0.001). However, we find no association for current smoking status, and poultry intake, and only a trend toward association with physical activity (t_{meta}=-1.70, p_{meta}=0.09). In the subset of WHI participants with circulating carotenoid measurements, we also find that mean carotenoid levels are associated with EEAA (t_{meta}=-4.34, p_{meta}=1x10^{-5}, **Figure 1-5A**).

Figure 1-4. Meta-analysis of multivariable linear models of EEAA and IEAA. EEAA (panel **A**) and IEAA (panel **B**) were regressed on potential confounding factors, fish and poultry intake, and current drinker status for the ethnic strata with sufficient sample sizes (n>100). Individual columns correspond to coefficient estimates (β) colored blue or red for negative and positive values respectively, and p-values (p) colored in green according to magnitude of significance, with the exception of the last two columns which denote Stouffer's method meta-t and meta-p values. Models are adjusted for originating dataset (WHI BA23, WHI AS315, or InCHIANTI) and for sex.

A

EEAA	WHI								InCHIANTI		Meta-analysis	
	Caucasian		African		Hispanic		Asian					
n	1807		1058		618		100		402		3985	
	β	p	β	p	β	p	β	p	β	p	meta-t	meta-p
log2(1 + Fish)	-0.83	0.19	-1.05	0.12	-1.13	0.24	-0.55	0.78	-3.61	0.02	-2.92	3E-3
log2(1 + Poultry)	-0.72	0.20	0.41	0.51	-0.15	0.83	2.06	0.29	-1.35	0.35	-0.73	0.46
Current drinker	-0.66	0.02	-0.01	0.99	-1.58	9E-4	-2.13	0.07	-0.11	0.87	-3.23	1E-3
Education	-0.19	0.01	-0.20	0.05	-0.23	0.03	-0.03	0.91	-0.13	0.06	-4.14	3E-5
BMI	0.09	5E-4	0.10	3E-3	0.07	0.10	0.17	0.13	0.02	0.82	4.86	1E-6
Physically active	-0.61	0.09	-0.11	0.82	-0.09	0.87	-2.69	0.08	-0.23	0.78	-1.70	0.09
Current smoker	0.19	0.66	1.09	0.06	-0.80	0.23	-0.25	0.93	-2.34	0.02	0.07	0.94

B

IEAA	WHI								InCHIANTI		Meta-analysis	
	Caucasian		African		Hispanic		Asian					
n	1807		1058		618		100		402		3985	
	β	p	β	p	β	p	β	p	β	p	meta-t	meta-p
log2(1 + Fish)	0.61	0.24	-0.85	0.12	0.38	0.60	1.03	0.53	-2.51	0.03	-0.38	0.71
log2(1 + Poultry)	-1.20	0.01	-0.15	0.76	-0.42	0.43	-2.28	0.17	-2.86	0.01	-3.30	1E-3
Current drinker	-0.09	0.70	-0.11	0.75	-0.43	0.23	1.04	0.29	0.61	0.25	-0.36	0.72
Education	0.01	0.84	-0.10	0.21	0.03	0.73	-0.34	0.16	-0.01	0.78	-0.68	0.49
BMI	0.08	3E-5	0.04	0.17	0.02	0.60	0.35	5E-4	-0.03	0.56	4.14	4E-5
Physically active	0.06	0.84	-0.40	0.28	-0.15	0.71	-2.22	0.09	-0.43	0.48	-1.06	0.29
Current smoker	0.07	0.83	0.23	0.62	0.30	0.56	0.01	1.00	-1.25	0.11	0.12	0.91

Figure 1-5. Meta-analysis of linear models of EEAA and IEAA including carotenoid levels. Analogous to **Figure 1-4** except including mean carotenoid levels: EEAA (panel **A**) and IEAA (panel **B**) were regressed on potential confounding factors, fish and poultry intake, mean across standardized measures of carotenoids, and current drinker status for the ethnic strata with sufficient sample sizes (n>100). Individual columns correspond to coefficient estimates (β) colored blue or red for negative and positive values respectively, and p-values (p) colored in green according to magnitude of significance, with the exception of the last two columns which denote Stouffer's method meta-t and meta-p values. Models are adjusted for originating dataset (WHI BA23 or WHI AS315).

A

	WHI								Meta-analysis	
EEAA	Caucasian		African		Hispanic		Asian			
n	886		481		259		100		886	
	β	p	β	p	β	p	β	p	meta-t	meta-p
log2(1 + Fish)	-1.18	0.18	-0.99	0.28	-0.91	0.50	-0.67	0.73	-1.88	0.06
log2(1 + Poultry)	0.03	0.97	-0.03	0.98	0.47	0.65	1.82	0.35	0.41	0.68
Mean carotenoids	-0.98	**2E-3**	-0.93	**0.02**	-0.68	0.17	-1.25	0.16	-4.34	**1E-5**
Current drinker	-0.80	**0.04**	0.17	0.77	-0.95	0.18	-2.36	**0.04**	-2.31	**0.02**
Education	-0.05	0.62	-0.17	0.22	-0.47	**2E-3**	0.04	0.89	-2.20	**0.03**
BMI	0.03	0.43	0.14	**2E-3**	0.07	0.28	0.11	0.37	2.85	**4E-3**
Physically active	-0.54	0.28	0.47	0.44	1.06	0.19	-2.49	0.11	-0.24	0.81
Current smoker	1.15	0.06	0.75	0.35	-2.29	**0.02**	-0.82	0.76	0.87	0.38

B

	WHI								Meta-analysis	
IEAA	Caucasian		African		Hispanic		Asian			
n	886		481		259		100		886	
	β	p	β	p	β	p	β	p	meta-t	meta-p
log2(1 + Fish)	0.20	0.78	-0.90	0.24	-0.21	0.85	1.05	0.52	-0.35	0.73
log2(1 + Poultry)	-0.50	0.43	0.13	0.85	-0.84	0.30	-2.23	0.18	-1.20	0.23
Mean carotenoids	-0.62	**0.02**	-0.43	0.20	-0.14	0.71	0.24	0.75	-2.47	**0.01**
Current drinker	-0.27	0.40	0.42	0.38	-0.42	0.46	1.09	0.27	-0.16	0.87
Education	-0.02	0.80	-0.22	0.06	-0.06	0.63	-0.35	0.15	-1.72	0.09
BMI	0.04	0.22	0.05	0.19	0.04	0.41	0.36	**8E-4**	2.72	**0.01**
Physically active	-0.10	0.82	0.05	0.92	0.91	0.16	-2.26	0.09	0.02	0.99
Current smoker	-0.17	0.74	-0.70	0.30	-0.55	0.48	0.13	0.96	-1.05	0.29

Multivariate linear models were used to examine whether variations in cardiometabolic biomarkers and/or the number of symptoms for metabolic syndrome accounted for any of the associations between EEAA and lifestyle factors. The inclusion of biomarkers in an unstratified model shows that EEAA positively relates to CRP (log2, β=0.31, $p=3x10^{-4}$, **Figure 1-6A**, model 3) and that this is accompanied by a concomitant diminishing in the effect size of BMI (67% decrease in coefficient magnitude, **Figure 1-6A**, model 2 vs. model 5), suggesting that higher CRP may partially explain the positive association between BMI and EEAA. When metabolic syndrome (**MetS**) was included in the model, results showed that higher EEAA is positively associated with the number of metabolic syndrome symptoms (β=0.29, p=0.002, **Figure 1-6A**, model 4). In the subset of participants with both biomarker and carotenoid measurements, EEAA was negatively associated with mean carotenoid levels (β=-1.10, $p=1x10^{-4}$) while appearing to diminish associations with biomarkers (**Figure 1-7A**, model 5).

Figure 1-6. Multivariate linear models of EEAA and IEAA with and without biomarkers. EEAA (panel **A**) and IEAA (panel **B**) were regressed on potential confounding factors, fish and poultry intake and current drinker status, and select biomarkers. Individual columns list the corresponding coefficient estimates (β) and p-values (p) for each fitting. Coefficients are colored according to sign (positive = red, negative = blue) and significance according to magnitude (green). Models 1 through 5 correspond to a minimal model, a model including dietary intake variables, a model including potential explanatory biomarkers, a model including number of metabolic syndrome symptoms and a complete model with all of the variables above, respectively. Models are adjusted for originating dataset (WHI BA23, or WHI AS315).

A

EEAA	Model 1 *Minimal*		Model 2 *Food*		Model 3 *Biomarkers*		Model 4 *MetS*		Model 5 *Full*	
n=2725	β	p	β	p	β	p	β	p	β	p
log2(1 + Fish)			-1.23	0.01					-1.13	0.02
log2(1 + Poultry)			-0.03	0.94					-0.05	0.90
Current drinker			-0.63	0.01					-0.54	0.03
Education	-0.27	7E-6	-0.22	2E-4	-0.24	5E-5	-0.25	2E-5	-0.20	8E-4
BMI	0.09	1E-5	0.09	2E-5	0.02	0.46	0.06	0.01	0.03	0.26
Physically active	-0.54	0.06	-0.44	0.13	-0.41	0.15	-0.49	0.08	-0.33	0.25
Current smoker	0.30	0.38	0.33	0.32	0.23	0.50	0.26	0.43	0.28	0.41
African American	-2.31	1E-17	-2.32	5E-17	-2.34	6E-16	-2.22	3E-16	-2.36	1E-15
Hispanic	1.07	8E-4	0.98	2E-3	0.93	4E-3	1.11	5E-4	0.83	0.01
log2(C-reactive protein)					0.31	3E-4			0.31	3E-4
log2(Insulin)					0.19	0.25			0.19	0.24
log2(Triglycerides)					0.25	0.24			0.34	0.14
log2(Glucose)					0.34	0.32			0.41	0.27
HDL Cholesterol					-0.01	0.33			-0.01	0.28
Systolic blood pressure					0.01	0.13			0.01	0.08
Diastolic blood pressure					0.00	0.95			0.00	0.75
log2(Waist-to-hip ratio)					0.25	0.80			0.48	0.63
Metabolic syndrome symptoms							0.29	2E-3	-0.16	0.30

B

IEAA	Model 1 *Minimal*		Model 2 *Food*		Model 3 *Biomarkers*		Model 4 *MetS*		Model 5 *Full*	
n=2725	β	p	β	p	β	p	β	p	β	p
log2(1 + Fish)			-0.04	0.92					0.01	0.99
log2(1 + Poultry)			-0.70	0.03					-0.71	0.03
Current drinker			-0.07	0.72					0.00	0.99
Education	-0.03	0.58	-0.02	0.72	-0.01	0.82	-0.01	0.75	-0.01	0.90
BMI	0.06	2E-4	0.06	9E-5	0.02	0.19	0.03	0.07	0.03	0.20
Physically active	-0.24	0.30	-0.23	0.31	-0.16	0.49	-0.19	0.39	-0.16	0.48
Current smoker	0.26	0.33	0.26	0.34	0.26	0.33	0.23	0.38	0.25	0.35
African American	-0.56	0.01	-0.50	0.02	-0.51	0.03	-0.48	0.03	-0.45	0.05
Hispanic	-1.51	3E-9	-1.47	1E-8	-1.60	6E-10	-1.47	8E-9	-1.54	3E-9
log2(C-reactive protein)					0.10	0.13			0.11	0.12
log2(Insulin)					0.13	0.31			0.13	0.34
log2(Triglycerides)					0.40	0.02			0.35	0.06
log2(Glucose)					0.35	0.20			0.31	0.30
HDL Cholesterol					0.00	0.89			0.00	0.72
Systolic blood pressure					0.00	0.67			0.00	0.93
Diastolic blood pressure					0.01	0.42			0.01	0.39
log2(Waist-to-hip ratio)					-0.17	0.83			-0.24	0.77
Metabolic syndrome symptoms							0.27	4E-4	0.07	0.58

Figure 1-7. Multivariate linear models of EEAA and IEAA including carotenoid levels. Analogous to **Figure 1-6** except including mean carotenoid levels: EEAA (panel A) and IEAA (panel B) were regressed on potential confounding factors, fish and poultry intake and current drinker status, and select biomarkers. Individual columns list the corresponding coefficient estimates (β) and p-values (p) for each fitting. Coefficients are colored according to sign (positive = red, negative = blue) and significance according to magnitude (green). Models 1 through 5 correspond to a minimal model, a model including dietary intake variables, a model including potential explanatory biomarkers, a model including number of metabolic syndrome symptoms and a complete model with all of the variables above, respectively. Models are adjusted for originating dataset (WHI BA23 or WHI AS315).

A

EEAA	Model 1 *Minimal*		Model 2 *Food*		Model 3 *Biomarkers*		Model 4 *MetS*		Model 5 *Full*	
n=922	β	p	β	p	β	p	β	p	β	p
log2(1 + Fish)			-1.45	0.05					-1.41	0.06
log2(1 + Poultry)			0.55	0.39					0.54	0.39
Mean carotenoids			-1.25	**5E-6**					-1.10	**1E-4**
Current drinker			-0.60	0.12					-0.52	0.18
Education	-0.33	**3E-4**	-0.24	**0.01**	-0.30	**1E-3**	-0.32	**5E-4**	-0.23	**0.02**
BMI	0.12	**9E-5**	0.07	**0.03**	0.03	0.36	0.09	**0.01**	0.02	0.55
Physically active	-0.17	0.71	0.14	0.76	-0.04	0.93	-0.17	0.71	0.21	0.64
Current smoker	0.65	0.23	0.26	0.63	0.50	0.36	0.61	0.26	0.25	0.65
African American	-3.22	**3E-14**	-3.13	**3E-13**	-3.43	**2E-14**	-3.19	**5E-14**	-3.29	**6E-13**
Hispanic	0.24	0.64	0.34	0.51	-0.05	0.92	0.26	0.62	0.09	0.86
log2(C-reactive protein)					0.32	**0.02**			0.24	0.08
log2(Insulin)					0.54	**0.03**			0.45	0.08
log2(Triglycerides)					-0.18	0.58			-0.01	0.97
log2(Glucose)					-0.75	0.22			-0.50	0.44
HDL Cholesterol					-0.03	0.06			-0.02	0.17
Systolic blood pressure					0.02	0.09			0.02	0.06
Diastolic blood pressure					-0.01	0.58			-0.01	0.55
log2(Waist-to-hip ratio)					-0.77	0.61			-1.22	0.43
Metabolic syndrome symptoms							0.27	0.07	-0.13	0.62

B

IEAA	Model 1 *Minimal*		Model 2 *Food*		Model 3 *Biomarkers*		Model 4 *MetS*		Model 5 *Full*	
n=922	β	p	β	p	β	p	β	p	β	p
log2(1 + Fish)			-0.34	0.57					-0.22	0.71
log2(1 + Poultry)			-0.59	0.25					-0.60	0.23
Mean carotenoids			-0.41	0.06					-0.40	0.07
Current drinker			0.17	0.58					0.16	0.59
Education	-0.15	**0.04**	-0.13	0.07	-0.13	0.08	-0.14	**0.05**	-0.12	0.12
BMI	0.08	**1E-3**	0.07	**0.01**	0.05	0.10	0.06	**0.02**	0.06	0.05
Physically active	-0.16	0.65	-0.06	0.87	-0.13	0.71	-0.16	0.65	-0.04	0.91
Current smoker	-0.16	0.72	-0.32	0.46	-0.13	0.77	-0.18	0.68	-0.23	0.61
African American	-0.80	**0.02**	-0.64	0.06	-0.69	**0.05**	-0.78	**0.02**	-0.52	0.15
Hispanic	-1.48	**4E-4**	-1.32	**2E-3**	-1.54	**3E-4**	-1.47	**4E-4**	-1.42	**1E-3**
log2(C-reactive protein)					0.08	0.46			0.04	0.70
log2(Insulin)					0.25	0.22			0.26	0.21
log2(Triglycerides)					0.54	**0.04**			0.76	**0.01**
log2(Glucose)					-0.39	0.42			-0.08	0.88
HDL Cholesterol					0.01	0.53			0.00	0.83
Systolic blood pressure					0.01	0.15			0.02	**0.05**
Diastolic blood pressure					-0.02	0.42			-0.01	0.46
log2(Waist-to-hip ratio)					0.19	0.88			0.38	0.76
Metabolic syndrome symptoms							0.15	0.21	-0.30	0.14

Additionally, we find that for the small subset of individuals for whom we have EEAA measurements at two time points (n=239, mean time interval = 2.7 years), increase in BMI (β=0.40, p=0.002) but not initial BMI (β=-0.01, p=0.81) is significantly associated with increased EEAA (higher follow-up EEAA after adjusting for the initial EEAA, dataset, and ethnicity).

Associations with IEAA

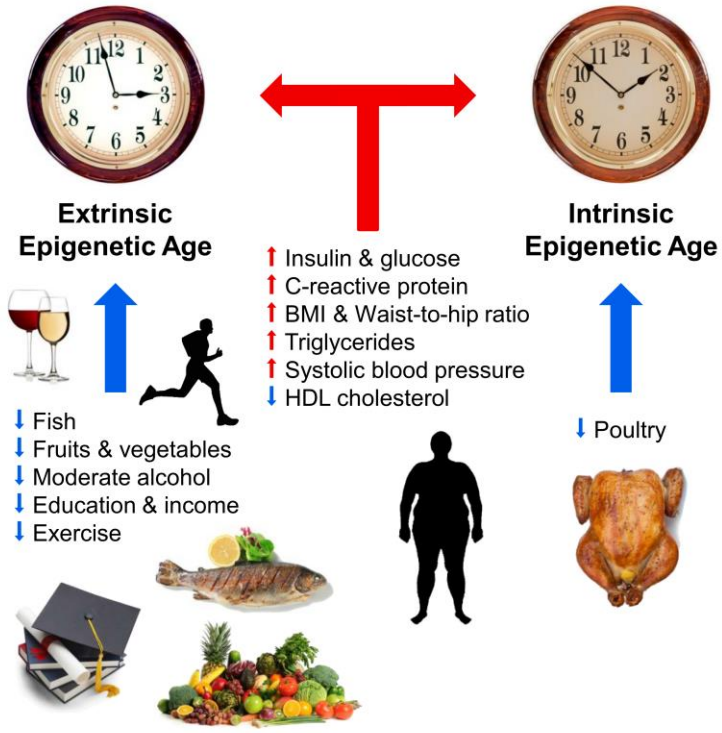
We conducted an analogous meta-analysis of ethnically-stratified linear models of IEAA and found that lower IEAA was significantly associated with poultry intake (t_{meta}=-3.30, p_{meta}=0.001) and lower BMI (t_{meta}=4.14, p_{meta}=4x10^{-5}), after adjusting for potential confounders (**Figure 1-4B**). In the subset of participants with measured carotenoids, IEAA was significantly associated with mean carotenoid levels (t_{meta}=-2.47, p_{meta}=0.01, **Figure 1-5B**). When regressed on clinical biomarkers IEAA was significantly associated with triglycerides (log2, β=0.40, p=0.02, **Figure 1-6B**, model 3). Their inclusion diminished the association between IEAA and BMI (60% decrease in coefficient magnitude, **Figure 1-6B**, model 2 vs. model 5). Number of metabolic syndrome symptoms was also significantly associated with IEAA (β=0.27, p=4x10^{-4}, **Figure 1-6B**, model 4), and diminished the association between IEAA and BMI by 50%. In the subset of WHI participants with circulating carotenoid measurements, we find a trend toward

association between IEAA and mean carotenoid levels (β=-0.40, p=0.07, **Figure 1-7B**, model 5). Finally, in the participants with epigenetic profiling at two time points, increase in BMI (β=0.22, p=0.03) but not initial BMI (β=-0.22, p=0.44) is significantly associated with increased IEAA (higher follow-up IEAA after adjusting for the initial IEAA, dataset, and ethnicity).

DISCUSSION

To our knowledge, this is the first study to examine associations between lifestyle factors and measures of epigenetic age acceleration in blood. Our main findings are summarized graphically in **Figure 1-8**. Overall, our dietary results are consistent with some of the current Dietary Guidelines for Americans [67, 68], reflecting potential health benefits associated with higher intake of fish, poultry, and fruits and vegetables. The weak correlations between dietary factors and epigenetic aging rates probably reflect that a relatively large proportion of the variance in aging rates (around 40 percent) is explained by genetic factors [20, 66, 69]. We find that education, physical activity, low body mass index are associated with a slow extrinsic age acceleration both in univariate correlation tests (**Figure 1-1**) and in multivariate regression models (**Figure 1-4A to 1-7A**). However, consistent with our previous work, smoking status was not associated with epigenetic age acceleration [27], which highlights that not every poor lifestyle choice is associated with an increased epigenetic aging effect in blood tissue.

Figure 4. Pictorial summary of our main findings. The blue and red arrows depict anti-aging and pro-aging effects in blood respectively. The two clocks symbolize the extrinsic epigenetic clock (enhanced version of the Hannum estimate) and the intrinsic epigenetic clock (Horvath 2013) which are dependent and independent of blood cell counts, respectively.



EEAA, inflammation, and metabolic functioning

The age-related changes in immune functioning and inflammation are believed to contribute to increased susceptibility of a wide range of diseases later in life, including diabetes, some cancers, cardiovascular, neurodegenerative, auto-immune, and infectious diseases [70, 71]. In our analysis, EEAA, a biomarker which explicitly incorporates aspects of immune system aging such as age-related changes in blood cell counts, was associated with cardiometabolic biomarkers, fish, fruit, vegetable, and alcohol intake.

Our finding that fish intake was negatively associated with EEAA is consistent with prospective studies suggesting that fish consumption is protective against various age-related diseases [72-74]. The benefits of fish intake may be mediated in part through the omega-3 fatty acids, eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), which stimulate the synthesis of anti-inflammatory cytokines [75]. This is further supported by our finding that

CRP—a well-known marker of inflammation—was the most significant explanatory biomarker of EEAA. This suggests that one reason higher fish consumption may lower EEAA is because it has beneficial anti-inflammatory or metabolic effects. The consensus between these associations also appears to converge on MetS as a potential mediating factor; this was further supported by our results showing that the number of MetS characteristics significantly relates to EEAA. Though CRP is not included in most MetS diagnostic criteria, the association between the two has been previously established [76].

We also find that alcohol consumption was negatively associated with EEAA even after adjusting for potential confounders such as socioeconomic status; this is consistent with prospective studies which have identified light to moderate alcohol intake as a protective factor against all-cause and CHD-related mortality [77, 78] and is supported by a recent publication that also found an association between epigenetic age and alcohol intake in Caucasian and African American individuals (n=656, n=180, respectively) [79]. In our study, we find that the potential benefits of alcohol consumption are observed using a threshold of more than one serving per month, though the effect size of this variable was also stable when adding weekly and daily intake levels (**Figure 1-3A-D**). The association appears to be driven by wine consumption though there is also a trend towards association with beer (**Figures 1-3E-H**). This is consistent with other studies have suggested that wine may have added benefits compared to light alcohol consumption [80]. This finding may also be related to the anti-inflammatory effects of light alcohol consumption, which are associated with decreased circulating levels of inflammatory markers such as IL-6 and CRP [81]. Alternatively, this may be the result of reverse causation, whereby individuals suffering from health issues abstain from alcohol

consumption [82], though interventional studies support a causal protective effect of moderated alcohol intake on cardiovascular blood biomarkers [83].

Though EEAA only trended toward association with reported fruit and vegetable intake, we find significant associations with blood carotenoid levels, which are quantitative surrogates of fruit and vegetable intake; this is likely a reflection of the bias and inaccuracy of self-reported diet. This is in agreement with the wide range of literature supporting the protective effects of high fruit and vegetable intake against age-related diseases CHD [84, 85], stroke [86], type-2 diabetes [87], breast cancer [88], and all-cause mortality [89]. The association between fruit and vegetables with aging of the blood immune system may be partially mediated by anti-inflammatory [90, 91] and cardiometabolic effects, however it is interesting to note that the explanatory power of mean carotenoid levels remained even after including the other explanatory factors into the model, suggesting the possibility of independent anti-aging mechanisms (**Figure 1-7A**, model 5).

Our results for EEAA also share similarities with previously reported findings showing that LTL relates to BMI [62], metabolic factors, vegetable consumption [92], and dietary intake of foods high in omega-3 fatty acids [93]. This agreement is likely a reflection of the shared immunological basis, which is supported by the weak negative correlation between EEAA and age-adjusted LTL. In contrast, IEAA is not significantly associated with LTL, supporting the idea that these measures represent different aspects of aging.

Intrinsic epigenetic aging and metabolic health

Results showed that BMI has a positive association with IEAA (**Figure 1-6B**, model 2). The statistically significant but weak correlations between BMI and epigenetic age acceleration in blood ($r<0.10$) are much smaller than those we recently reported for human liver ($r=0.42$)

[36], suggesting that associations between aging signatures and risk factors may vary in strength depending on the tissue, and may be stronger in organs/tissues most affected by the risk. Interestingly, IEAA was also associated with number of metabolic syndrome characteristics, suggesting a role in tracking metabolic aging processes (**Figure 1-6B**, model 4).

We did find that reported poultry intake was negatively associated with IEAA, even after adjusting for potential confounders and explanatory factors (**Figure 1-6B**, model 5). Given the relative inert behavior of IEAA, the mechanism by which poultry may affect aging is unclear.

Generalization to the InCHIANTI

Our results from the InCHIANTI show some validation of our findings from the WHI: fish intake was related to EEAA, and poultry was related to IEAA. Associations with available biomarkers of cardiometabolic health, however, were not found to be validated, and in a few cases were reversed in directions, within the InCHIANTI (data not shown). The discrepancies between the WHI and InCHIANTI cohorts may be due to numerous differences in the study population (cultural, demographic, genetic, health status, **Table 1-1**) and data collection methodology (dietary assessment). Despite being younger, on average, US participants from the WHI had higher body mass indexes (BMI), and worse metabolic health than their Italian counterparts—as indicated by their greater prevalence of metabolic syndrome (23% in the WHI versus 7.6% in the InCHIANTI).

The InCHIANTI study is also arguably underpowered (n=402) when it comes to detecting the weak associations with epigenetic age acceleration. According to sample size calculations (*PASS* software), we find that n=1820 samples are needed to provide 80% power to detect a correlation of r=0.08 at a two-sided significance level of α=0.01. Similarly, n=1163 samples are needed to detect a correlation of r=0.10.

Limitations

While our study of the WHI benefits from having a relatively large sample size, associations with epigenetic aging may not be detectable in smaller studies given the weak effect sizes observed here. This situation is exacerbated by self-reported lifestyle habits which are notorious for bias and inaccuracy, limiting their ability to represent true lifestyle habits and potentially producing false negative results. Further, as evidenced by our results from the ethnic strata and the InCHIANTI, studies conducted in different ethnic populations may not be entirely consistent due to fundamental differences in age, diet, culture, and demographics. There were several potential limitations to this study, which include the assumption of non-confounding from unmeasured variables such as existing patient co-morbidities and the assumption of accuracy and long-term consistency of reported dietary habits. This is the first longitudinal study to show that an increase in BMI is associated with an increase in epigenetic age acceleration but larger longitudinal studies will be needed to dissect causal relationships between epigenetic aging rates and dietary measures, education, exercise, and lifestyle factors.

Conclusions about epigenetic age acceleration

Our large sample size (n>4500) provides sufficient statistical power for one of our main conclusions: diet has only a weak effect on epigenetic aging rates in blood. These findings will be valuable for researchers who plan to use epigenetic biomarkers in dietary intervention studies. The wide range of associations found with EEAA suggest that immune system aging may be closely linked to conventional notions of metabolic health and may be sensitive to variations in environment and lifestyle. In contrast, IEAA has few associations, which is consistent with the hypothesis that cell-intrinsic aging remains relatively stable, more likely being determined by an intrinsic aging or developmental process under genetic control. Further, using longitudinal data

in the WHI, we found that change in both EEAA and IEAA are significantly associated with change in BMI, suggesting that both modes of epigenetic aging may respond to changes in lifestyle, at least with respect to change in obesity. Overall, our results are consistent with previous literature supporting the protective effects of fish, poultry, & alcohol consumption, exercise, education, as well as the risk of obesity and dyslipidemia.

METHODS

Estimation of DNA methylation age

DNAm age (also referred to as epigenetic age) was calculated from human samples profiled with the Illumina Infinium 450K platform, described in detail in [20]. Briefly, the epigenetic clock is defined as a prediction method of age based on the DNAm levels of 353 CpGs. Predicted age, referred to as DNAm age, correlates with chronological age in sorted cell types (CD4+ T cells, monocytes, B cells, glial cells, neurons), tissues, and organs, including: whole blood, brain, breast, kidney, liver, lung, saliva [20]. We also applied the Hannum measure of DNAm age based on 71 CpGs which was developed using DNA methylation data from blood [19]. Despite high correlations, DNAm age estimates can deviate substantially from chronological age at the individual level, and adjusting for age we can arrive at measures of epigenetic age acceleration as described in the following.

Estimation of Intrinsic and Extrinsic Epigenetic Age Acceleration (IEAA, EEAA)

In this article, we consider two measures of epigenetic age acceleration. These measures, referred to as intrinsic and extrinsic age acceleration only apply to blood. IEAA is derived from the Horvath measure of DNAm age based on 353 CpGs [20], and is defined as the residual resulting from regressing Horvath DNAm age on chronological age and estimates of plasmablasts, naive and exhausted CD8+ T cells, CD4+ T cells, natural killer cells, monocytes,

and granulocytes. Thus, IEAA is independent of chronological age and most of the variation in blood cell composition. IEAA is meant to capture cell-intrinsic properties of the aging process that exhibits preservation across various cell types and organs.

EEAA can be interpreted as an enhanced version of the Hannum measure of DNAm age based on 71 CpGs [19]. EEAA up-weights the contributions of age related blood cell counts [25]. Specifically, EEAA is defined using the following three steps. First, we calculated the epigenetic age measure from Hannum et al, which already correlated with certain blood cell types [22]. Second, we increased the contribution of immune blood cell types to the age estimate by forming a weighted average of Hannum's estimate with 3 cell types that are known to change with age: naïve (CD45RA+CCR7+) cytotoxic T cells, exhausted (CD28-CD45RA-) cytotoxic T cells, and plasmablasts using the Klemera Doubal approach [94]. The weights used in the weighted average are determined by the correlation between the respective variable and chronological age. The weights were chosen on the basis of the WHI data and the same (static) weights were used for all data sets. Finally, EEAA was defined as the residual variation resulting from a univariate model regressing the resulting age estimate on chronological age. Thus, EEAA tracks both age related changes in blood cell composition and intrinsic epigenetic changes.

In a recent large scale meta-analysis involving over 13 thousand subjects from 13 cohorts, we have shown that both IEAA and EEAA are predictive of mortality, independent of chronological age, even after adjusting for additional risk factors, and within the racial/ethnic groups that we examined (Caucasians, Hispanics, African Americans) [25].

IEAA and EEAA can be obtained from the online DNAm age calculator (http://labs.genetics.ucla.edu/horvath/dnamage/), where they are denoted as *AAHOAdjCellCounts* and *BioAge4HAStaticAdjAge*, respectively.

Dietary assessment in the Women's Health Initiative (WHI)

Participants were selected from the WHI, a national study that began in 1993 and enrolled postmenopausal women between the ages of 50-79 years into either randomized clinical trials (RCTs) or into an observational study [95]. Participants completed self-administered questionnaires at baseline which provided personal information on a wide range of topics, including sociodemographic information (age, education, race, income), and current health behaviors (recreational physical activity, tobacco and alcohol exposure, and diet). Participants also visited clinics at baseline where certified Clinical Center staff collected blood specimens and performed anthropometric measurements including weight, height, hip and waist circumferences, and systolic and diastolic blood pressures; body mass index and waist to hip ratio were calculated from these measurements (**Table 1-1**).

Dietary intake levels were assessed at baseline using the WHI Food Frequency Questionnaire [96]. Briefly, participants were asked to report on dietary habits in the past three months, including intake, frequency, and portion sizes of foods or food groups, along with questions concerning topics such as food preparation practices and types of added fats. Nutrient intake levels were then estimated from these responses. For current drinker, we use the threshold of more than one serving equivalent (14g) within the last 28 days.

Estimation of blood cell counts based on DNA methylation levels

We estimate blood cell counts using two different software tools. First, Houseman's estimation method [97], which is based on DNA methylation signatures from purified leukocyte samples, was used to estimate the proportions of CD8+ T cells, CD4+ T, natural killer, B cells, and granulocytes (also known as polymorphonuclear leukocytes). Second, the advanced analysis option of the epigenetic clock software [20, 34] was used to estimate the percentage of

exhausted CD8+ T cells (defined as CD28-CD45RA-) and the number (count) of naïve CD8+ T cells (defined as CD45RA+CCR7+). We and others have shown that the estimated blood cell counts have moderately high correlations with corresponding flow cytometric measures [97, 98]. For example, flow cytometric measurements from the MACS study correlate strongly with DNA methylation based estimates: r=0.63 for CD8+T cells, r=0.77 for CD4+ T cells, r=0.67 for B cell, r=0.68 for naïve CD8+ T cell, r=0.86 for naïve CD4+ T, and r=0.49 for exhausted CD8+ T cells [98].

Blood biomarkers and DNA methylation in the WHI

Two separate subsamples were aggregated for our study within the WHI (BA23 and AS315). Both had baseline blood specimens collected after an overnight fast in EDTA tubes and stored at -70C. These samples were processed at the WHI core laboratory and select nutrient and cardiovascular biomarkers were measured including lycopene, alpha- & beta-carotene, alpha- & gamma-tocopherol, C-reactive protein, triglycerides, total, LDL, and HDL cholesterol.

For the first subsample (BA23) consisting of 2098 samples, DNA methylation levels were measured using the Illumina Infinium HumanMethylation450 BeadChip at the HudsonAlpha Institute of Biotechnology. This platform uses bisulfite conversion to quantify methylation levels at 485,577 specific CpG sites genome-wide. Samples were prepared according to the standard Illumina protocol, and β methylation values were calculated from the intensity ratio between methylated and total (methylated and unmethylated) probe fluorescence intensities. Methylation data was processed as described in [20]. In order to test the quality of these array measurements, we perform correlation measures with duplicates within this dataset and with a "gold" standard which is an average of many samples previously collected.

Correlation between duplicates and with the gold standard were high (r>0.9), indicative of high quality measurements. The second WHI data set is described in the following.

WHI-EMPC Description

The Women's Health Initiative – Epigenetic Mechanisms of PM-Mediated CVD (WHI-EMPC, AS315) is an ancillary study of epigenetic mechanisms underlying associations between ambient particulate matter (PM) air pollution and cardiovascular disease (CVD) in the Women's Health Initiative clinical trials (CT) cohort. The WHI-EMPC study population is a stratified, random sample of 2,200 WHI CT participants who were examined between 1993 and 2001; had available buffy coat, core analytes, electrocardiograms, and ambient concentrations of PM; but were not taking anti-arrhythmic medications at the time. As such, WHI-EMPC is representative of the larger, multiethnic WHI CT population from which it was sampled: n = 68,132 participants aged 50-79 years who were randomized to hormone therapy, calcium/vitamin D supplementation, and / or dietary modification in 40 U.S. clinical centers at the baseline exam (1993-1998) and re-examined in the fasting state one, three, six, and nine years later [99].

Illumina Infinium HumanMethylation450 BeadChip data from the Northwestern University Genomics Core Facility for WHI-EMPC participants sampled in stages 1a (800 participants), 1b (1200 participants), and 2 (200 participants x 2 samples each) was quality controlled and batch adjusted. Batch adjustment involved applying empirical Bayes methods of adjusting for stage and plate as implemented in ComBat [100].

Dietary assessment in the Invecchiare nel Chianti (InCHIANTI)

The InCHIANTI Study is a population-based prospective cohort study of residents ages 30 or older from two areas in the Chianti region of Tuscany, Italy. Data on demographic and lifestyle factors such as smoking, years of education, BMI, and physical activity were collected

during the baseline interview. Physical activity in the previous year was categorized as sedentary or active. Smoking was categorized into current smoker versus former or non-smokers (**Table 1-1**).

In the InCHIANTI study, dietary intake for the past year was assessed using a 236 item food frequency questionnaire (FFQ) for the European Prospective Investigation on Cancer and nutrition (EPIC) study, previously validated in the InCHIANTI population [101]. The FFQ was administered by a trained interviewer and collected information on how frequently (weekly, monthly, yearly) each specific food was generally consumed. Participants were asked to specify the size of the portion usually consumed, in comparison to a range of portion that are shown in colored photographs. Nutrient data for specific foods were obtained from the Food Composition Database for Epidemiological Studies in Italy [102]. Dietary information was judged as unreliable and excluded from further analysis if reported energy intakes were <600 kcal/day or >4,000 kcal/day and >4,200 kcal/day in women and men, respectively.

Blood biomarkers and DNA methylation in the InCHIANTI

Sampling and data collection procedures have been described elsewhere [103]. Briefly, participants were enrolled between 1998 and 2000 and were examined at three-year intervals. Serum samples obtained from blood collected in evacuated tubes without anticoagulant were centrifuged at 2000g for 10 min, and stored at -80 °C for measurement of glucose, total, LDL, and HDL, cholesterol, triglycerides, CRP, and creatinine. DNA methylation was assayed using the Illumina Infinium HumanMethylation450 platform for n=407 participants with sufficient DNA at both baseline (years 1998-2000) and year 9 follow-up visits (2007-2009).

Assessment of metabolic syndrome

Metabolic syndrome status was assessed using the ATPIII NCEP 2004 criteria defined by the presence of 3 or more of the following characteristics: waist circumference >88cm (if male, >102cm), systolic blood pressure >130mmHg or diastolic blood pressure >85mmHg, fasting plasma glucose >100mg/dL, HDL cholesterol <50mg/dL (if male, <40mg/dL), and triglycerides >150mg/dL. In regression models, we use total number of metabolic syndrome characteristics as an ordinal variable, ranging from 0 to 5.

Statistical Analyses

Dietary analysis

Biweight midcorrelation, an outlier-robust correlation measure, was used to assess marginal linear relationships between epigenetic aging measures and dietary, cardiometabolic, and socioeconomic factors. To adjust for possible socioeconomic and lifestyle confounders, we fit ethnically-stratified multivariable linear models adjusting for education, exercise, BMI, and current drinker and smoker status. We used Stouffer's method to infer the meta-analytic significance of each variable over the different ethnic strata using the square-root of the sample size as the Z-score weighting factor. Specifically for the WHI, the age acceleration measures were adjusted for differences in originating dataset and within the InCHIANTI the measures were adjusted for sex. Models including regression on biomarkers, and number of metabolic syndrome symptoms are not stratified by ethnicity due to lack of coverage for biomarker profiling. Models were designed based on common prior knowledge and in cases where there was co-linearity between confounding variables, choice for adjustment was selected based on variable commonality in order to improve comparability with other studies, e.g. BMI was chosen over WHR because BMI is more commonly measured and reported. Variables with skewness >1

were log transformed (possibly adding +1 to avoid forming the logarithm of zero). Mean carotenoids was computed as the mean across standardized measures of lycopene, log2(alpha-carotene), log2(beta-carotene), log2(lutein + zeaxanthin), and log2(beta-cryptoxanthin). Repeat measurements on the same individuals were omitted from the analysis.

Chapter 2: Transcriptomic analysis of monocytes in HIV-associated neurocognitive disorders

ABSTRACT

Events leading to and propagating neurocognitive impairment (NCI) in HIV-1-infected (HIV+) persons are largely mediated by peripheral blood monocytes. We previously identified expression levels of individual genes and gene networks in peripheral blood monocytes that correlated with neurocognitive functioning in HIV+ adults. Here, we expand upon those findings by examining if gene expression data at baseline is predictive of change in neurocognitive functioning two years later. We also attempt to validate the original findings in a new sample of HIV+ patients and determine if the findings are HIV-specific by including HIV-uninfected (HIV-) participants as a comparison group.

At two time points, mRNA was isolated from the monocytes of 123 HIV+ and 60 HIV- adults enrolled in the Multicenter AIDS Cohort Study and analyzed with the Illumina HT-12 v4 Expression BeadChip. All participants received baseline and follow-up neurocognitive testing two years after mRNA analysis. Data were analyzed using standard gene expression analysis and weighted gene co-expression network analysis with correction for multiple testing. Gene sets were analyzed for GO term enrichment.

Only weak reproducibility of associations of single genes with neurocognitive functioning was observed, indicating that such measures are unreliable as biomarkers for HIV-related NCI; however, gene networks were generally preserved between time points and largely reproducible, suggesting that these may be more reliable. Several gene networks associated with variables related to HIV infection were found (e.g., MHC I antigen processing, TNF signaling, interferon gamma signaling, and antiviral defense); however, no significant associations were

found for neurocognitive function. Furthermore, neither individual gene probes nor gene networks predicted later neurocognitive change.

This study did not validate our previous findings and does not support the use of monocyte gene expression profiles as a biomarker for current or future HIV-associated neurocognitive impairment.

INTRODUCTION

HIV-associated neurocognitive disorders (HAND) represent a significant public health issue as they affect as many as half of the estimated 1.2 million HIV-1 infected individuals within the United States alone [104, 105]. A key aspect of the neuropathogenic process leading to HAND is the increased migration across the blood-brain barrier of monocytes [106, 107] driven both by chemokine gradients originating in the CNS and from a peripheral immune response [108-110]. Once in the CNS compartment, monocytes typically differentiate into macrophages which can release pro-inflammatory cytokines and chemokines; if infected with HIV, they may also release viral proteins that are harmful to nearby neurons and other cells [110-115]. Macrophage density in brain is associated with severity of HAND [116], further underscoring the important role of monocyte/macrophages in HAND.

Because the crosstalk between the CNS and circulating blood monocytes is a central mechanism underlying HAND neuropathogenesis, monocytes may hold useful biomarkers of impending or current HAND. For example, CD14+/CD69+ monocytes were a strong indicator of neurologic injury among patients with HIV-associated dementia in the pre-HAART era [106], although this relationship appears to be weaker in the current HAART era [117]. Considering that the vast majority of HAND cases are mild [104, 105], our group previously examined global gene expression within peripheral blood monocytes to identify transcriptional changes

associated with not only in HIV-associated dementia, but neurocognitive functioning in general [118]. By focusing on peripheral molecular genetic mechanisms that may be prodromal to HAND or indicative of mild HAND, this approach was potentially useful because it might enable deeper understanding of early neuropathogenic processes, and open the possibility of preventative therapies. Findings from our cross-sectional study of 86 HIV+ cases implicated a variety of dysregulated genes, most notably Kelch-like ECH-associated protein-1 (KEAP1), Hypoxia up-regulated-1, and interleukin 6 receptor, implicating oxidative stress as an underlying pathogenic process. In addition, weighted gene co-expression network analysis (WGCNA) [119, 120], a systems biologic approach devised to arrive at a biologically meaningful reduction of high dimensional transcriptomic data, implicated mitotic cell cycle and translational elongation as biological processes correlated with neurocognitive functioning. Those results led successful preclinical trials of compounds that elicit broad anti-oxidant and anti-inflammatory responses in monocytes, enhance neuroprotective factors, and decrease viral replication (unpublished data presented by Gruenewald et al., at the 14th meeting of the International Society on NeuroVirology, 2016). Here we expanded upon the previous findings in three ways. First, we attempted to validate the original findings in an independent sample of HIV+ adults. Second, we determined if gene expression changes within monocytes at baseline predicted neurocognitive status two years later. Third, we included a HIV-uninfected comparison group, which allowed us to determine if any associations between the biological signals and clinical variables are HIV-specific. Our hypotheses were: 1) the findings from the initial study would be validated; 2) baseline gene expression characteristics would be predictive of neurocognitive change measured two years later, and; 3) these findings would be HIV-specific; that is, they would not be observed in the HIV- group.

MATERIALS & METHODS

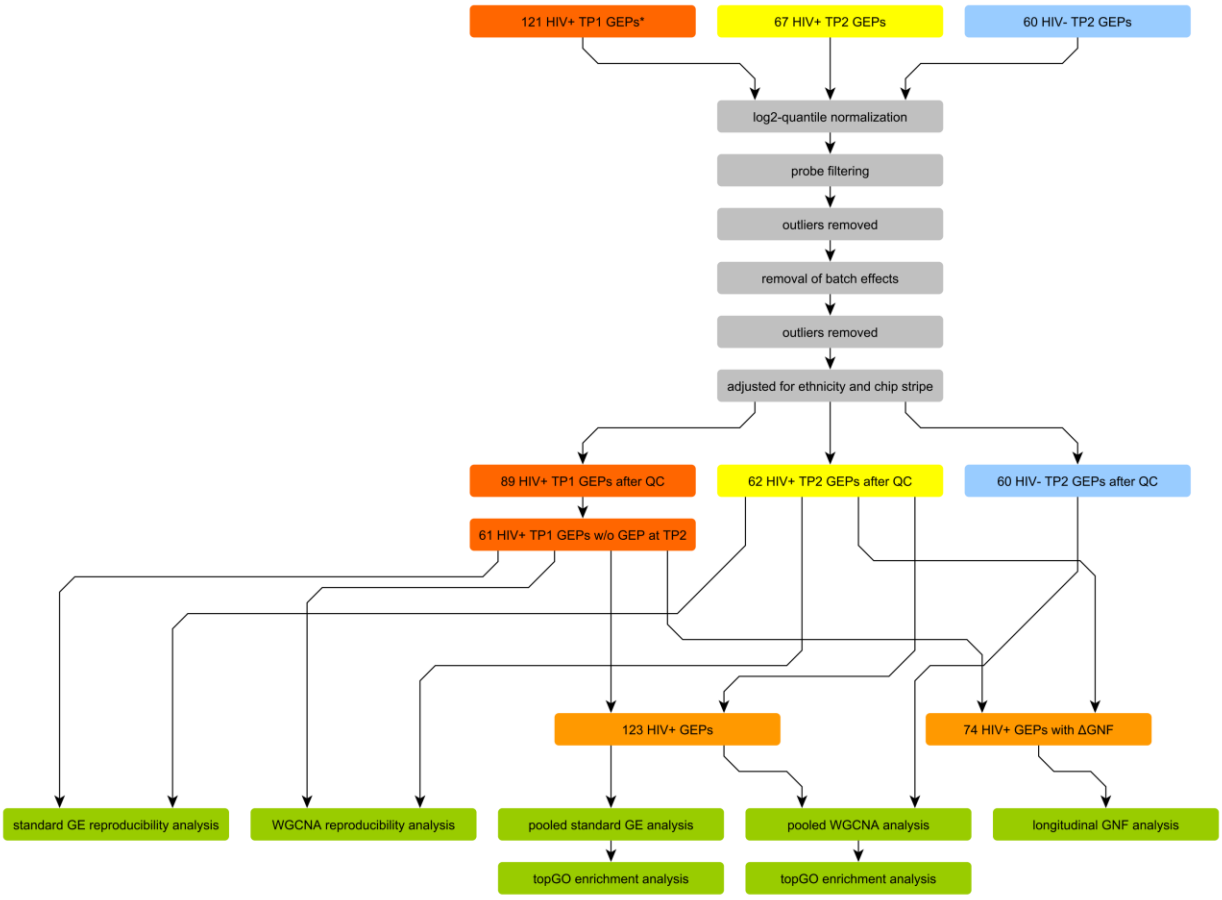
Participants

This study was conducted in accordance with the University of California, Los Angeles Medical Institutional Review Board rules and regulations (IRB#10-001099). All MACS participants who completed the full neuropsychological test battery within 3 weeks of blood draw were eligible. Between 2011-2015, 206 participants in the Multicenter AIDS Cohort Study (MACS) in Los Angeles, California were recruited for this sub-study. The total sample was composed of middle-aged males from white, black, and Hispanic racial groups, all of whom were on ART at the time of the study. Of these, 146 were HIV+ and 60 HIV-seronegative. Monocytes were extracted from the blood of 121 HIV+ cases at baseline (herein referred to as time point 1), and then 67 HIV+ (39 new and 28 returning) and 60 HIV-uninfected cases approximately two years later (herein referred to as time point 2). Due to specific procedural issues (platelets or red blood cell contamination and/or mRNA degradation) several samples were omitted from further analysis. After additional data quality control steps (described below), gene expression data from time point 1 included 89 HIV+ cases and from time point 2 included 62 HIV+ cases (28 of whom were also seen at time point 1) and 60 HIV- cases. Group characteristics are shown in **Table 2-1**, and participant and sample flow from baseline and follow-up visits are detailed in **Figure 2-1**. All participants completed comprehensive self-report questionnaires assessing drug use, medication use, and medical co-morbidities, as well as comprehensive neuropsychological testing and assessment of activities of daily living from which their HAND status was determined. All participants returned after 2-years for follow-up questionnaires and procedures. Procedures and assays were identical to those described in the previous study [118].

Table 2-1. Descriptive statistics of sample sets. Statistics are listed for the time point 1 HIV+, time point 2 HIV+, and time point 2 HIV- sample sets as denoted in the columns of the table. The top portion of the table lists the sample sizes, means, and standard deviations for numeric traits, whereas the bottom portion of the table lists the counts and percentages of categorical or ordinal variables as labeled on the left.

		Time point 1 HIV+			Time point 2 HIV+			Time point 2 HIV-		
		N	**Mean**	**St dev**	**N**	**Mean**	**St dev**	N	**Mean**	**St dev**
Age		89	52.74	9.08	62	51.71	10.50	60	57.20	10.29
Global neurocognitive function (GNF)		89	50.12	7.14	62	49.12	6.67	60	52.00	6.26
Log10 viral load		89	1.41	0.99	55	1.44	1.01	0		
CD4 count		88	601.09	189.81	54	644.22	265.46	60	965.83	271.71
Nadir CD4 count		89	258.79	164.29	62	290.05	164.38	60	616.40	192.77
Duration of HIV-infection, years		89	19.87	8.67	62	17.13	10.15	0		
CNS penetration effectiveness (CPE)		66	1.36	0.80	31	1.13	0.88	0		
		N	**Percent**		**N**	**Percent**		**N**	**Percent**	
Viral load	Undetectable	69	78%		42	76%				
	Detectable	20	22%		13	24%				
HAND	0	65	73%		44	71%		51	85%	
	1	10	11%		13	21%		6	10%	
	2	12	13%		3	5%		3	5%	
	3	2	2%		2	3%				
Education	< 8 years	4	4%		1	2%				
	< 12 years	5	6%		3	5%		3	5%	
	12 years	10	11%		6	10%		5	8%	
	< 16 years	25	28%		28	45%		13	22%	
	16 years	20	22%		13	21%		17	28%	
	> 16 years	25	28%		11	18%		22	37%	
Ethnic group	White non-Hispanic	54	61%		27	44%		42	70%	
	White Hispanic	11	12%		9	15%		4	7%	
	Black non-Hispanic	12	13%		12	19%		5	8%	
	Black Hispanic				1	2%				
	Other				2	3%		1	2%	
	Other Hispanic	12	13%		11	18%		8	13%	
Smoke	Never	18	21%		12	20%		17	29%	
	Former	49	56%		33	56%		34	59%	
	Current	20	23%		14	24%		7	12%	
Alcohol	< Monthly	47	54%		30	51%		29	52%	
	Monthly	14	16%		9	15%		7	13%	
	Weekly	12	14%		13	22%		11	20%	
	Daily	14	16%		7	12%		9	16%	
Hash	< Monthly	64	74%		45	76%		47	84%	
	Monthly	4	5%		4	7%		3	5%	
	Weekly	7	8%		4	7%		4	7%	
	Daily	12	14%		6	10%		2	4%	
Cocaine	< Monthly	85	98%		57	97%		54	96%	
	Monthly				1	2%				
	Weekly	2	2%		1	2%				
	Daily							2	4%	

Figure 2-1. Study workflow diagram. The workflow for the time point 1 (TP1) HIV+ (dark orange), time point 2 (TP2) HIV+ (yellow), and time point 2 HIV- (blue) sample set are illustrated in the workflow diagram. The gene expression profiles (GEPs) for the three sample sets all undergo processing steps (grey); some GEPs are omitted after quality control (QC) steps. More information on these steps can be found in the **Methods**. The input sample sets to the various analyses (green) are denoted by arrows. *****The samples from our previous transcriptome study are included in this sample set.



Blood processing, Monocyte Isolation, mRNA extraction, and gene expression profiling

24ml of fresh blood was collected from participants. Blood was drawn into three 8 mL Cell Preparation Tubes (CPT) containing sodium citrate. Peripheral blood mononuclear cells (PBMCs) were then isolated through centrifugation within 6 hours of collection [121]. PBMCs were washed with phosphate buffered saline, and then monocytes were isolated through Rosette

separation (RosetteSep®; Stem Cell Technologies, British Columbia, Canada) according to the manufacturer instructions. Monocytes were then pelleted, lysed, and RNA extracted using the Qiagen RNeasy kit including a DNase treatment to eliminate any potentially confounding genomic DNA contamination [122]. RNA was stored at -80°C and sent in batches to the Southern California Genotyping Consortium (SCGC) for microarray analysis, which was performed with the Illumina Human HT-12 v4 gene expression BeadChip. The expression data and sample characteristics, including all information required by the MIAME standard, are available from the NCBI Gene Expression Omnibus.

Variables Included in the Gene Expression Analysis

Neurocognitive functioning

Participants completed a comprehensive battery of neuropsychological tests as part of the standard MACS protocol, as previously described [123]. This includes measures of working memory, learning, memory, executive functioning, motor functioning, and information processing speed. T-scores were calculated using normative data derived from the HIV-seronegative MACS cohort, with demographic corrections for age, education, ethnicity, and number of times they had undergone neurocognitive testing. For this study, we calculated a Global Neurocognitive Functioning (GNF) score based on the average of all available domain T-scores. GNF was our primary phenotype.

HAND Severity

HAND status was determined via an algorithm developed by MACS investigators. The algorithm is based on neurocognitive test performance and self-reported deficits in activities of daily living [124] in accordance with current research criteria [125]. Participants were rated as neurocognitively normal, mildly impaired, moderately impaired, or severely impaired. The latter

three correspond to established research criteria; respectively, Asymptomatic Neurocognitive Impairment, Minor Neurocognitive Disorder, and HIV-Associated Dementia. Because of the poor reliability and specificity of the HAND from a diagnostic standpoint [126], we limited this variable to secondary analyses.

CNS Penetration Effectiveness (CPE)

CPE scores for the regimen reported at the time of neurocognitive testing were calculated [127]. Higher scores indicate a regimen with increased penetration of the blood-brain barrier.

Substance Use

We considered the effects of alcohol, marijuana, and cocaine use on gene expression. MACS participants completed a substance use questionnaire that assesses frequency of use during the six months prior to the visit. Participants were considered *active* users of alcohol, stimulants, or marijuana if they report daily or weekly use and *non-users* if they report monthly or less use in the six months preceding the visit. Tobacco use was also considered.

Depression

Depression was determined with the Center for Epidemiologic Studies Depression Scale (CES-D) [128]. Scores on the CES-D were entered as a continuous variable, with higher scores indicating greater degree of depression.

Virologic measures

The percentage of lymphocytes that were CD4+ T-cells was determined by flow cytometry. HIV viral load was determined via either the COBAS TaqMan HIV-1 Test, Version 2.0 or the Roche Amplicor HIV-1 MONITOR Test, Version 1.5. Both tests quantify HIV-1 RNA based on in vitro amplification of the highly conserved HIV-1 gag gene. Nadir CD4+ T cell count was obtained either by self-reports or, for those who seroconverted during the course of the

study, their lowest CD4+ count according to study records. Duration of infection was calculated based on self-reported year of conversion or study records if they seroconverted while in the MACS.

Statistical Analysis

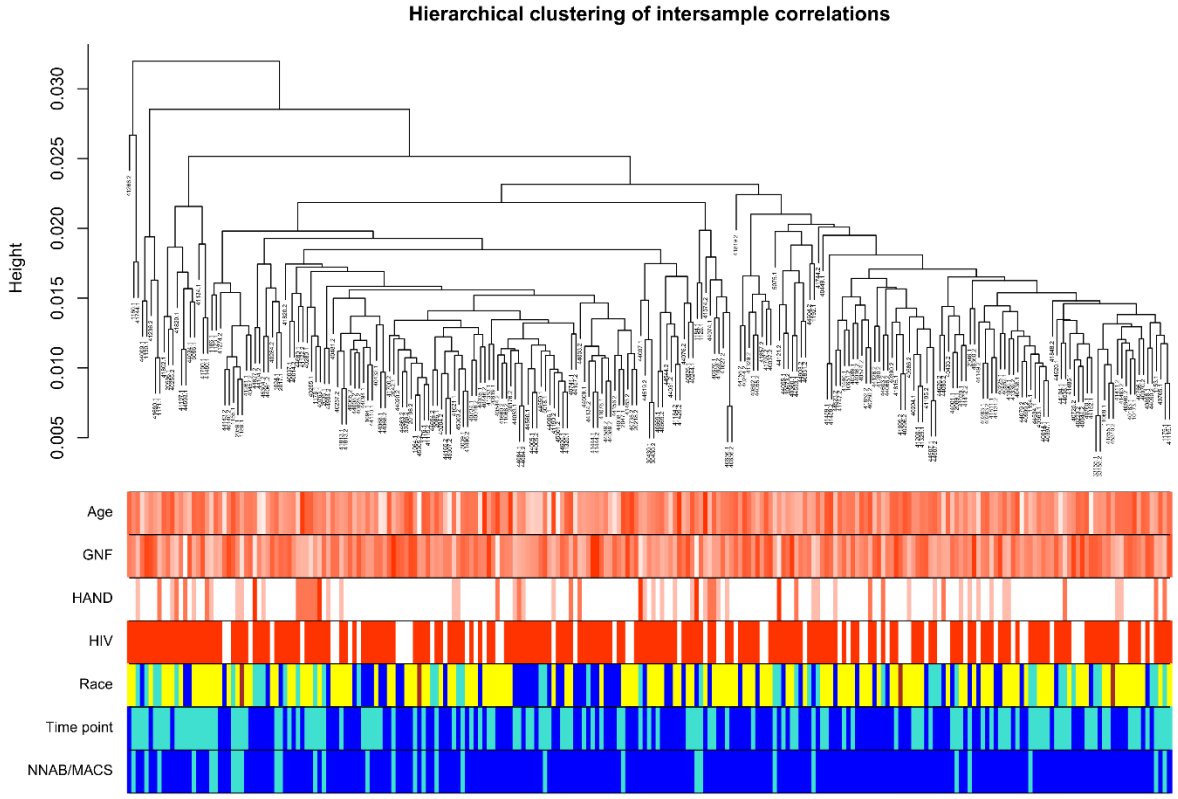
Data Preprocessing

Raw gene expression data was processed in Illumina BeadStudio software and the *lumi* R package was used to log2-transform and quantile normalize the expression profiles to stabilize variance and to normalize inter-sample expression profile distributions, respectively. Probe reannotations provided by the illuminaHumanv3.db R package were used to filter out poor probe hybridization specificity. Probes with significant detection in less than 80% of samples were omitted from further analysis. The data was then batch-corrected for sample chip effects using the *ComBat* R function from the *R package sva* (freely available from http://www.bioconductor.org).

Outliers identified by hierarchical clustering of samples using standardized Euclidean distance and single linkage were removed both before and after batch correction. The expression data was then adjusted for race and chip stripe by retaining the residuals from robust multivariable linear regression on these covariates.

As a final quality control measure, we determined the correlation between the gene expression profiles of all samples. We found strong consistency between the gene expression profiles within and between individuals (**Figure 2-1**). Inter-individual variation was greater than the variation between repeat measurements on the same individual between time points, however even then the lowest inter-sample correlation was strong ($r = 0.93$).

Figure 2-1. Agreement of gene expression profiles between all samples. We found strong consistency between the gene expression profiles within and between individuals. Inter-individual variation was greater than the variation between repeat measurements on the same individual between time points, however even then the lowest inter-sample correlation was strong (r = 0.93).



Differential expression analysis

In our previous study, we found significant correlations between several gene transcript and GNF in a HIV+ sample [118]. Here we assessed the consistency of these findings in an independent sample of HIV+ participants, and also in the HIV- participants in order to determine if the correlations were specific to HIV. Towards these ends, we first correlated gene expression with GNF in the time point 1 samples (excluding samples with repeat measurements at follow-up), and in the HIV+ and HIV- samples at time point 2. These probe-GNF correlations were then

correlated among these subsets to determine the reproducibility of between different HIV+ samples and the agreement between HIV+ and HIV- samples.

In order to maximize power, we then proceeded to test for differential gene expression across all HIV+ samples (excluding repeat measurements) using correlation tests with the variables of interest including GNF, HAND rating, CPE, CES-D, substance use (separately: alcohol, tobacco, marijuana, and cocaine), nadir CD4, and log10 viral load. To address our multiple testing across gene probes, we use a Bonferroni corrected significance threshold.

To examine whether or not individual gene probes measured at time point 1 (for the original sample of 89 HIV+ individuals) or time point 2 (for the second sample of 62 HIV+ individuals and the HIV- comparison group) predicted change in neurocognitive functioning at follow-up visits, we calculated the change in GNF by regressing follow-up GNF on current GNF, retaining the residuals to adjust for the potential confounding effects of regression to the mean. Change in GNF was then subject to correlation with individual gene probes, and module eigengenes in the WGCNA analyses (below).

Weighted Gene Co-Expression Network Analysis (WGCNA)

WGCNA was employed in our previous study to reduce the data into smaller groups of co-expressing genes (modules) which generally represent biologically meaningful pathways [129, 130]. In WGCNA, highly correlated module genes are represented and summarized by the module eigengene, or ME [131], which can then be used in standard statistical analyses. In this study, we first attempted to reproduce the WGCNA results from our previous study by assessing the reproducibility of the gene coexpression network results. This was accomplished by computing the preservation of modules found in the first HIV+ sample (from time point 1) in the second, independent HIV+ sample (from time point 2), as described elsewhere [132]. Briefly, we

use the modulePreservation function from the WGCNA package which computes a module preservation statistic for modules in a reference dataset within a new set of data along with an accompanying significance level (permutation test p-value).

In order to examine associations between modules and variables of interest, we then used the entire sample of HIV+ and HIV- expression profiles (excluding repeat measurements) to construct a gene network using the WGCNA parameter settings power=4 and deepCut=4 which were chosen based on their qualitative optimality for scale-free topology and resolution of finer modules, respectively. We then correlated the identified modules with the variables of interest.

Gene-annotation enrichment analysis

The biological meaning of gene and module associations with GNF and other variables can be elucidated by gene annotation enrichment analysis. For this, we used the topGO R package. For the differential expression analyses (which consider correlations between individual gene probes and variables of interest), we conducted enrichment analysis on the top 5% genes associated with GNF (and change in GNF) in the HIV+ samples and in the HIV- samples, regardless of statistical significance. We conducted an analogous enrichment analysis on the gene coexpression modules identified by the WGCNA analyses. TopGO was run using the Fisher's exact and Kolmogorov-Smirnov significance tests and the weight01 algorithm which takes into account the dependencies present in the GO topology and thus can be considered corrected for multiple testing.

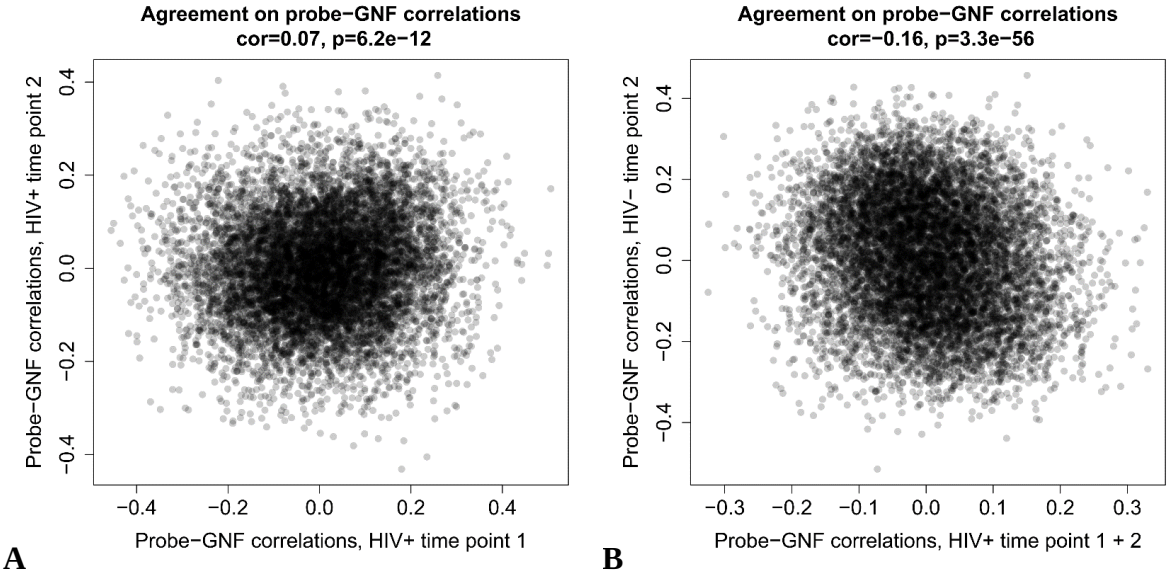
RESULTS

Cross-sectional and longitudinal associations between GNF and gene expression

Agreement between time points

We first sought to assess the reproducibility of the findings of our previous study by comparing gene expression probe-GNF associations between the previous and new study samples. After excluding the repeated measurements on the same individuals to avoid statistical dependency, the sample sizes for time point 1 HIV+ and time point 2 HIV+ groups were 61 and 62, respectively. Of the 89 HIV+ participants from time point 1, 28 also provided blood samples for gene expression analysis at time point 2; we did not include duplicate cases in this analysis, thus our sample size for time point 1 HIV+ is 89 - 28 = 61. None of the top genes identified in our previous study were validated in the independent HIV+ group. Furthermore, the correlation between all probe-GNF correlations for the two different groups was weak (r=0.07), indicating that the reproducibility of the differential expression at the single probe level was unreliable (**Figure 2-2, Panel A**). In comparison, the probe-GNF correlations between the HIV+ groups and HIV- group indicated an inverse association of slightly greater magnitude, either when the HIV+ samples from time point 1 and time point 2 were combined (r = -0.16) or analyzed separately (r = -0.09 and r = -0.15, respectively) (**Figure 2-2, Panel B**). None of these correlations are statistically significant, as the listed p-value for the correlation of correlations is massively inflated since it treats each GNF-probe correlation as independent (n > 10k probes) when in reality there is only a sample size of 2 (HIV+ correlations vs. HIV- correlations). As such, we find poor validation for gene expression between the HIV+ groups, whereas this correlation was somewhat stronger, yet inverse, between HIV+ and HIV- groups.

Figure 2-2. Agreement of probe-GNF correlations from different sample sets. Correlations between the gene expression probe levels in peripheral monocytes and global neurocognitive function (GNF) from different sample sets were plotted against each other in order to assess inter-set agreement. Each point on the scatterplot represents a single gene expression probe with the correlation coefficient with GNF denoted on the x- and y-axes. Correlation coefficients are listed above each scatterplot. A) Correlations in the time point 1 HIV+ samples (n=61) were plotted against correlations in time point 2 HIV+ samples (n=62). B) Probe-GNF correlations in all HIV+ samples (n=123) were plotted against correlations in all HIV- samples (n=60). The HIV+ individuals with repeat measurements at both time points were excluded to avoid artificial inflation of agreement.



Correlations between GNF and gene probe levels among combined sample

In order to maximize statistical power, we combined the HIV+ samples from time points 1 & 2 (excluding repeat measurements) and correlated expression levels with GNF. No significant associations with GNF were found ($p>1.9\times10^{-4}$) at the Bonferroni adjusted significance threshold, $\alpha<5\times10^{-6}$ (**Figure 2-3**). Similarly, no significant associations between probes and GNF were found for the HIV- samples ($p > 2.5\times10^{-5}$) at the Bonferroni adjusted significance threshold.

Figure 2-3. Top correlations between gene probes and HIV status, viral load, and GNF. Traits of interest are listed in the leftmost column with each grouped set of gene probes described in the middle columns. These probes have the top 10 most significant correlations with their respective traits. Correlation coefficients are colored in blue and red for negative and positive correlations respectively. P-values are denoted in green with p-values surpassing transcriptome-wide significance denoted in bold.

	Illumina ID	Gene name	Symbol	cor	p
HIV+ status	ILMN_1763207	basic leucine zipper ATF-like transcription factor 3	BATF3	-0.38	**1E-7**
	ILMN_1655163	serine/threonine kinase 24	STK24	-0.36	**5E-7**
	ILMN_2103841	aryl hydrocarbon receptor interacting protein	AIP	-0.34	**3E-6**
	ILMN_1746704	tripartite motif containing 8	TRIM8	-0.34	**3E-6**
	ILMN_2373010	transmembrane protein 70	TMEM70	0.33	**4E-6**
	ILMN_1706273	MOB kinase activator 2	MOB2	-0.32	9E-6
	ILMN_1738938	translocase of inner mitochondrial membrane 8 homolog B	TIMM8B	0.32	9E-6
	ILMN_1739032	transmembrane protein 70	TMEM70	0.32	1E-5
	ILMN_2411897	Kruppel like factor 10	KLF10	0.32	1E-5
	ILMN_1793950	POTE ankyrin domain family member M	POTEM	0.31	2E-5
Viral laod	ILMN_1711030	5-oxoprolinase (ATP-hydrolysing)	OPLAH	0.42	**3E-6**
	ILMN_2132599	ankyrin repeat domain 22	ANKRD22	0.36	6E-5
	ILMN_1708672	acetyl-CoA acetyltransferase 2	ACAT2	0.35	9E-5
	ILMN_1762725	eukaryotic translation initiation factor 3 subunit L	EIF3L	-0.35	1E-4
	ILMN_1670305	serpin family G member 1	SERPING1	0.35	1E-4
	ILMN_2388547	epithelial stromal interaction 1	EPSTI1	0.35	1E-4
	ILMN_1713285	NSF attachment protein alpha	NAPA	0.35	1E-4
	ILMN_1748650	mitochondrial ribosomal protein L45	MRPL45	-0.35	1E-4
	ILMN_1655497	eukaryotic translation initiation factor 4B	EIF4B	-0.35	1E-4
	ILMN_1749629	cullin 1	CUL1	0.34	2E-4
GNF in HIV+	ILMN_1723020	mitogen-activated protein kinase kinase kinase 1	MAP3K1	0.33	2E-4
	ILMN_2137066	zinc finger protein 7	ZNF7	0.33	2E-4
	ILMN_1740716	RNA binding motif protein 26	RBM26	0.33	2E-4
	ILMN_1763663	HEAT repeat containing 3	AF086132	-0.32	2E-4
	ILMN_1807633	reactive intermediate imine deaminase A homolog	HRSP12	-0.32	3E-4
	ILMN_1801766	mitochondrial calcium uniporter dominant negative beta subunit	CCDC109B	0.32	3E-4
	ILMN_2151048	stromal antigen 1	STAG1	0.32	4E-4
	ILMN_1683313	ST3 beta-galactoside alpha-2,3-sialyltransferase 1	ST3GAL1	0.31	4E-4
	ILMN_1805646	SS18, nBAF chromatin remodeling complex subunit	SS18	0.31	4E-4
	ILMN_1679881	Werner syndrome RecQ like helicase	WRN	0.31	5E-4

To further leverage our data, we then focused on the top 5% genes with the strongest positive and negative correlations with GNF and change in GNF (regardless of statistical significance) and performed gene annotation enrichment analysis using the topGO package. Using this method, genes positively correlated with GNF in HIV+ subjects were found to be enriched for annotations related to c*omplement activation* and consistent with monocyte activation and proliferation (see Table 2-2 below, and Supplemental Table 1 for full details). *Mitochondrial outer membrane permeability* was also a notable finding. Significant GO term

enrichment observed for genes negatively correlated with GNF largely involved *regulation of transcription* and *negative regulation of production miRNA involved in gene silencing*, as well as other seemingly innocuous biological processes. GNF in HIV- cases was positively correlated genes related to *mitochondrial activation*, whereas negatively correlated genes were enriched for *morphogenic activities* (**Table 2-2**).

Table 2-2. GO term enrichment of top genes correlated with GNF. The top 7 enriched GO terms for gene sets comprised of the top 5% of genes most negatively and positively with GNF within sample subsets are presented. The top and bottom halves of the table show the enriched terms for the negatively and positively GNF-correlated gene sets respectively. These halves are divided by HIV+ and HIV- sample sets as labeled on the left along with the trait of interest (GNF or Change in GNF). Enrichment statistics are reported in the rightmost columns including Fisher's exact test p-values.

Table 2. GO term enrichment of top genes correlated with GNF

		GO ID	Term	Annotated	Significant	Expected	Fold enrichment	Fisher's p-value	
HIV+ subjects	GNF	GO:0006958	complement activation, classical pathway	12	6	0.6	10.0	1E-5	GO terms enriched in negatively correlated genes
		GO:0006957	complement activation, alternative pathway	6	4	0.3	13.3	9E-5	
		GO:0097345	mitochondrial outer membrane permeabilization	47	5	2.34	2.1	2E-4	
		GO:0014066	regulation of phosphatidylinositol 3-kinase signaling	75	9	3.74	2.4	7E-4	
		GO:1901299	negative regulation of hydrogen peroxide-mediated programmed cell death	5	3	0.25	12.0	1E-3	
		GO:0038203	TORC2 signaling	5	3	0.25	12.0	1E-3	
		GO:0045916	negative regulation of complement activation	5	3	0.25	12.0	1E-3	
	Change in GNF	GO:2001223	negative regulation of neuron migration	6	5	0.29	17.2	2E-6	
		GO:0060441	epithelial tube branching involved in lung morphogenesis	14	6	0.68	8.8	3E-5	
		GO:0060259	regulation of feeding behavior	7	4	0.34	11.8	2E-4	
		GO:0006953	acute-phase response	21	6	1.01	5.9	4E-4	
		GO:0001656	metanephros development	27	7	1.3	5.4	7E-4	
		GO:0043303	mast cell degranulation	38	5	1.84	2.7	9E-4	
		GO:0007098	centrosome cycle	47	9	2.27	4.0	1E-3	
HIV- subjects	GNF	GO:0045930	negative regulation of mitotic cell cycle	156	8	7.65	1.0	1E-5	
		GO:0060571	morphogenesis of an epithelial fold	8	5	0.39	12.8	1E-5	
		GO:0019896	axonal transport of mitochondrion	5	4	0.25	16.0	3E-5	
		GO:0001922	B-1 B cell homeostasis	6	4	0.29	13.8	8E-5	
		GO:0032909	regulation of transforming growth factor beta2 production	6	4	0.29	13.8	8E-5	
		GO:0002052	positive regulation of neuroblast proliferation	11	5	0.54	9.3	1E-4	
		GO:0009855	determination of bilateral symmetry	34	8	1.67	4.8	1E-4	
HIV+ subjects	GNF	GO:0060065	uterus development	6	4	0.31	12.9	1E-4	GO terms enriched in positively correlated genes
		GO:0000122	negative regulation of transcription from RNA polymerase II promoter	410	38	21.04	1.8	3E-4	
		GO:0045944	positive regulation of transcription from RNA polymerase II promoter	569	52	29.2	1.8	3E-4	
		GO:0006355	regulation of transcription, DNA-templated	1886	143	96.8	1.5	6E-4	
		GO:0007064	mitotic sister chromatid cohesion	19	6	0.98	6.1	1E-3	
		GO:0051056	regulation of small GTPase mediated signal transduction	165	19	8.47	2.2	1E-3	
		GO:1903799	negative regulation of production of miRNAs involved in gene silencing by miRNA	5	3	0.26	11.5	1E-3	
	Change in GNF	GO:0048841	regulation of axon extension involved in axon guidance	9	4	0.45	8.9	6E-4	
		GO:0032007	negative regulation of TOR signaling	28	7	1.39	5.0	7E-4	
		GO:0046323	glucose import	43	4	2.13	1.9	1E-3	
		GO:0007602	phototransduction	24	5	1.19	4.2	1E-3	
		GO:0006417	regulation of translation	305	26	15.14	1.7	3E-3	
		GO:0071380	cellular response to prostaglandin E stimulus	13	4	0.65	6.2	3E-3	
		GO:0032094	response to food	16	5	0.79	6.3	4E-3	
HIV- subjects	GNF	GO:0070125	mitochondrial translational elongation	90	18	4.52	4.0	4E-7	
		GO:0006418	tRNA aminoacylation for protein translation	34	11	1.71	6.4	5E-7	
		GO:0070126	mitochondrial translational termination	89	17	4.47	3.8	2E-6	
		GO:0042776	mitochondrial ATP synthesis coupled proton transport	29	9	1.46	6.2	8E-6	
		GO:0030099	myeloid cell differentiation	232	12	11.66	1.0	3E-4	
		GO:0043985	histone H4-R3 methylation	8	4	0.4	10.0	4E-4	
		GO:0009584	detection of visible light	17	6	0.85	7.1	5E-4	

Predicting change in GNF

We were largely interested in identifying gene expression signals that might predict later neurocognitive change. Seventy-four HIV+ participants with baseline gene expression profiling at either time point 1 or time point 2 were assessed for neurocognitive function again approximately two years later (mean interval=1.9 years). Correlations between gene expression at time point 1 and change in GNF across this period were determined (**Table 2-2**). After adjusting for multiple comparisons, no significant associations were detected between probe levels and change in GNF ($p>2.5\times10^{-5}$). The top GO terms for the top 5% of genes correlated with change in GNF in HIV+ subjects were *negative regulation of neuron migration* and *regulation of axon extension involved in axon guidance* for negatively and positively correlated genes, respectively. (**Table 2-2** and **Table 2-3**).

Table 2-3. GO term enrichment of top genes correlated with GNF and HAND. The top 3 enriched GO terms for gene sets comprised of the top 5% of genes most negatively and positively with neurocognitive traits are presented. The left and right halves of the table show the enriched terms for the positively and negatively correlated gene sets respectively. These halves are divided by HIV+ and HIV- sample sets as labeled on the left, which is subdivided based on correlation with GNF, HAND severity, or HAND status. Enrichment statistics are reported in the rightmost columns of each half including Fisher's exact test p-values.

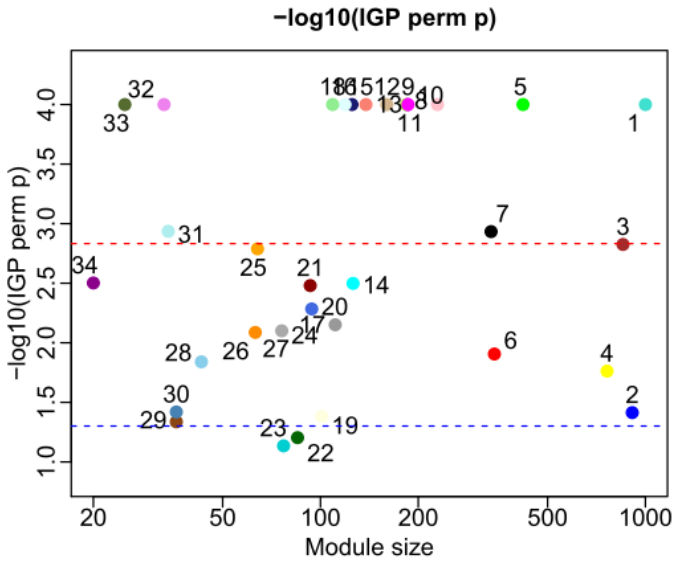
		GO terms enriched in positively correlated genes							GO terms enriched in negatively correlated genes						
		GO ID	Term	Annotated	Significant	Expected	Fold enrichment	Fisher's p-value	GO ID	Term	Annotated	Significant	Expected	Fold enrichment	Fisher's p-value
HIV+ subjects	GNF	GO:0060065	uterus development	6	4	0.31	12.9	1E-4	GO:0006958	complement activation, classical pathway	12	6	0.6	10.0	1E-5
		GO:0000122	negative regulation of transcription from RNA polyr	410	38	21.04	1.8	3E-4	GO:0006957	complement activation, alternative pathway	6	4	0.3	13.3	9E-5
		GO:0045944	positive regulation of transcription from RNA polyr	569	52	29.2	1.8	3E-4	GO:0097345	mitochondrial outer membrane permeabilization	47	5	2.34	2.1	2E-4
	HAND severity	GO:0040012	regulation of locomotion	387	36	19.31	1.9	5E-6	GO:0006614	SRP-dependent cotranslational protein targeting to	181	32	9.2	3.5	4E-10
		GO:0006958	complement activation, classical pathway	12	6	0.6	10.0	1E-5	GO:0000184	nuclear-transcribed mRNA catabolic process, nons	212	35	10.78	3.2	4E-10
		GO:0045916	negative regulation of complement activation	5	4	0.25	16.0	3E-5	GO:0006413	translational initiation	274	38	13.93	2.7	2E-9
	HAND status	GO:0006810	transport	2706	140	138.24	1.0	3E-6	GO:0006413	translational initiation	274	39	13.9	2.8	2E-8
		GO:0040012	regulation of locomotion	387	32	19.77	1.6	6E-6	GO:0006614	SRP-dependent cotranslational protein targeting to	181	27	9.18	2.9	4E-7
		GO:0070527	platelet aggregation	46	14	2.35	6.0	8E-6	GO:0019083	viral transcription	245	32	12.43	2.6	5E-7
	Change in GNF	GO:0048841	regulation of axon extension involved in axon guid	9	4	0.45	8.9	6E-4	GO:2001223	negative regulation of neuron migration	6	5	0.29	17.2	2E-6
		GO:0032007	negative regulation of TOR signaling	28	7	1.39	5.0	7E-4	GO:0060441	epithelial tube branching involved in lung morpho	14	6	0.68	8.8	3E-5
		GO:0046323	glucose import	43	4	2.13	1.9	1E-3	GO:0060259	regulation of feeding behavior	7	4	0.34	11.8	2E-4
HIV- subjects	GNF	GO:0070125	mitochondrial translational elongation	90	18	4.52	4.0	4E-7	GO:0045930	negative regulation of mitotic cell cycle	156	8	7.65	1.0	1E-5
		GO:0006418	tRNA aminoacylation for protein translation	34	11	1.71	6.4	5E-7	GO:0060571	morphogenesis of an epithelial fold	8	5	0.39	12.8	1E-5
		GO:0070126	mitochondrial translational termination	89	17	4.47	3.8	2E-6	GO:0019896	axonal transport of mitochondrion	5	4	0.25	16.0	3E-5
	HAND severity	GO:0031666	positive regulation of lipopolysaccharide-mediatec	9	5	0.44	11.4	3E-5	GO:0019083	viral transcription	245	28	11.93	2.3	2E-5
		GO:0030574	collagen catabolic process	19	6	0.93	6.5	2E-4	GO:0006614	SRP-dependent cotranslational protein targeting to	181	24	8.81	2.7	3E-5
		GO:1902166	negative regulation of intrinsic apoptotic signaling	14	5	0.69	7.2	4E-4	GO:0006364	rRNA processing	317	31	15.43	2.0	4E-5
	HAND status	GO:1900045	negative regulation of protein K63-linked ubiquitine	7	4	0.36	11.1	2E-4	GO:0042776	mitochondrial ATP synthesis coupled proton trans	29	7	1.36	5.1	3E-4
		GO:0002606	positive regulation of dendritic cell antigen proces	5	3	0.25	12.0	1E-3	GO:0006499	N-terminal protein myristoylation	5	3	0.23	13.0	1E-3
		GO:0016064	immunoglobulin mediated immune response	60	6	3.06	2.0	2E-3	GO:0044861	protein transport into plasma membrane raft	5	3	0.23	13.0	1E-3

Weighted Gene Coexpression Network Analysis

Preservation of gene modules between two separate HIV+ samples

We first conducted a WGCNA module preservation analysis between time points 1 and 2 for the nonoverlapping HIV+ participants. The majority of modules from the original sample exhibit significant preservation as indicated by their significant permutation p-values (**Figure 2-4**). These results indicate that at the network level, expression data is reproducible between these two small HIV+ samples.

Figure 2-4. Module preservation statistics between different HIV+ sample sets. The preservation of gene co-expression modules from time point 1 HIV+ samples were assessed by comparing in-group proportion of modules from time point 2 HIV+ samples versus permutated gene expression values to arrive at a permutation p-value. Points on the scatter plot represent modules as denoted by their color and label, where their vertical position denotes increasing negative log-scaled significance and their horizontal position indicates the number of genes in the module. The blue and red dotted lines represent nominal (p=0.05) and Bonferroni significance (p=0.00015) levels respectively. Ten thousand permutations were used in the computation, leading to achievable maximum significance of p=0.0001, which was attained by a number of modules aligned horizontally at the top of the graph.

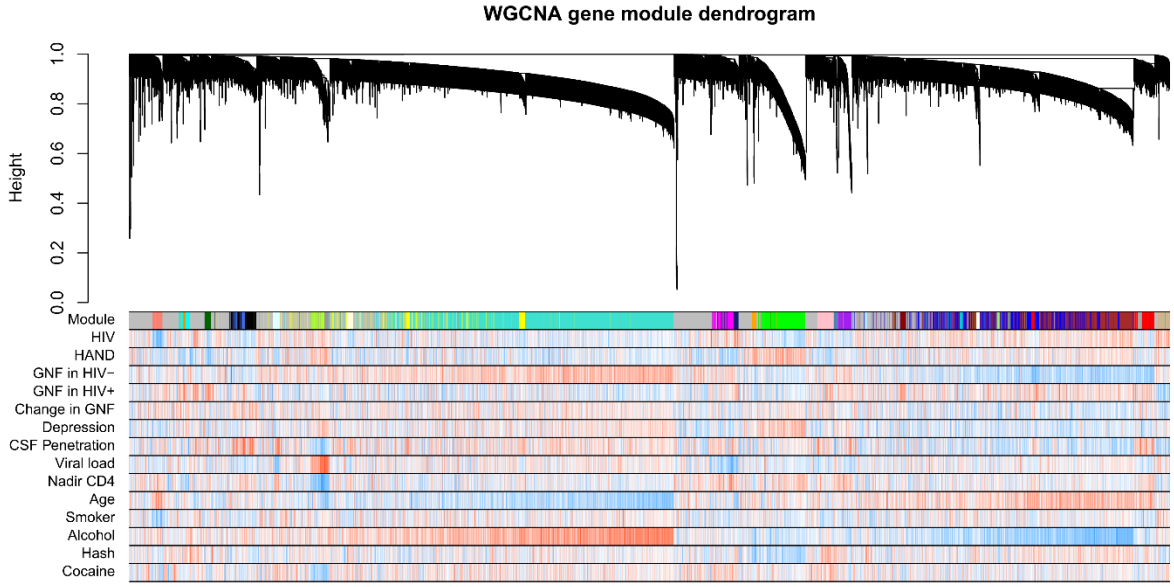


Cross-sectional WGCNA analysis

We conducted a WGCNA analysis of the gene expression data from all samples (HIV+ and HIV-, excluding repeat measurements). The dendrogram of the gene expression WGCNA analysis is shown in **Figure 2-5**. There are several variables showing qualitative relationships with gene clusters. For example, module 1 is negatively correlated with age and positively correlated with reported alcohol intake and GNF in HIV- subjects, whereas modules 2 and 3 appear to have the reverse relationship; they are positively associated with age and negatively

associated with alcohol and GNF in HIV- participants. Globally, the gene expression profiles of the HIV+ and HIV- cases show qualitatively different associations with GNF and HAND (as indicated by opposing red and blue bands on the heatmap).

Figure 2-5. Dendrogram of WGCNA gene modules from pooled HIV+ and HIV- samples. The clustering of genes based on coexpression is represented in the dendrogram with individual gene probes represented as the vertical leafs (black lines) and descending branches indicating coexpression gene clusters. Module labels are shown in the first row by color along with numeric labels displayed above. Subsequent rows show correlations between traits and individual gene probe levels with blue and red denoting negative and positive correlations according to their magnitude.



The resulting eigengenes, each a quantitative value representing the level of a gene module, were then analyzed for correlations with virologic, immunologic, neurocognitive, and drug use variables (**Figure 2-6**). With the Bonferroni-corrected significance threshold of $p < 0.001$, significant associations were found between modules 12 & 18 and viral load (and Nadir CD4 for module 12), and between modules 14, 16, & 24 and HIV status. Gene ontology analyses

for these modules are shown in **Table 2-4**. More comprehensive results are provided in **Table 2-5**.

Figure 2-6. Heatmap of correlations between modules and traits. Correlations between are illustrated in this grid with blue and red representing negative and positive correlations, respectively, according to magnitude as the color scale shows on the right. Module eigengenes are listed in the rows as labeled on the left and traits are listed in the columns as labeled at the bottom with sample numbers described in parentheses. The correlation p-values are printed within the grid; here the Bonferroni significance threshold is p<0.0015.

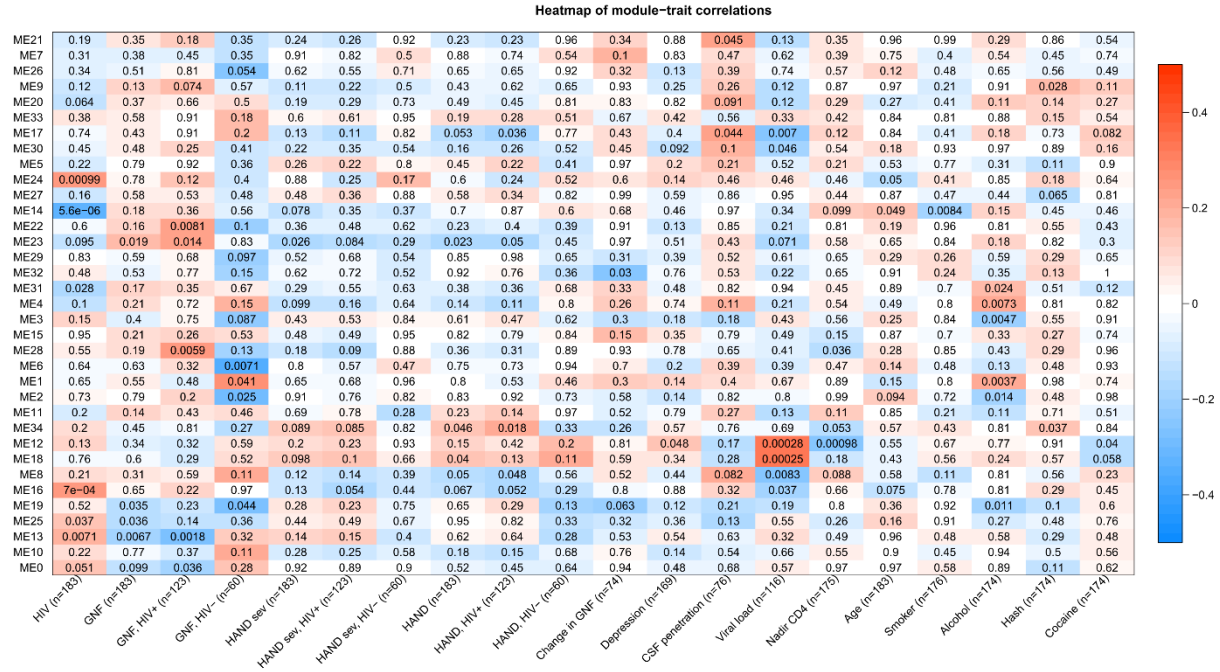


Table 2-4. GO term enrichment of gene modules. The top 7 enriched GO terms for module gene sets are presented. The horizontal portions of the table correspond to the modules with significant trait correlations and are labeled on the left. Enrichment statistics are reported in the rightmost columns including Fisher's exact test p-values.

Module	GO ID	Term	Annotated	Significant	Expected	Fold enrichment	Fisher's p-value
12	GO:0033209	tumor necrosis factor-mediated signaling pathway	138	17	1.75	9.7	3E-14
	GO:0002479	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	80	15	1.02	14.7	4E-14
	GO:0060333	interferon-gamma-mediated signaling pathway	68	16	0.86	18.6	1E-12
	GO:0060337	type I interferon signaling pathway	69	13	0.88	14.8	2E-9
	GO:0006521	regulation of cellular amino acid metabolic process	65	10	0.83	12.0	7E-9
	GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycl...	86	11	1.09	10.1	9E-9
	GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	84	10	1.07	9.3	9E-8
14	GO:0043117	positive regulation of vascular permeability	5	3	0.05	60.0	1E-5
	GO:0038084	vascular endothelial growth factor signaling pathway	8	3	0.08	37.5	6E-5
	GO:0050672	negative regulation of lymphocyte proliferation	31	5	0.32	15.6	1E-4
	GO:0007219	Notch signaling pathway	87	6	0.91	6.6	2E-4
	GO:0050853	B cell receptor signaling pathway	44	5	0.46	10.9	3E-4
	GO:0002250	adaptive immune response	214	8	2.23	3.6	9E-4
	GO:0030035	microspike assembly	5	2	0.05	40.0	1E-3
16	GO:0006413	translational initiation	274	38	2.04	18.6	1E-30
	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	181	36	1.35	26.7	1E-30
	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	212	36	1.58	22.8	1E-30
	GO:0019083	viral transcription	245	36	1.82	19.8	1E-30
	GO:0006364	rRNA processing	317	36	2.36	15.3	1E-30
	GO:0000027	ribosomal large subunit assembly	34	7	0.25	28.0	4E-9
	GO:0075713	establishment of integrated proviral latency	9	3	0.07	42.9	3E-5
18	GO:0051607	defense response to virus	186	29	1.7	17.1	9E-25
	GO:0045071	negative regulation of viral genome replication	42	16	0.38	42.1	6E-23
	GO:0035455	response to interferon-alpha	19	10	0.17	58.8	2E-12
	GO:0039530	MDA-5 signaling pathway	9	5	0.08	62.5	3E-6
	GO:0033159	negative regulation of protein import into nucleus, translocation	5	3	0.05	60.0	7E-6
	GO:0010847	regulation of chromatin assembly	5	3	0.05	60.0	7E-6
	GO:0034341	response to interferon-gamma	112	12	1.02	11.8	9E-6
24	GO:0032467	positive regulation of cytokinesis	12	2	0.07	28.6	2E-3
	GO:1902600	hydrogen ion transmembrane transport	88	5	0.52	9.6	3E-3
	GO:1900153	positive regulation of nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay	14	2	0.08	25.0	1E-2
	GO:0042776	mitochondrial ATP synthesis coupled proton transport	29	2	0.17	11.8	1E-2
	GO:0006278	RNA-dependent DNA biosynthetic process	56	2	0.33	6.1	2E-2
	GO:0000398	mRNA splicing, via spliceosome	281	6	1.65	3.6	2E-2
	GO:0018279	protein N-linked glycosylation via asparagine	39	2	0.23	8.7	2E-2

Regarding GNF, several additional modules indicated trends towards significance ($p < 0.01$). For example, GNF in HIV+ individuals is positively correlated with modules 22 ($p = 0.008$) and 28 ($p = 0.006$), which appear to be enriched for genes involved in *protein ubiquitinylation process*, whereas module 13 has a negative correlation with GNF ($p = 0.002$) and is enriched for *gluconeogenic activity*. For HIV- individuals, only module 6 has a negative correlation ($p = 0.007$) with GNF and is enriched for adaptive immune response.

Table 2-5. GO term enrichment of gene modules. The top 3 enriched GO terms for all module gene sets are presented. The horizontal portions of the table correspond to the modules as labeled on the left. Enrichment statistics are reported in the rightmost columns including Fisher's exact test p-values.

Module	GO ID	Term	Annotated	Significant	Expected	Fold enrichment	Fisher's p-value
0	GO:0006614	SRP-dependent cotranslational protein targ	181	64	42.54	1.5	1E-4
	GO:0019886	antigen processing and presentation of exo	80	34	18.8	1.8	1E-4
	GO:0006413	translational initiation	274	77	64.4	1.2	2E-4
1	GO:0009165	nucleotide biosynthetic process	167	51	42.2	1.2	1E-5
	GO:0006418	tRNA aminoacylation for protein translation	34	20	8.59	2.3	3E-5
	GO:0070126	mitochondrial translational termination	89	40	22.49	1.8	4E-5
2	GO:0000122	negative regulation of transcription from RN	410	66	38.89	1.7	2E-5
	GO:2000052	positive regulation of non-canonical Wnt sig	9	5	0.85	5.9	8E-5
	GO:1904953	Wnt signaling pathway involved in midbrain	10	6	0.95	6.3	1E-4
3	GO:1990573	potassium ion import across plasma memb	6	6	0.56	10.7	6E-7
	GO:0006883	cellular sodium ion homeostasis	11	8	1.02	7.8	7E-7
	GO:0050711	negative regulation of interleukin-1 secretion	9	7	0.84	8.3	2E-6
4	GO:0043968	histone H2A acetylation	19	5	0.81	6.2	1E-3
	GO:0060213	positive regulation of nuclear-transcribed m	12	4	0.51	7.8	1E-3
	GO:0071285	cellular response to lithium ion	6	3	0.25	12.0	1E-3
5	GO:0070527	platelet aggregation	46	19	1.88	10.1	2E-12
	GO:0007229	integrin-mediated signaling pathway	62	17	2.54	6.7	2E-10
	GO:0002576	platelet degranulation	72	18	2.94	6.1	4E-10
6	GO:0006817	phosphate ion transport	12	4	0.24	16.7	8E-6
	GO:0042088	T-helper 1 type immune response	19	4	0.39	10.3	2E-4
	GO:0061088	regulation of sequestering of zinc ion	7	3	0.14	21.4	3E-4
7	GO:0010923	negative regulation of phosphatase activity	39	6	0.65	9.2	4E-5
	GO:0060263	regulation of respiratory burst	10	3	0.17	17.6	5E-4
	GO:1902166	negative regulation of intrinsic apoptotic sig	14	3	0.24	12.5	1E-3
8	GO:0006614	SRP-dependent cotranslational protein targ	181	44	2.87	15.3	1E-30
	GO:0006413	translational initiation	274	51	4.34	11.8	1E-30
	GO:0000184	nuclear-transcribed mRNA catabolic proces	212	44	3.36	13.1	1E-30
9	GO:0050684	regulation of mRNA processing	94	9	1.25	7.2	1E-4
	GO:0043507	positive regulation of JUN kinase activity	45	4	0.6	6.7	1E-3
	GO:0009408	response to heat	115	6	1.53	3.9	6E-3
10	GO:0006120	mitochondrial electron transport, NADH to u	39	9	0.56	16.1	3E-9
	GO:0032981	mitochondrial respiratory chain complex I a	52	9	0.75	12.0	4E-8
	GO:0001302	replicative cell aging	5	4	0.07	57.1	2E-7
11	GO:0006955	immune response	969	32	13.13	2.4	3E-6
	GO:0006968	cellular defense response	32	5	0.43	11.6	6E-5
	GO:0019835	cytolysis	17	4	0.23	17.4	7E-5
12	GO:0033209	tumor necrosis factor-mediated signaling pa	138	17	1.75	9.7	3E-14
	GO:0002479	antigen processing and presentation of exo	80	15	1.02	14.7	4E-14
	GO:0060333	interferon-gamma-mediated signaling pathv	68	16	0.86	18.6	1E-12
13	GO:0043456	regulation of pentose-phosphate shunt	7	5	0.08	62.5	3E-9
	GO:0006094	gluconeogenesis	58	7	0.63	11.1	3E-6
	GO:0045899	positive regulation of RNA polymerase II tran	12	4	0.13	30.8	6E-6
14	GO:0043117	positive regulation of vascular permeability	5	3	0.05	60.0	1E-5
	GO:0038084	vascular endothelial growth factor signaling	8	3	0.08	37.5	6E-5
	GO:0050672	negative regulation of lymphocyte proliferati	31	5	0.32	15.6	1E-4
15	GO:1900169	regulation of glucocorticoid mediated signa	5	3	0.04	75.0	6E-6
	GO:0044770	cell cycle phase transition	439	4	3.74	1.1	5E-4
	GO:0021762	substantia nigra development	43	4	0.37	10.8	5E-4
16	GO:0006413	translational initiation	274	38	2.04	18.6	1E-30
	GO:0006614	SRP-dependent cotranslational protein targ	181	36	1.35	26.7	1E-30
	GO:0000184	nuclear-transcribed mRNA catabolic proces	212	36	1.58	22.8	1E-30
17	GO:0090201	negative regulation of release of cytochrome	17	3	0.15	20.0	4E-4
	GO:1901077	regulation of relaxation of muscle	5	2	0.04	50.0	8E-4
	GO:2000377	regulation of reactive oxygen species metab	119	5	1.07	4.7	1E-3
18	GO:0060337	type I interferon signaling pathway	69	22	0.63	34.9	5E-25
	GO:0051607	defense response to virus	186	29	1.7	17.1	9E-25
	GO:0045071	negative regulation of viral genome replicati	42	16	0.38	42.1	6E-23
19	GO:0006499	N-terminal protein myristoylation	5	2	0.03	66.7	4E-4
	GO:0035970	peptidyl-threonine dephosphorylation	9	2	0.06	33.3	1E-3
	GO:0007183	SMAD protein complex assembly	10	2	0.06	33.3	2E-3
20	GO:0071569	protein ufmylation	6	2	0.04	50.0	7E-4
	GO:0071300	cellular response to retinoic acid	26	3	0.17	17.6	7E-4
	GO:0030262	apoptotic nuclear changes	28	3	0.19	15.8	8E-4
21	GO:0016567	protein ubiquitination	602	14	4.62	3.0	1E-3
	GO:0051457	maintenance of protein location in nucleus	9	2	0.07	28.6	2E-3
	GO:0035518	histone H2A monoubiquitination	10	2	0.08	25.0	3E-3
22	GO:0018146	keratan sulfate biosynthetic process	14	3	0.09	33.3	9E-5
	GO:0085020	protein K6-linked ubiquitination	5	2	0.03	66.7	4E-4
	GO:0071763	nuclear membrane organization	6	2	0.04	50.0	6E-4
23	GO:0023014	signal transduction by protein phosphorylati	521	8	3.75	2.1	4E-7
	GO:0003214	cardiac left ventricle morphogenesis	6	3	0.04	75.0	7E-6
	GO:0070989	oxidative demethylation	7	3	0.05	60.0	1E-5
24	GO:0006123	mitochondrial electron transport, cytochrom	11	2	0.06	33.3	2E-3
	GO:0032467	positive regulation of cytokinesis	12	2	0.07	28.6	2E-3
	GO:1902600	hydrogen ion transmembrane transport	88	5	0.52	9.6	3E-3
25	GO:0006527	arginine catabolic process	6	2	0.04	50.0	6E-4
	GO:0009084	glutamine family amino acid biosynthetic pr	6	2	0.04	50.0	6E-4
	GO:0034214	protein hexamerization	7	2	0.04	50.0	8E-4
26	GO:0006390	transcription from mitochondrial promoter	7	2	0.03	66.7	4E-4
	GO:0031167	rRNA methylation	17	2	0.08	25.0	3E-3
	GO:0006352	DNA-templated transcription, initiation	144	4	0.64	6.3	4E-3
27	GO:0015671	oxygen transport	8	5	0.03	166.7	5E-11
	GO:0042744	hydrogen peroxide catabolic process	15	3	0.06	50.0	3E-5
	GO:0051881	regulation of mitochondrial membrane pote	47	4	0.2	20.0	4E-5
28	GO:0006991	response to sterol depletion	13	2	0.05	40.0	4E-3
	GO:0045944	positive regulation of transcription from RNA	569	8	2.32	3.4	5E-3
	GO:0010501	RNA secondary structure unwinding	35	2	0.14	14.3	9E-3
29	GO:0071356	cellular response to tumor necrosis factor	179	8	0.69	11.6	6E-9
	GO:0070098	chemokine-mediated signaling pathway	24	5	0.09	55.6	2E-8
	GO:0002675	positive regulation of acute inflammatory res	14	4	0.05	80.0	2E-7
30	GO:0032287	peripheral nervous system myelin maintena	5	2	0.02	100.0	1E-4
	GO:0045725	positive regulation of glycogen biosynthetic p	5	2	0.02	100.0	1E-4
	GO:0032891	negative regulation of organic acid transport	5	2	0.02	100.0	1E-4
31	GO:0070206	protein trimerization	16	2	0.05	40.0	1E-3
	GO:0021766	hippocampus development	37	2	0.12	16.7	7E-3
	GO:0035023	regulation of Rho protein signal transduction	47	2	0.16	12.5	1E-2
32	GO:0050920	regulation of chemotaxis	81	5	0.3	16.7	2E-7
	GO:0002407	dendritic cell chemotaxis	19	4	0.07	57.1	6E-7
	GO:0090050	positive regulation of cell migration involved	6	3	0.02	150.0	9E-7
33	GO:0022417	protein maturation by protein folding	8	2	0.02	100.0	1E-4
	GO:0034975	protein folding in endoplasmic reticulum	9	2	0.02	100.0	2E-4
	GO:0036500	ATF6-mediated unfolded protein response	10	2	0.02	100.0	2E-4
34	GO:0045746	negative regulation of Notch signaling pathv	12	2	0.03	66.7	4E-4
	GO:0043306	positive regulation of mast cell degranulatio	13	2	0.03	66.7	5E-4
	GO:0046627	negative regulation of insulin receptor signa	23	2	0.06	33.3	1E-3

WGCNA at time point 1 as a predictor of later neurocognitive change

Change in GNF was not significantly associated with any time point 1 modules (**Figure 2-6**).

DISCUSSION

In this study, we attempted to replicate our previous findings that neurocognitive functioning in HIV+ persons was correlated with the expression of several oxidative-stress-related genes in peripheral blood monocytes. We also sought to expand those findings by determining if gene expression profiles in such cells could predict neurocognitive status two years later, and whether or not any associations or predictive markers were specific to HIV+ persons or were also observed in an HIV- comparison sample.

Contrary to our hypotheses, we were unable to replicate the findings from our earlier study [118], which had implicated several genes involved in anti-oxidant response. Despite some overlap between the current and previous study, there was a substantial number of samples that were different in the current study—only 61 out of the 123 samples were from the original analysis. The lack of reproducibility of our previous top associations is consistent with the weak agreement found between our two cross-sectional samples at the single gene level. Also contrary to our hypotheses, gene expression characteristics determined at baseline did not predict neurocognitive decline as measured two years later. This includes both individual gene transcripts, modules consisting of co-varying gene networks, and biological ontologies based on top correlations. These results, although unexpected, provide strong evidence that a useful concurrent or predictive biomarker of HIV-associated neurocognitive impairment is unlikely to be found in the gene expression profiles of monocytes, a finding also supported by past studies [133], as also reviewed in [134, 135].

An alternative explanation for the null results may be that our primary phenotype (global neurocognitive functioning) is affected not only by HIV, but by other factors including substance use, HCV co-infection, pre-existing cognitive deficits, and error due to psychometric characteristics of the tests and participant effort [125, 136, 137]. This is especially true of mild neurocognitive deficits, which would generally be seen in the relatively healthy MACS participants [105]. We chose GNF as our primary outcome variable because the diagnosis of HAND is unreliable, as demonstrated by Woods et al. [126] and further indicated by the near equal number of HIV-seronegative control cases that meet criteria for this condition [105, 137]. Therefore, if one were to focus advanced HAND cases (e.g., HIV-associated dementia) in analyses such as ours, more consistent signals are more likely to be found. The problem with this approach, however, is that advanced cases are increasingly rare, thus being statistical underpowered for similarly sized studies. A power analysis indicates that in order to have 80% power to detect a weak correlation of $r = 0.3$ at a transcriptome-wide significance level of $p < 5x10^{-6}$, we would need approximately 300 samples; analogously a modular approach with a less stringent significance threshold of $p < 0.001$ would still require at least 170 samples. However, because we were searching for biomarkers of HAND, the value of weak associations would be insubstantial considering that biomarkers require medium to large effect sizes.

Despite these null results, there are several indications that the findings from this are valid and meaningful. For example, we found that alcohol intake and GNF in HIV- subjects appeared to have anti-aging gene expression signatures (increased mitochondrial function and decreased transcriptional activity), which is consistent with a growing body of literature establishing the healthful effects of moderate alcohol consumption [138, 139]. Additionally, the WGCNA results related to our other variables of interest as expected. The strong effects of HIV

infection and viral load yielded clear correlations between HIV viral load and modules enriched for gene networks involved in immune response (e.g., MHC I antigen processing and presentation, TNF signaling, and interferon gamma signaling) and antiviral defense. Furthermore, HIV infection was associated with glycoprotein functioning and translation/transcription processes (e.g., SRP-dependent co-translational protein targeting to membrane, translation initiation, and viral transcription). Finally, the module preservation analysis showed that gene coexpression structure was preserved between our two samples, indicating that though the expression of individual genes is inconsistent, gene modules are reproducible.

It is worth noting that the non-significant trends between GNF and modules 6, 13, 22, and 28, broadly suggest a potential relationship with regulation of glucose metabolism and ubiquitin-proteasomal based protein. It is unclear what relation this may have with previous studies of proteasomal regulation in brains of HIV+ cases with HIV-associated dementia [140], but our results suggest that upregulation of this process in monocytes is associated with better neurocognitive function. Additional biological functions associated with GNF that were implicated by the GO analysis, and that also have some support via previous studies, include activation of NFκβ-inducing kinase activity [141], tumor necrosis factor-mediated signaling pathway [141], and positive regulation of canonical Wnt signaling pathway and beta-catenin-TCF complex assembly [142]. However, while our findings may provide support for dysregulation of these processes in association with HAND, they strongly indicate that none are so crucial that they could serve as biomarkers, at least not based on transcript levels.

In summary, the results from our study show that monocyte transcriptional profiles are not significantly predictive of future GNF or reliably associated with current GNF. While this

may be due in part to an imperfect neurocognitive phenotype or underpowered sample, our results suggest that there are no strong relationships between gene expression in peripheral blood monocytes and GNF in HIV+ individuals.

Chapter 3: Transcriptomic signatures of epigenetic aging in blood

ABSTRACT

The epigenetic clock is highly predictive of chronological age and has been shown to relate to many age-related phenotypes and outcomes. Though the link between epigenetic aging and markers of inflammation is becoming increasingly clear, the answer to how these two are related is remains elusive. In order address this, we analyze global gene expression and DNA methylation profiles from 2,188 peripheral leukocytes samples in the Framingham Heart Study (FHS) and from 1,202 purified monocytes samples in the Multi-Ethnic Study of Atherosclerosis.

Epigenetic age acceleration in peripheral leukocytes were associated with increased granulocyte count estimates ($r=0.23$, $p=2x10^{-27}$), DNAm plasma biomarker signatures ($r>0.35$, $p<2x10^{-65}$), and female sex ($r=-0.25$, $p=2x10^{-33}$). Epigenetic aging in monocytes was associated with DNAm plasma biomarker signatures ($r>0.22$, $p<5x10^{-14}$), and female sex ($r=-0.27$, $p=2x10^{-21}$). Associations between epigenetic aging and global gene expression and DNA methylation appeared to be dominated by cell composition, race, or sex-based effects. Desmocollin 2 (DSC2), a gene which plays a role in the formation of cell-cell junctions, was found to be among the top associations with epigenetic aging in both datasets. GO term analysis revealed the enrichment of interferon signaling among transcripts associated with epigenetic age acceleration in both the peripheral leukocyte and purified monocyte datasets.

Overall, this study supports the multi-factorial etiology of this phenomenon, and further suggests a possible role of interferon-mediated cellular senescence as a mechanism for cell-intrinsic aging of the epigenome.

INTRODUCTION

There is a growing body of work investigating the molecular mechanisms underlying the epigenetic clock. The Horvath clock, which tracks chronological age across nearly all tissues, consists of CpGs that are over-represented near Polycomb-group target genes [5]. A GWAS study in brain tissue identified genome-wide significant SNPs that implicate mTOR and DNA topology [143]. Another GWAS in leukocytes found an association between genetic variants in the telomerase reverse transcriptase gene and intrinsic epigenetic age acceleration [144], however epigenetic age acceleration was found to only weakly correlate with telomere length [145]. A study of senescent and immortalized cells in vitro showed that replicative and oncogenic induced senescence and immortalized proliferation in culture are accompanied by epigenetic aging of the cells however DNA damage induced senescence did not [146]. Though much progress has been made in understanding the molecular underpinning of the epigenetic clock, there still does not appear to be a unified explanation for this phenomenon.

Classically, DNA methylation marks are thought to coordinate the accessibility of chromatin with hypomethylation being associated with open chromatin and increased local gene transcription and hypermethylation being associated with closed heterochromatin and gene silencing. The interplay between DNA methylation and gene expression now understood to be much more complex with the "epigenetic code" being context specific both with respect to epigenomic machinery and other local epigenetic marks. Given the strong conceptual relationship between DNA methylation and transcription, we asked whether the epigenetic clock was associated with characteristic changes in gene expression. Here I describe the results from work examining the relationship between genome-wide gene expression and epigenetic aging in

peripheral blood mononuclear cells and in purified CD14+ monocytes in order to elucidate any potential relationships between these two processes.

METHODS

Data collection and preprocessing

To investigate the associations between gene expression and epigenetic aging in blood, we obtained data from the Framingham Heart Study (**FHS**) based near Framingham, Massachusetts. The initial study was established in 1948 and enrollment of their offspring began in 1971 with in-person evaluations every 4-8 years. This study is limited to the consenting offspring that survived until the 8th examination cycle 2005-2008 when the peripheral blood samples were collected. DNA methylation and gene expression data were acquired from these isolated PBMCs (**FHS**, n=2188) [147] using the Illumina HumanMethylation450 Beadchip array and Affymetrix Human Exon 1.0 ST Array platforms respectively. Another dataset was acquired from individuals in the Multi-Ethnic Study of Atherosclerosis study (**MESA**, n=1202) [148]. Purified CD14+ monocytes were used to profile global DNA methylation (**DNAm**) the using the Illumina HumanMethylation450 Beadchip and global gene expression (**GEx**) using the Illumina HumanHT12v4 platform, respectively.

Both the GEx and DNAm underwent analogous data preprocessing steps: missing values were imputed using k-nearest neighbors, profiles were quantile normalized, and samples were adjusted for batch effects using the ComBat R package. Estimates of DNAm age, epigenetic age acceleration, and cell counts were computed using the online DNAm age calculator (http://labs.genetics.ucla.edu/horvath/dnamage/) or from custom models. The variables analyzed here include the intrinsic epigenetic age acceleration measures AgeAccelerationResidual (developed in multiple tissues), the extrinsic epigenetic age acceleration measures

BioAge1HAAdjAge (developed in blood), AgeAccelPheno (developed based on age-related phenotypes and disease outcomes), AgeAccelImmuno (developed on plasma protein biomarkers), AgeAccelSkinClock (developed on skin cells and blood), and AgeAccelPC (the first principal component of all aforementioned AgeAccel measures), and the Houseman cell count estimates for Granulocytes, Monocytes, B, NK, CD4 T, and CD8 T cells, in addition to the Horvath estimates for plasmablasts, naive CD4 and CD8 T cells, and exhausted CD8 T cells (CD8pCD28nCD45RAn).

Age-adjusted surrogate plasma biomarkers are also computed from the DNAm data including GDF15, B2M, cystatin C, TIMP1, adrenomedullin (adm), plasminogen activator inhibitor type 1 (PAI), and leptin. Briefly, elastic net regularized regression was used to develop DNAm-based models to predict immunoassay-measured plasma biomarkers in the FHS cohort. These estimates were then adjusted by rescaling them to be proportional with chronologic age, and then regressing out chronological age and retaining the residuals, arriving at the age-adjusted surrogate biomarkers e.g. "AgeAccelleptin". AgeAccelImmuno is a weighted average of the DNAm age estimates based on these plasma proteins.

Data analysis

Weighted gene correlation network analysis was used to infer co-expression and co-methylation modules and compute their module eigenvalues using the WGCNA R package. Principal components analysis was used to infer the major axes of covariance in the GEx and DNAm data. GO term enrichment was tested using Fisher's exact and Kolgorov-Smirnov tests implemented in the TopGO R package.

To explore the relationship between gene expression, epigenetic aging, and other variables, probes and modules were tested for pairwise correlations. The most significant probe

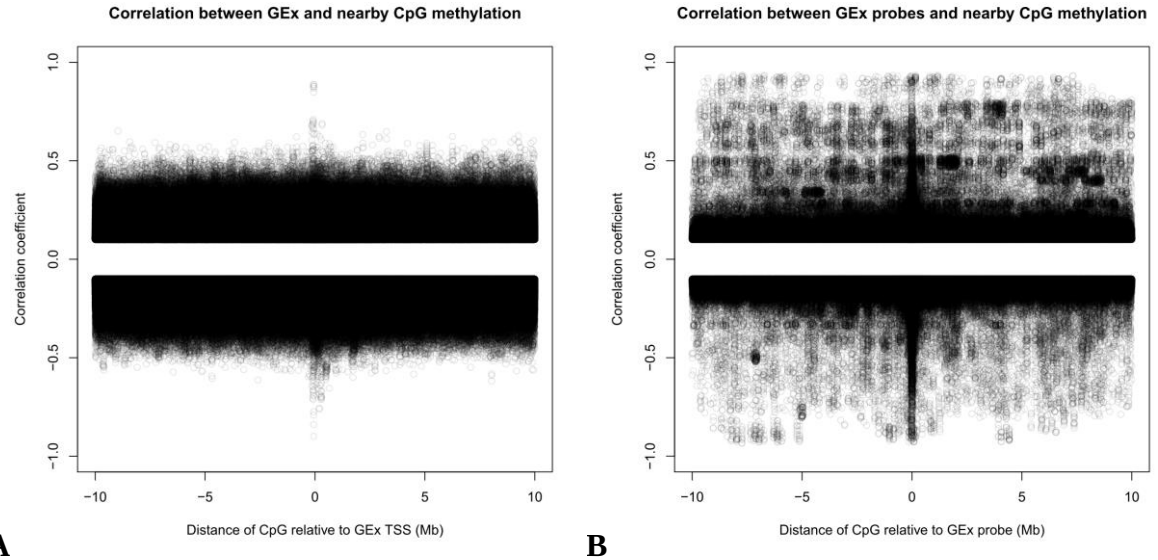
and module associations are presented in tabular format. Correlations are visualized using heatmaps colored in accordance with their sign and magnitude.

RESULTS AND DISCUSSION

Relationships between gene expression and DNA methylation levels in blood

Despite the theoretical ties between gene expression and DNA methylation, it is unclear whether these layers of cellular information have high correspondence in practice. In order to address this issue, we analyzed GEx and DNAm data from PBMCs and purified monocytes. In PBMCs, gene transcript levels exhibit moderate to strong correlations ($r<0.9$) with local DNAm levels (<1Mb away) and weak to moderate strength correlations ($r<0.5$) with distal CpGs up to 10Mb away (**Figure 3-1**). In contrast, gene expression probes in purified monocytes range in correlations with DNA methylation at distant CpG sites ($r<0.9$) though generally distal correlations are weak ($r<0.3$) with strong relationships punctuated throughout the examined range; there is a clear enrichment for strong associations occurring within 1 Mb of the transcription site ($r<0.9$). These results suggest that the relationships between individual DNAm and GEx markers may be "averaged out" when examining heterogeneous cell populations. The subtle epigenetic interactions present in one cell type may not exist in another, resulting in the diminished detectability of these relationships when these subpopulations are combined. In addition, entire cell-type specific epigenetic signatures would be associated with transcripts also specific to the cell-type, causing an apparent genome-wide inflation of distal associations. Overall, these results reiterate the complexity of the epigenetic code and demonstrate the loss of information due to cell type confounding in studies of cell mixtures.

Figure 3-1. Associations between gene expression and nearby DNA methylation. Correlation coefficients between transcript expression and CpG methylation levels are plotted against CpG distance from the gene location. **A**. Gene-level correlations in PBMCs from the FHS are presented. Moderate correlations are observed between transcript levels and CpG methylation across the entire surrounding 20Mb region; local effects (<1Mb) are only slightly stronger than observed distal effects. **B**. Gene probe-level correlations in purified monocytes from MESA are presented. The majority of CpG-transcript correlations are weak (r<0.3) though strong correlations are observed throughout the 20Mb region. Positive correlations appear to be more punctuated compared to negative correlations, indicating complexity of, and likely gene-specific, epigenetic interactions. Correlations less than r<0.1 are excluded from both scatterplots.



A **B**

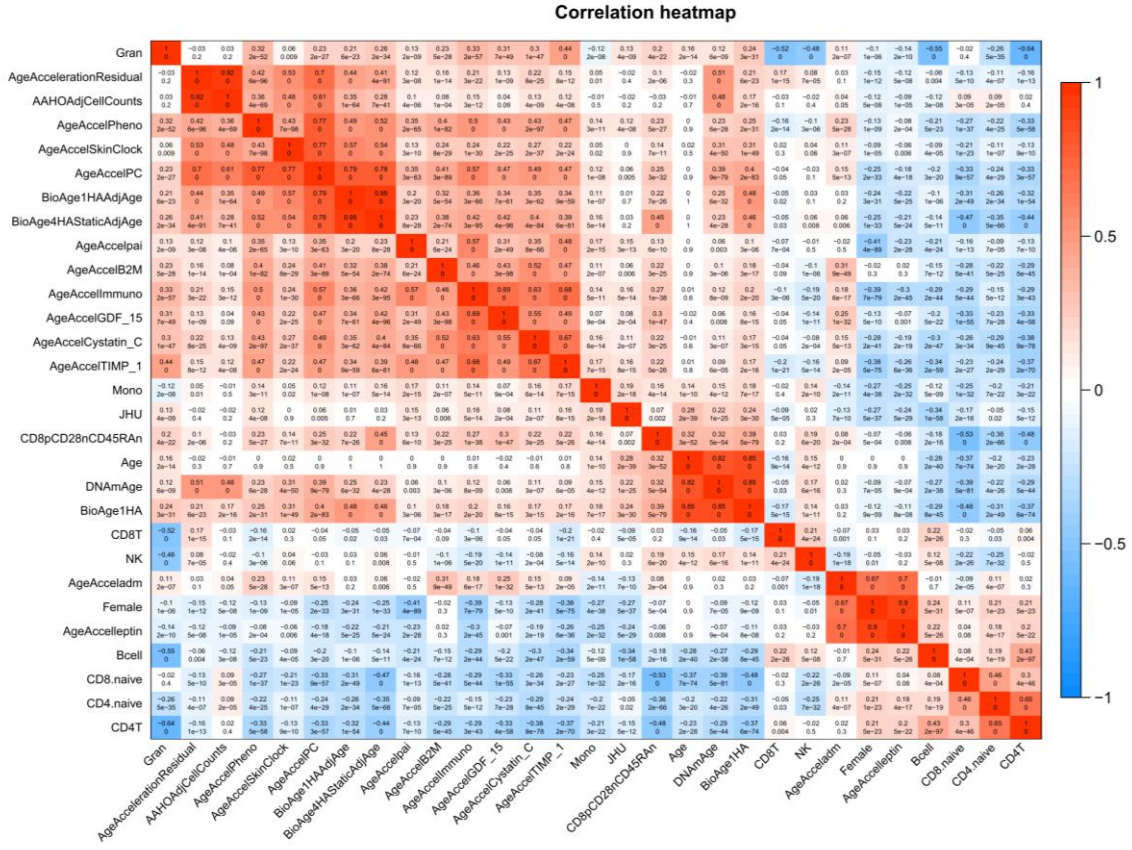
Characterizing epigenetic age acceleration in blood

Epigenetic age acceleration (EAA) measures, cell count, and biomarkers estimates based on DNA methylation were found to aggregate into distinct groups after correlation based hierarchical clustering. In the PBMC data, there are appear to be major five clusters corresponding roughly to epigenetic age acceleration measures, plasma biomarker surrogates, age, sex, and non-granulocyte cell count estimates (**Figure 3-2**). There are a number of interesting observations that can be noted from this analysis. Granulocytes, and to a lesser extend

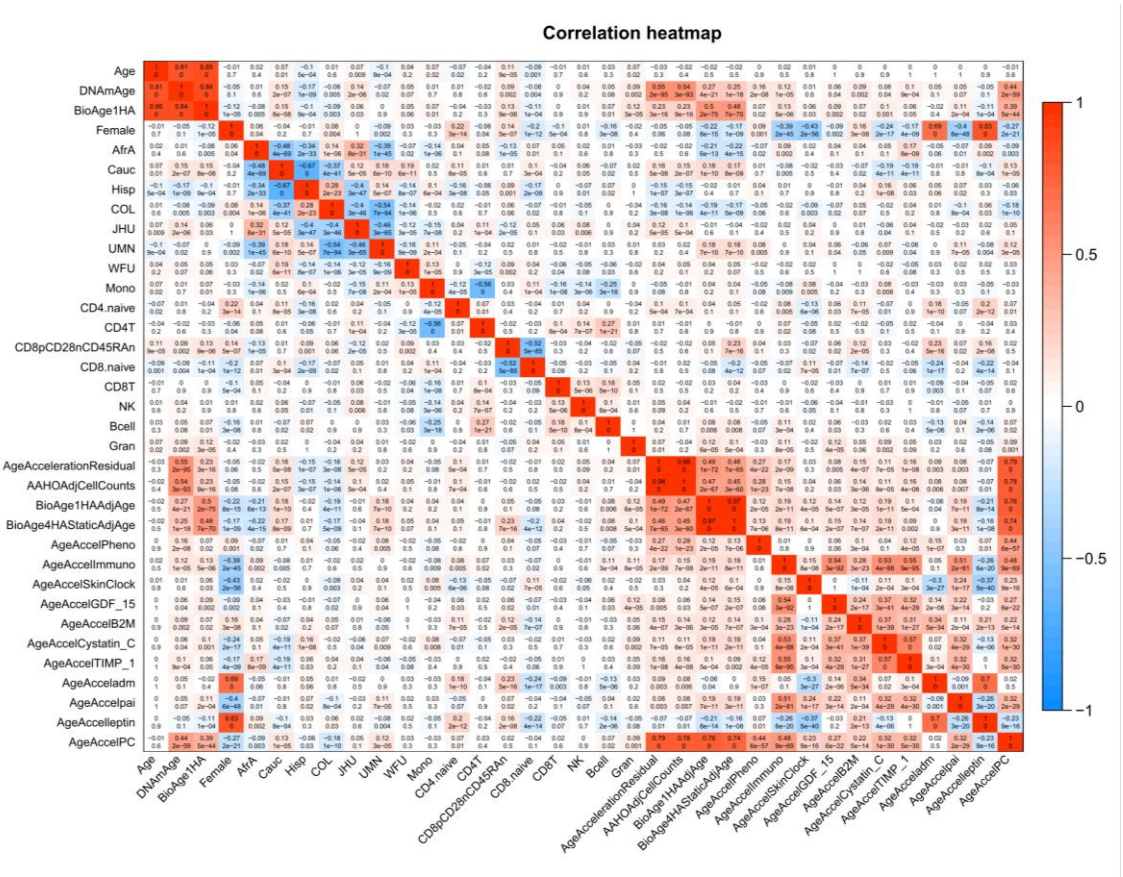
monocytes and exhausted CD8 T cells, cluster together with the pro-aging block, whereas B cells, naive and normal CD8 and CD4 T cells cluster more closely with the anti-aging block. Female status is negatively associated with age acceleration measures, as do leptin and adrenomedullin surrogate biomarkers which exhibit sex-based differences [149, 150]. Finally, all epigenetic aging measures share similar correlation profiles with other sample characteristics however they vary in terms of their strength, from most to least "reactive": AgeAccelPC, AgeAccelImmuno, AgeAccelPheno, BioAge1HAAdjAge/BioAge4HAStaticAdjAge, AgeAccelSkinClock, and AgeAccelResidual/AAHOAdjCellCounts.

The monocyte dataset shows similar patterns except with decreased effect sizes. Again, there are five blocks corresponding roughly to sex, cell counts, plasma biomarkers, age, and epigenetic age acceleration estimates. Though cell count estimates do form a weak cluster, they are not associated with the other clusters and the clustering is likely due to residual cell contamination after the monocyte isolation procedure. Overall, these results suggest that epigenetic age acceleration in PBMCs is associated with a variety of factors including DNAm signatures of cell composition, sex, and plasma biomarkers, and that these associations are preserved to some extent when examining a purified sample of monocytes (except cell count associations).

Figure 3-2. Associations between sample characteristics and measures of epigenetic aging. Labeled correlation heatmaps are presented for the FHS PBMC dataset (**A**) and the MESA monocyte dataset (**B**). Sample characteristics are listed in the rows and columns. Positive and negative correlation coefficients are colored red and blue with intensity being proportional to magnitude (color scale on right). Individual cells are labeled with correlation coefficients and p-values (above and below within each cell).



A



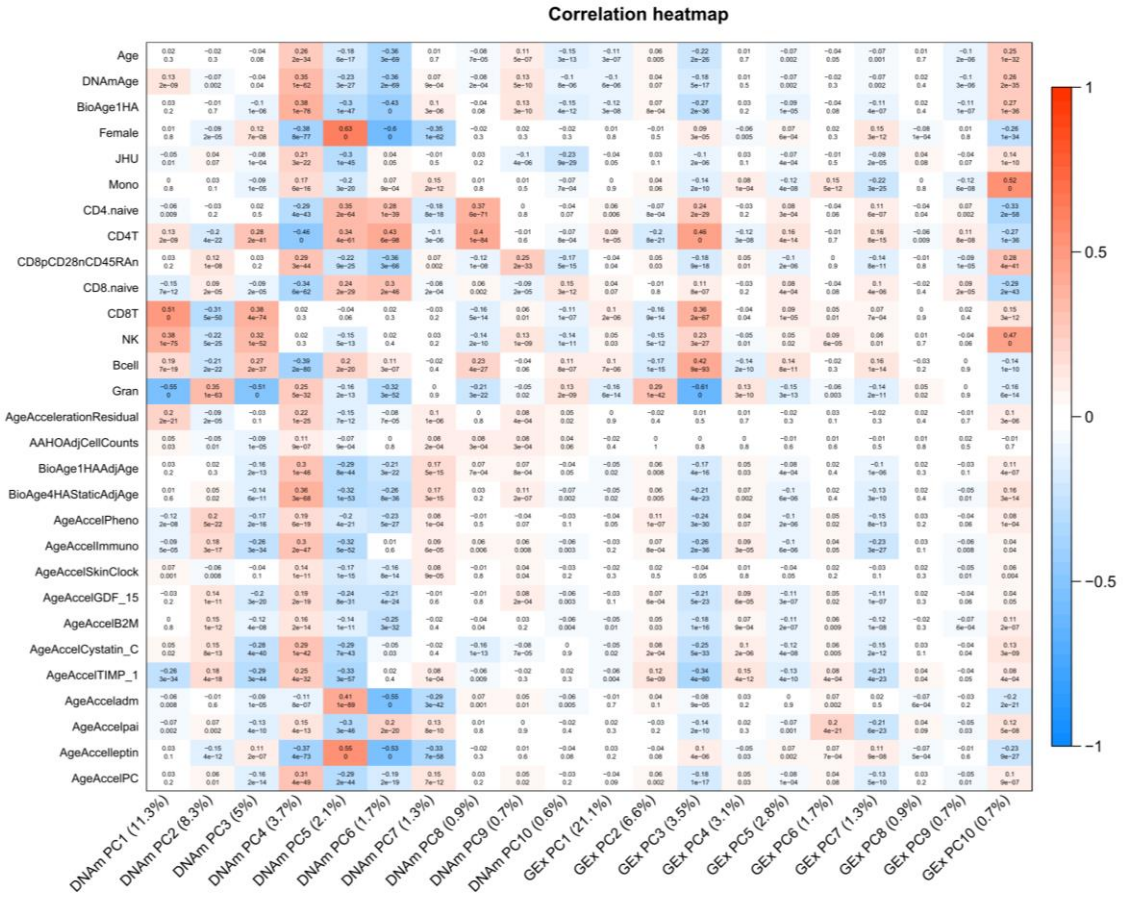
B

Transcriptomic analysis of the epigenetic age acceleration

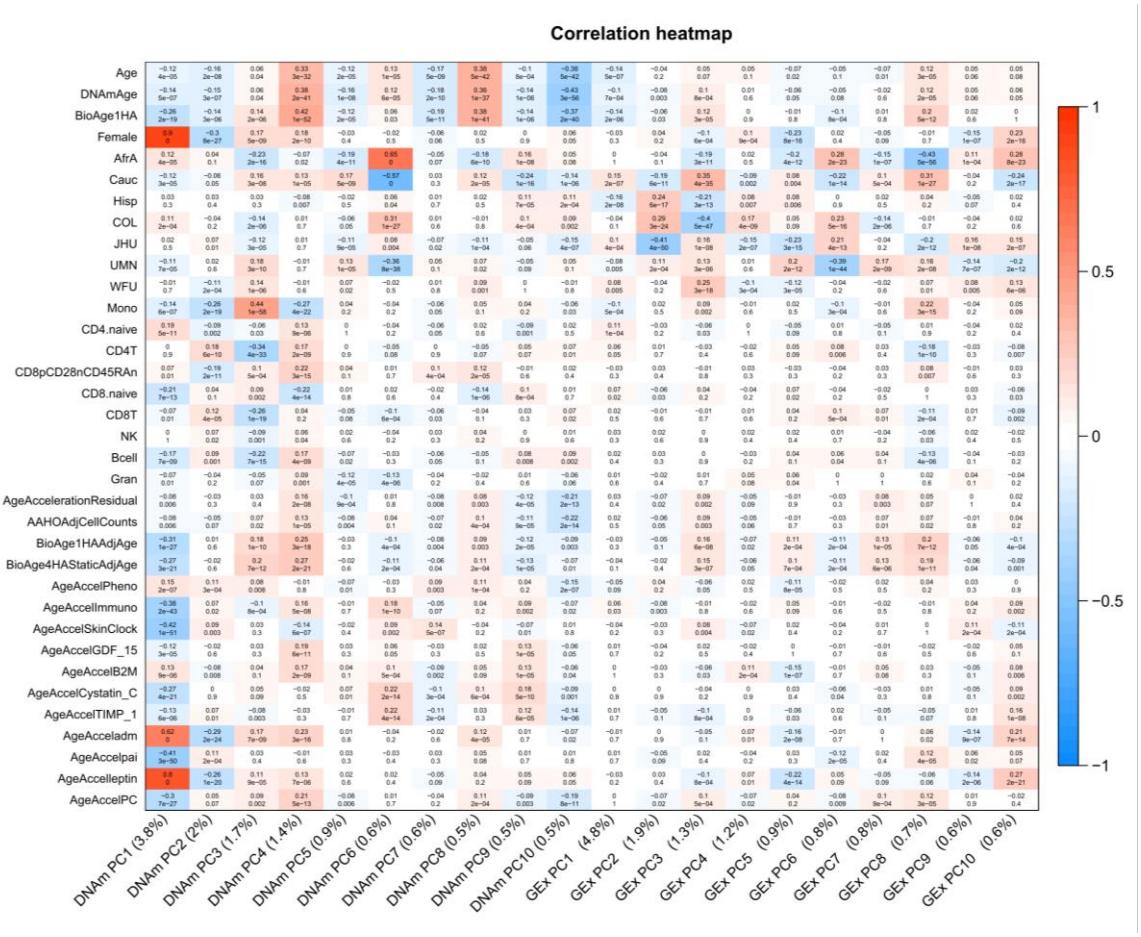
Principal Components Analysis

Examining the principal components (**PCs**) of the PBMC and monocyte datasets (**Figure 3-3**), we find that the PBMC data is more linearly compressible than the monocyte data both at the DNAm level (PC1 proportion of variance: 11.5% versus 3.8%) and at the GEx level (PC1 proportion of variance: 21.1% versus 4.8%). In the PBMC data, the first 3 DNAm PCs are moderately correlated with cell count estimates (r ~0.5), whereas PCs 4-6 appear to be related to age and sex (r~0.5). The PCs of the GEx data in the PBMCs have mostly weak correlations (r<0.3) with available variables with the exception of PC3 and PC10 which are correlated with cell counts (r~0.5).

Figure 3-3. Associations between sample characteristics and principal components. Labeled correlation heatmaps are presented for the PBMC dataset (A) and for the monocyte dataset (**B**). Sample characteristics are listed in the rows and DNAm and GEx PCs are listed in the columns. Proportions of variation captured by PCs are listed in in parentheses next to the column labels at the bottom. Positive and negative correlation coefficients are colored red and blue with intensity being proportional to magnitude (color scale on right). Individual cells are labeled with correlation coefficients and p-values (above and below within each cell).



A



B

The DNAm PCs of the monocyte data appear to reflect sex (PC1), residual non-monocyte cell contamination (PC3), race (PC6), and aging (PC4, PC8-10). The GEx PCs 2, 3, 6, and 8 are moderately correlated with race and study site with the remaining PCs only exhibiting weak correlations with available variables.

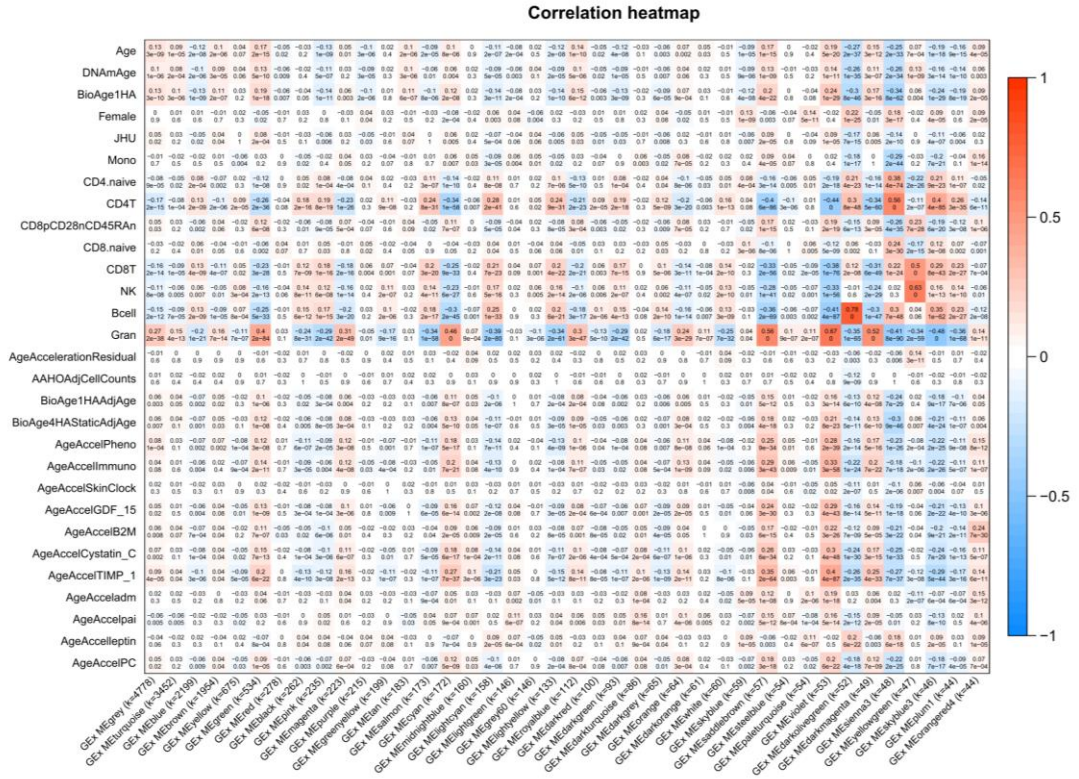
Overall these results demonstrate that the PCs of DNAm tend to capture global factors such as cell composition, age, sex, and race, whereas the PCs of GEx have little or no correlations with available variables. Comparing the PBMC and monocyte results in DNAm, cell composition effects dominate in the blood cell mixture whereas clearer sex and race signals are seen in the purified cell samples which in line with previous reports in the literature [97]. GEx data can also reflect these factors though the associations are generally not as strong or as

common in the top PCs as in DNAm (GEx PC1 in both PBMC and monocytes is not associated with any recorded variables). Taken together, these results suggest that the DNAm and GEx data may be fundamentally capturing different types of information, with the majority of global covariance in GEx remaining unexplained.

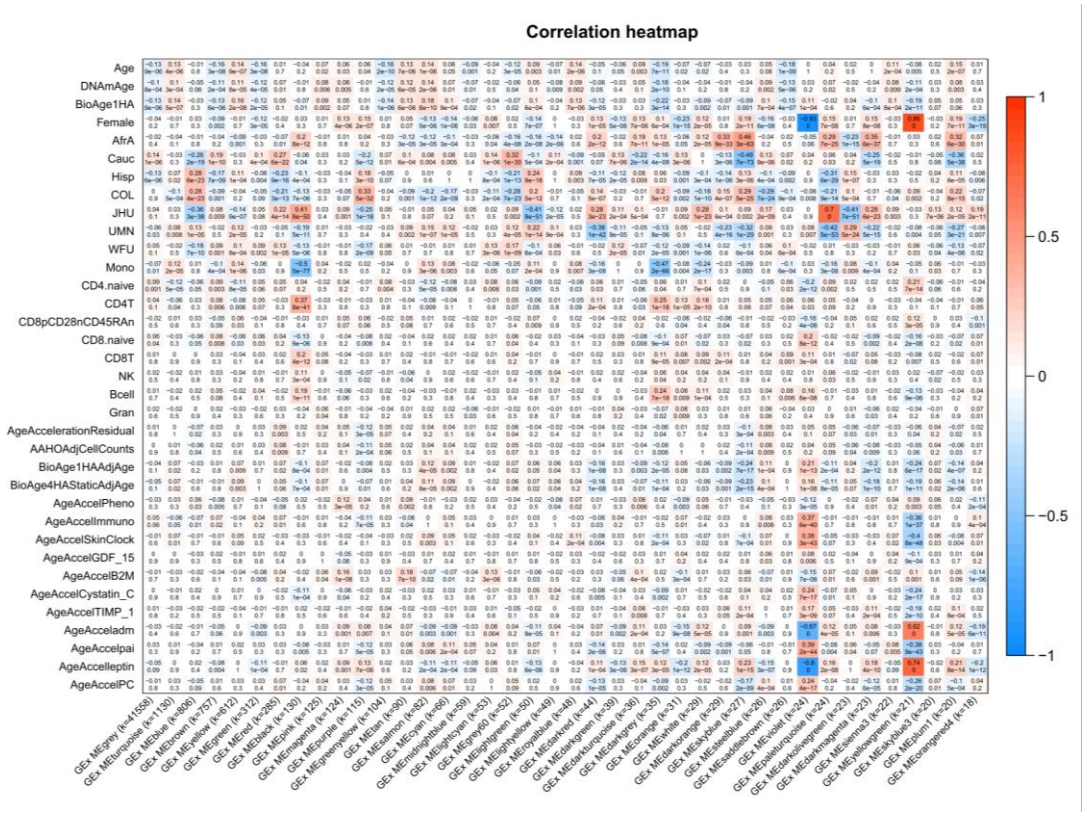
Coexpression and comethylation module analysis of epigenetic age acceleration

We also analyzed global GEx and DNAm using by constructing coexpression and comethylation modules and testing them for correlations with sample characteristics. We find that the majority of GEx modules have weak correlations with sample characteristics in PBMCs; the moderately strong correlations that do exist relate to cell count estimates (**Figure 3-4**). Though associations between modules and EAA measures are nominally significant, they appear to be reflections of cell type confounding as correlations mirrored across entire columns and the strongest within-column correlations are with cell count estimates. The monocyte coexpression modules similarly show weak correlations with apparent confounding attributable to sex, race, and residual cell contamination. Comethylation module analysis reveals similar results except with much stronger correlations with confounding factors. Overall, both the PCA and WGCNA analyses indicate that epigenetic aging associated with global transcriptomic and methylomic factors through confounding factors such as cell composition and sex.

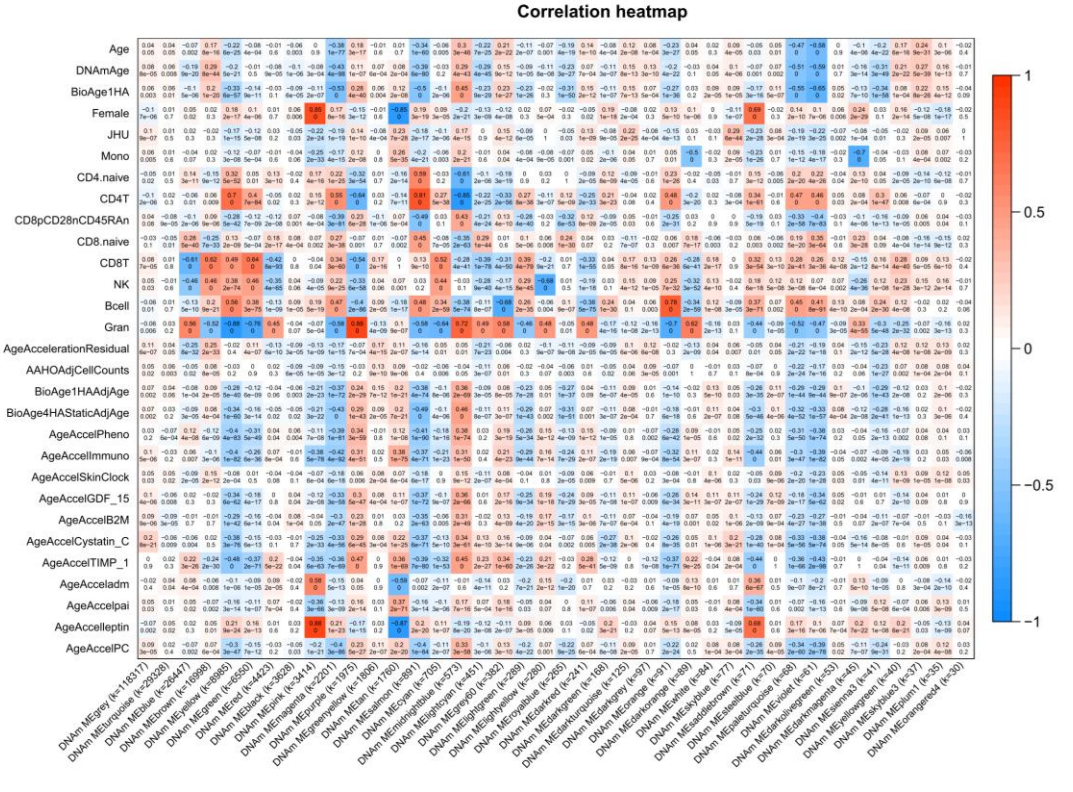
Figure 3-4. Associations between sample characteristics and WGCNA modules. Labeled correlation heatmaps are presented for the PBMC and monocyte datasets with coexpression (**A, B**) and comethylation modules (**C, D**). Sample characteristics are listed in the rows and GEx modules are listed in the columns. Modules are sorted from largest to smallest from left to right, with the number of module members being listed in parentheses after each module name. Positive and negative correlation coefficients are colored red and blue with intensity being proportional to magnitude (color scale on right). Individual cells are labeled with correlation coefficients and p-values (above and below within each cell).



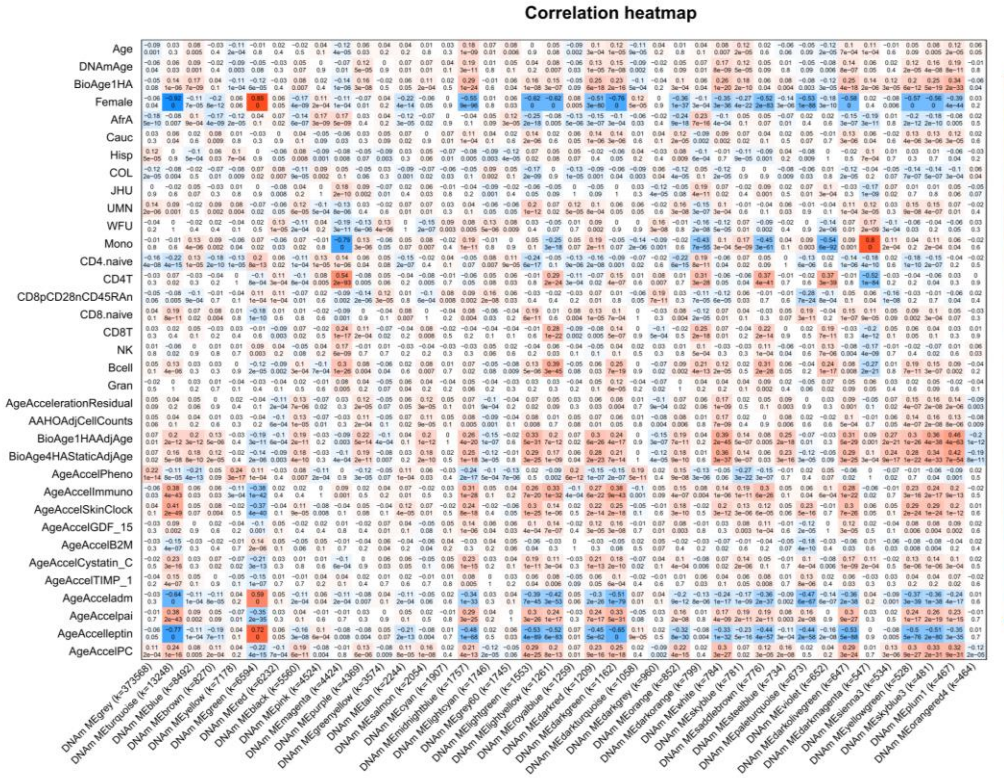
A


Correlation heatmap
Age
DNAmAge
BioAge1HA
Female
AfrA
Cauc
Hisp
COL
JHU
UMN
WFU
Mono
CD4.naive
CD4T
CD8pCD28nCD45RAn
CD8.naive
CD8T
NK
Bcell
Gran
AgeAccelerationResidual
AAHOAdjCellCounts
BioAge1HAAdjAge
BioAge4HAStaticAdjAge
AgeAccelPheno
AgeAccelImmuno
AgeAccelSkinClock
AgeAccelGDF_15
AgeAccelB2M
AgeAccelCystatin_C
AgeAccelTIMP_1
AgeAcceladm
AgeAccelpai
AgeAccelleptin
AgeAccelPC
GEx MEgrey (n=41558)
GEx MEturquoise (n=1130)
GEx MEblue (n=806)
GEx MEbrown (n=757)
GEx MEyellow (n=612)
GEx MEgreen (n=312)
GEx MEred (n=285)
GEx MEblack (n=130)
GEx MEpink (n=125)
GEx MEmagenta (n=124)
GEx MEpurple (n=115)
GEx MEgreenyellow (n=104)
GEx MEtan (n=90)
GEx MEsalmon (n=82)
GEx MEcyan (n=66)
GEx MEmidnightblue (n=59)
GEx MElightcyan (n=53)
GEx MEgrey60 (n=52)
GEx MElightgreen (n=50)
GEx MEroyalblue (n=49)
GEx MEdarkred (n=44)
GEx MEdarkgreen (n=39)
GEx MEdarkturquoise (n=36)
GEx MEdarkgrey (n=35)
GEx MEorange (n=31)
GEx MEdarkorange (n=29)
GEx MEwhite (n=29)
GEx MEskyblue (n=27)
GEx MEsaddlebrown (n=26)
GEx MEviolet (n=24)
GEx MEpaleturquoise (n=24)
GEx MEdarkolivegreen (n=23)
GEx MEdarkmagenta (n=23)
GEx MEsienna3 (n=22)
GEx MEyellowgreen (n=21)
GEx MEskyblue3 (n=20)
GEx MEplum1 (n=20)
GEx MEorangered4 (n=18)
1
0.5
0
−0.5
−1

B


Correlation heatmap
Age
DNAmAge
BioAge1HA
Female
JHU
Mono
CD4.naive
CD4T
CD8pCD28nCD45RAn
CD8.naive
CD8T
NK
Bcell
Gran
AgeAccelerationResidual
AAHOAdjCellCounts
BioAge1HAAdjAge
BioAge4HAStaticAdjAge
AgeAccelPheno
AgeAccelImmuno
AgeAccelSkinClock
AgeAccelGDF_15
AgeAccelB2M
AgeAccelCystatin_C
AgeAccelTIMP_1
AgeAcceladm
AgeAccelpai
AgeAccelleptin
AgeAccelPC
DNAm MEgrey (n=1183117)
DNAm MEturquoise (n=203287)
DNAm MEblue (n=264447)
DNAm MEbrown (n=169985)
DNAm MEyellow (n=89865)
DNAm MEgreen (n=65501)
DNAm MEred (n=44423)
DNAm MEblack (n=36281)
DNAm MEpink (n=34143)
DNAm MEmagenta (n=32201)
DNAm MEpurple (n=31975)
DNAm MEgreenyellow (n=31906)
DNAm MEtan (n=31760)
DNAm MEsalmon (n=30891)
DNAm MEcyan (n=30705)
DNAm MEmidnightblue (n=30573)
DNAm MElightcyan (n=30451)
DNAm MEgrey60 (n=30382)
DNAm MElightgreen (n=30289)
DNAm MEroyalblue (n=30280)
DNAm MEdarkred (n=30265)
DNAm MEdarkgreen (n=30168)
DNAm MEdarkturquoise (n=30125)
DNAm MEdarkgrey (n=30097)
DNAm MEorange (n=30049)
DNAm MEdarkorange (n=29968)
DNAm MEwhite (n=29841)
DNAm MEskyblue (n=29771)
DNAm MEsaddlebrown (n=29702)
DNAm MEviolet (n=29701)
DNAm MEdarkolivegreen (n=29681)
DNAm MEdarkmagenta (n=29553)
DNAm MEsienna3 (n=29445)
DNAm MEyellowgreen (n=29411)
DNAm MEskyblue3 (n=29340)
DNAm MEplum1 (n=29337)
DNAm MEorangered4 (n=29130)
1
0.5
0
−0.5
−1

C



D

Associations between the expression of individual genes and epigenetic aging

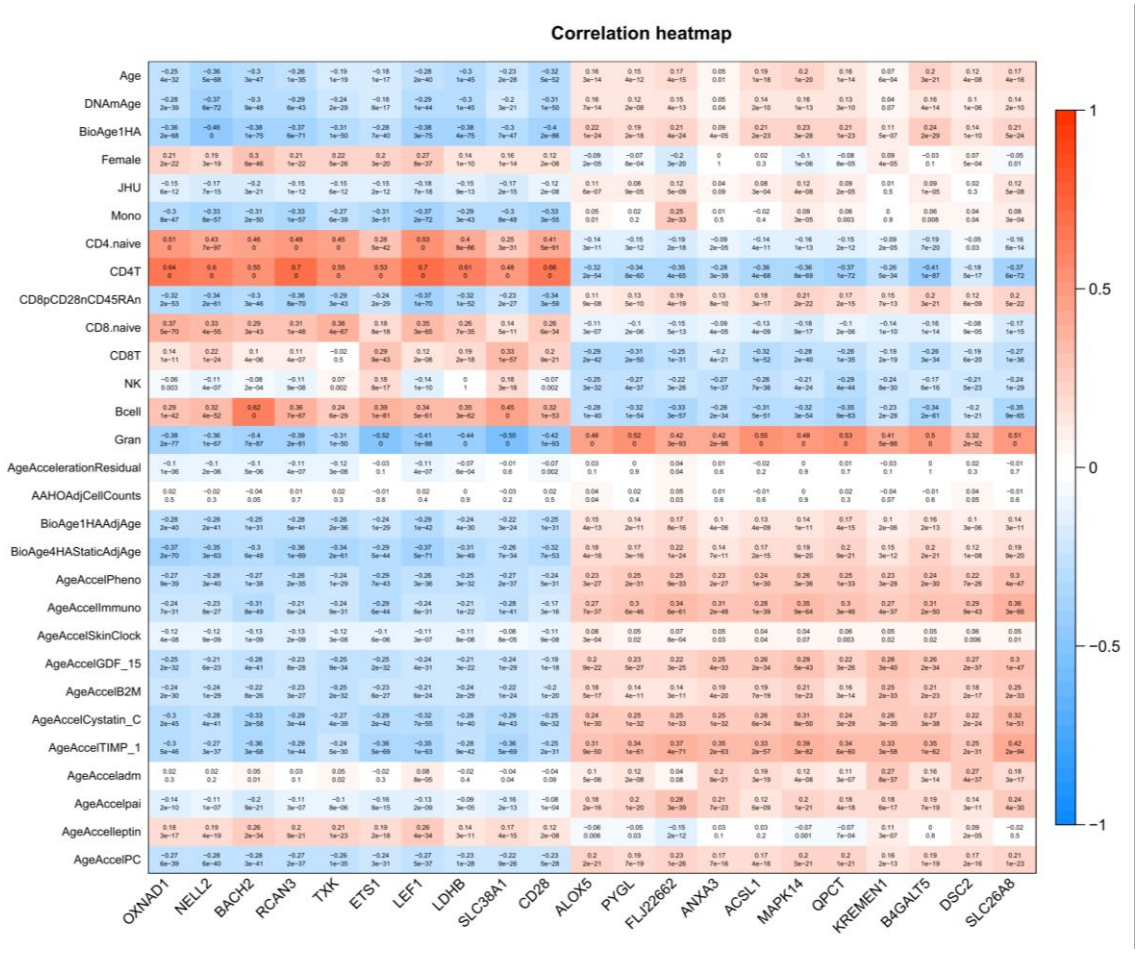
Although the above global analyses did not reveal associations between epigenetic aging measures which were not apparently attributable to confounding factors, we assessed whether individual transcripts might capture different types such relationships. PBMC and monocyte datasets were stratified by sex, study site, and race in both studies and individual GEx probes were tested for correlations with the following epigenetic age acceleration measures: AgeAccelerationResidual, BioAge1HAAdjAge, AgeAccelPheno, AgeAccelImmuno, and AgeAccelSkinClock. Transcript associations were averaged across all age acceleration measures, and were also combined across all strata by a weighted average (Stouffer's Z-score method weighting by sample size). The results of the screens from these two data sets are presented in **Table 3-1**; the correlation between these top genes and available sample characteristics is shown in **Figure 3-6**. In the PBMC dataset there are more than 20 genes which are significantly

associated with epigenetic age acceleration however these are likely the result of cell composition-based confounding. Though there are no significant correlations in the monocyte dataset, the most significant gene, desmocollin 2, was positively associated with epigenetic aging in both PBMCs and monocytes. To assess whether this association could be explained by confounding factors a composite measure of epigenetic age acceleration, AgeAccelPC, was regressed on DSC2 expression and adjusted for potential confounders including granulocyte count, sex, race, and study site and stratified based on sex and race (**Table 3-2**). DSC2 was positively associated with AgeAccelPC among most of these models with the exception of the African American stratum within the monocyte dataset.

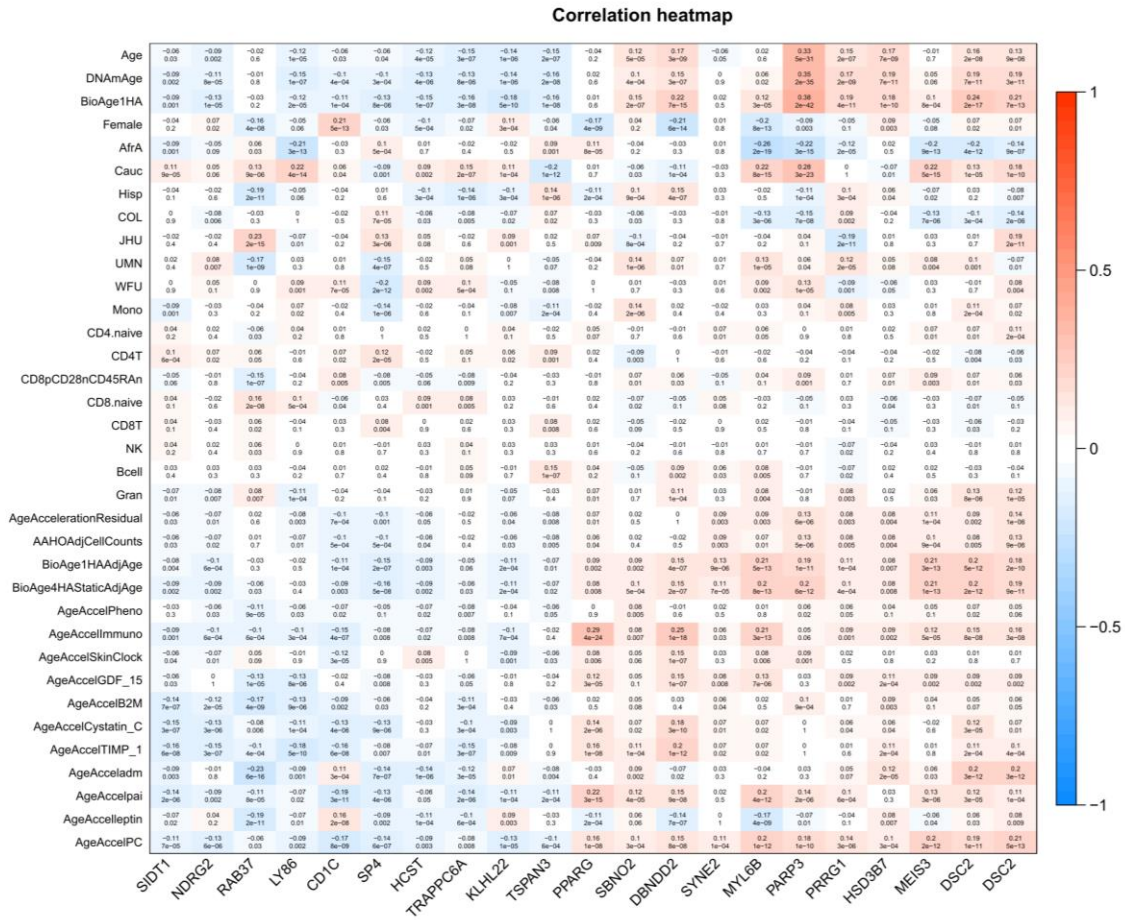
Table 3-1. Associations between the individual transcripts and epigenetic age acceleration. The top positive and negative associations are presented for the FHS peripheral leukocyte dataset (top) and the MESA monocyte dataset. Stouffer's Z-score method was used to combine Z-scores for gene correlations with the major epigenetic age acceleration measures and among groups stratified by sex, race, and site. Meta-analytic Z-scores and p-values are presented in red/blue and green respectively. The Bonferroni corrected significance threshold here is $p<5x10^{-7}$.

	Symbol	Name	meta Z	meta p	Location	Probe ID
Peripheral leukocyte dataset	SLC26A8	solute carrier family 26, member 8	7.86	4E-15	chr6:35998347-36104507	2951730
	DSC2	desmocollin 2	7.55	4E-14	chr18:26899404-26936375	3802980
	B4GALT5	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferas	7.39	2E-13	chr20:47682889-47809709	3908963
	KREMEN1	kringle containing transmembrane protein 1	7.06	2E-12	chr22:27799076-27894321	3941793
	QPCT	glutaminyl-peptide cyclotransferase	7.00	2E-12	chr2:37423494-37496720	2477438
	MAPK14	mitogen-activated protein kinase 14	6.86	7E-12	chr6:36103487-36186989	2904877
	ACSL1	acyl-CoA synthetase long-chain family member 1	6.85	8E-12	chr4:185913758-186002178	2796553
	ANXA3	annexin A3	6.82	9E-12	chr4:79691717-79774504	2732844
	FLJ22662	hypothetical protein FLJ22662	6.79	1E-11	chr12:14547884-14612382	3445544
	PYGL	phosphorylase, glycogen, liver	6.72	2E-11	chr14:50441691-50480984	3564210
	ALOX5	arachidonate 5-lipoxygenase	6.70	2E-11	chr10:45189658-45261565	3244622
	CD28	CD28 molecule	-7.01	2E-12	chr2:204279443-204310801	2523801
	SLC38A1	solute carrier family 38, member 1	-7.20	6E-13	chr12:44863120-44952390	3452231
	LDHB	lactate dehydrogenase B	-7.34	2E-13	chr12:21679576-21802038	3446868
	LEF1	lymphoid enhancer-binding factor 1	-7.40	1E-13	chr4:109186451-109348127	2781138
	ETS1	v-ets erythroblastosis virus E26 oncogene homolog	-7.45	9E-14	chr11:127833879-127963056	3397589
	TXK	TXK tyrosine kinase	-7.54	5E-14	chr4:47762719-47831030	2768354
	RCAN3	RCAN family member 3	-7.63	2E-14	chr1:24701913-24740112	2325479
	BACH2	BTB and CNC homology 1, basic leucine zipper tra	-7.80	6E-15	chr6:90692975-91063182	2964553
	NELL2	NEL-like 2 (chicken)	-7.99	1E-15	chr12:43057588-43557137	3451814
	OXNAD1	oxidoreductase NAD-binding domain containing 1	-8.02	1E-15	chr3:16281732-16340969	2612625
Monocyte dataset	DSC2	desmocollin 2	4.65	3E-06	chr18:28646112:28646161:-	ILMN_1663119
	DSC2	desmocollin 2	4.21	3E-05	chr18:28649058:28649107:-	ILMN_2381257
	MEIS3	Meis homeobox 3 pseudogene 1	3.02	3E-03	chr17:15692693:15692742:+	ILMN_2205896
	HSD3B7	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and	2.88	4E-03	chr16:31000060:31000109:+	ILMN_1653042
	PRRG1	proline rich and Gla domain 1	2.78	6E-03	chrX:37316174:37316223:+	ILMN_1781791
	PARP3	poly(ADP-ribose) polymerase family member 3	2.71	7E-03	chr3:51982800:51982849:+	ILMN_2397954
	MYL6B	myosin light chain 6B	2.70	7E-03	chr12:56551486:56551519:+	ILMN_1713450
	SYNE2	spectrin repeat containing nuclear envelope protein	2.62	9E-03	chr14:64682014:64682063:+	ILMN_1677009
	DBNDD2	NA	2.58	1E-02	chr20:44039163:44039212:+	ILMN_1730612
	SBNO2	strawberry notch homolog 2	2.56	1E-02	chr19:1107866:1107915:-	ILMN_1808811
	PPARG	peroxisome proliferator activated receptor gamma	2.53	1E-02	chr3:12475653:12475702:+	ILMN_1800225
	TSPAN3	tetraspanin 3	-2.46	1E-02	chr15:77348148:77348197:-	ILMN_1655469
	KLHL22	kelch like family member 22	-2.52	1E-02	chr22:20795871:20795920:-	ILMN_1705390
	TRAPPC6A	trafficking protein particle complex 6A	-2.52	1E-02	chr19:45666264:45666313:-	ILMN_1775703
	HCST	hematopoietic cell signal transducer	-2.53	1E-02	chr19:36395117:36395166:+	ILMN_2396991
	SP4	Sp4 transcription factor	-2.54	1E-02	chr7:21553605:21553654:+	ILMN_1721081
	CD1C	CD1c molecule	-2.68	7E-03	chr1:158263269:158263318:+	ILMN_1654210
	LY86	lymphocyte antigen 86	-2.73	6E-03	chr6:6655008:6655057:+	ILMN_1807825
	RAB37	RAB37, member RAS oncogene family	-2.74	6E-03	chr17:72743276:72743325:+	ILMN_2255579
	NDRG2	NDRG family member 2	-2.80	5E-03	chr14:21485404:21485453:-	ILMN_2361603
	SIDT1	SID1 transmembrane family member 1	-2.92	4E-03	chr3:113347803:113347852:+	ILMN_1795118

Figure 3-6. Association between individual gene transcripts and epigenetic age acceleration. Labeled correlation heatmaps are presented for the 10 most significantly associated genes with epigenetic aging measures and gene transcript levels for the PBMC (**A**) and monocyte datasets (**B**). Sample characteristics are listed in the rows and Positive and negative correlation coefficients are colored red and blue with intensity being proportional to magnitude (color scale on right). Individual cells are labeled with correlation coefficients and p-values (above and below within each cell).



A


Correlation heatmap
Age
DNAmAge
BioAge1HA
Female
AfrA
Cauc
Hisp
COL
JHU
UMN
WFU
Mono
CD4.naive
CD4T
CD8pCD28nCD45RAn
CD8.naive
CD8T
NK
Bcell
Gran
AgeAccelerationResidual
AAHOAdjCellCounts
BioAge1HAAdjAge
BioAge4HAStaticAdjAge
AgeAccelPheno
AgeAccelImmuno
AgeAccelSkinClock
AgeAccelGDF_15
AgeAccelB2M
AgeAccelCystatin_C
AgeAccelTIMP_1
AgeAcceladm
AgeAccelpai
AgeAccelleptin
AgeAccelPC
SIDT1
NDRG2
RAB37
LY86
CD1C
SP4
HCST
TRAPPC6A
KLHL22
TSPAN3
PPARG
SBNO2
DBNDD2
SYNE2
MYL6B
PARP3
PRRG1
HSD3B7
MEIS3
DSC2
DSC2
1
0.5
0
−0.5
−1

B

Table 3-2. Association between desmocollin 2 expression and AgeAccelPC. AgeAccelPC was regressed on desmocollin 2 (DSC2) expression levels in models stratified based on sex and race while adjusting for study site. University of Minnesota served as reference for the PBMC dataset. Caucasians and Columbia University (COL) served as the reference race and site for the monocyte dataset. JHU = John Hopkins University, UMN = University of Minnesota Twin Cities, WFU = Wake Forest University. Associations from linear models are represented by t-values (colored red and blue for positive and negative associations) and p-values (colored green for significance).

	AgeAccelPC models																	
	Peripheral leukocytes						Purified monocytes											
	All		Female		Male		All		Female		Male		Caucasian		AfricanAm		Hispanic	
n	2188		1191		997		1202		606		596		582		234		386	
	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p	t	p
DSC2	6.33	3E-10	5.14	3E-07	3.72	2E-04	7.71	3E-14	5.51	5E-08	5.46	7E-08	6.03	3E-09	0.75	5E-01	4.97	1E-06
Granulocytes	3.38	7E-04	2.20	3E-02	2.66	8E-03	4.54	6E-06	2.33	2E-02	3.81	2E-04	2.11	4E-02	1.30	2E-01	4.38	2E-05
Female	-10.52	3E-25					-10.04	8E-23					-6.43	3E-10	-1.01	3E-01	-6.15	2E-09
AfricanAm							-0.12	9E-01	-1.03	3E-01	0.83	4E-01						
Hispanic							-1.16	2E-01	-1.12	3E-01	-0.60	5E-01						
JHU	-0.27	8E-01	0.81	4E-01	-0.98	3E-01	1.98	5E-02	1.54	1E-01	1.33	2E-01	2.08	4E-02	0.84	4E-01		
UMN							4.69	3E-06	2.41	2E-02	4.08	5E-05	3.54	4E-04			3.36	9E-04
WFU							1.51	1E-01	1.24	2E-01	0.76	4E-01	1.63	1E-01	0.87	4E-01		

Desmocollin 2 is a protein which participates in the formation of desmosomes, a type of cell-cell junction. Mutations in this gene are associated with cardiomyopathogenic risk including conditions such as arrhythmogenic right ventricular cardiomyopathy. Expression of DSC2 has previously been found to be associated with increased epigenetic age acceleration in acute myeloid leukemia by RNA-Seq [151]. Coincidentally, a supercentenarian was found to carry a mutation in these gene with no apparent cardiomyopathogenic effects [152]. Altogether, these results suggest that DSC2 is associated with epigenetic aging though it is still unclear through what molecular mechanism it might exert its effects.

GO term enrichment of genes associated with epigenetic aging

The genes most associated with epigenetic aging were tested for enrichment in specific biological processes by GO term analysis (**Figure 3-7**). Both in the PBMC and monocyte data, genes positively associated epigenetic aging were enriched for functions related to interferon

signaling including type I interferon signaling pathway (Fisher $p=3x10^{-7}$, $8x10^{-10}$ for PBMC and monocyte data respectively), and interferon-gamma-mediated signaling pathway (Fisher $p=2x10^{-8}$, $3x10^{-8}$). Genes negatively associated with epigenetic aging were enriched for processes related to translational initiation including viral transcription (Fisher $p=3x10^{-10}$, $p<1x10^{-30}$), rRNA processing (Fisher $p=3x10^{-10}$, $p<1x10^{-30}$), translational initiation (Fisher $p=2x10^{-8}$, $p<1x10^{-30}$), SRP-dependent cotranslational protein targeting to membrane (Fisher $p=6x10^{-8}$, $p<1x10^{-30}$), and nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (Fisher $p=6x10^{-8}$, $2x10^{-29}$).

Interferon signaling is an endogenous cell-intrinsic inflammatory response to a variety of stresses including viral infection and genomic damage. This pathway appears to reduce viral proliferation and oncogenic risk by blocking major elements of gene expression and promoting cellular senescence. The GO term enrichment results described above are consistent with the growing body of literature describing the relationship between biological aging and interferon signaling. Interferon gamma expression has been reported to be induced by DNA damage and to promote senescence [153]. Further, progerin-induced replication stress found in Hutchinson-Gilford progeria syndrome has been linked to activated interferon-like cellular signaling [154]. Interferon signaling has also been associated with aging in the choroid plexus brain tissue, and its activity has been shown to result in cognitive impairment in mice [155]. These studies report that inhibition of the interferon signaling partially rescues progeric phenotypes, suggesting this pathway may play a causal role in biological aging.

The negative association between genes involved in translational initiation and epigenetic aging may be the signature of genes anti-correlated with interferon signaling. Replicative stress has been reported to decrease ribosomal RNA processing resulting in induction of senescence;

exogenous expression of rRNA processing genes was found to extend replicative lifespan [156]. Likewise, expression of genes related to translational initiation and viral transcription may indicate the absence of interferon signaling as it is associated translational arrest [157]. Altogether, the GO term enrichment analysis of gene expression suggests that epigenetic aging in leukocytes may primarily reflect interferon-mediated silencing of gene expression machinery and induction of cellular senescence.

Figure 3-7. GO term enrichment of genes most associated with epigenetic aging. Top genes positively and negatively associated with epigenetic age acceleration are tested for enrichment of GO term annotations. The top ten GO terms are listed here for positive (top) and negative (bottom) associations in the PBMC data (left) and in the monocyte data (right).

	Peripheral leukocytes								Purified monocytes							
	GO ID	Term	Annotated	Significant	Expected	Rank in KS	Fisher p	KS p	GO ID	Term	Annotated	Significant	Expected	Rank in KS	Fisher p	KS p
Positively associated	GO:0045087	innate immune response	690	120	38.08	12	2E-08	3E-07	GO:0060337	type I interferon signaling pathway	138	30	8.07	1	8E-10	7E-09
	GO:0060333	interferon-gamma-mediated signaling pathway	73	19	4.03	21	2E-08	8E-06	GO:0051607	defense response to virus	399	59	23.33	3	2E-08	4E-06
	GO:0060337	type I interferon signaling pathway	70	14	3.86	8	3E-07	4E-08	GO:0060333	interferon-gamma-mediated signaling pathway	153	31	8.95	4	3E-08	2E-05
	GO:0006955	immune response	1332	193	73.5	114	3E-07	1E-03	GO:0007259	JAK-STAT cascade	271	31	15.85	68	6E-06	2E-03
	GO:0002755	MyD88-dependent toll-like receptor signaling pa	31	11	1.71	35	4E-07	3E-05	GO:0035457	cellular response to interferon-alpha	15	7	0.88	35	1E-05	7E-04
	GO:0042742	defense response to bacterium	183	33	10.1	74	4E-07	4E-04	GO:0006958	complement activation, classical pathway	50	12	2.92	323	2E-05	2E-02
	GO:0006954	inflammatory response	610	93	33.66	3	1E-06	8E-11	GO:0035456	response to interferon-beta	40	9	2.34	6	5E-05	3E-05
	GO:0071260	cellular response to mechanical stimulus	66	15	3.64	26	2E-06	2E-05	GO:0039530	MDA-5 signaling pathway	13	6	0.76	21	5E-05	4E-04
	GO:0050707	regulation of cytokine secretion	131	32	7.23	120	4E-06	1E-03	GO:0045916	negative regulation of complement activation	13	6	0.76	80	5E-05	3E-03
	GO:0051607	defense response to virus	209	28	11.53	88	5E-06	5E-04	GO:0071901	negative regulation of protein serine/threonine k	226	26	13.22	163	1E-04	8E-03
Negatively associated	GO:0019083	viral transcription	138	27	7.35	4	3E-10	2E-12	GO:0006364	rRNA processing	753	130	42.25	5	1E-30	1E-30
	GO:0006364	rRNA processing	209	46	11.13	3	3E-10	6E-14	GO:0006614	SRP-dependent cotranslational protein targetin	488	105	27.38	1	1E-30	1E-30
	GO:0006355	regulation of transcription, DNA-templated	3139	227	167.2	1	1E-09	8E-27	GO:0019083	viral transcription	618	112	34.67	2	1E-30	1E-30
	GO:0006413	translational initiation	144	29	7.67	10	2E-08	1E-08	GO:0006413	translational initiation	682	120	38.26	3	1E-30	1E-30
	GO:0006614	SRP-dependent cotranslational protein targetin	60	16	3.2	14	6E-08	3E-08	GO:0000184	nuclear-transcribed mRNA catabolic process, r	533	104	29.9	4	2E-29	1E-30
	GO:0000184	nuclear-transcribed mRNA catabolic process, r	84	19	4.47	8	6E-08	1E-09	GO:0000027	ribosomal large subunit assembly	82	22	4.6	7	5E-10	6E-12
	GO:0006376	mRNA splice site selection	25	10	1.33	43	3E-07	3E-05	GO:0006283	transcription-coupled nucleotide-excision repair	135	26	7.57	14	3E-08	1E-05
	GO:0000398	mRNA splicing, via spliceosome	274	45	14.59	2	1E-05	4E-18	GO:0006296	nucleotide-excision repair, DNA incision, 5'-to le	73	16	4.1	27	2E-06	2E-04
	GO:0031295	T cell costimulation	69	14	3.68	149	1E-05	4E-03	GO:0042769	DNA damage response, detection of DNA dam	73	16	4.1	16	2E-06	3E-05
	GO:0050852	T cell receptor signaling pathway	153	25	8.15	147	3E-05	4E-03	GO:0002181	cytoplasmic translation	186	29	10.44	6	3E-06	5E-17

Limitations

In this study we use DNA methylation measurements to estimate a range of surrogates including epigenetic age, cell composition, and biomarker levels. In substituting real sample measurements for epigenetic signatures, it is possible that the results of this study could be confounded by artifacts in DNA methylation, in its measurement, or in the estimation of these variables. Similarly, though we attempt to analyze the transcriptomic data using robust methods, adjusting for confounding factors, and validating in two independent datasets, there is still a possibility of

residual confounding given the strong associations between epigenetic aging and cell counts and sex.

Conclusions

To my knowledge, this is the first study on relationship between epigenetic age acceleration and genome-wide transcription in normal leukocytes. The results from this study reinforce previous findings which suggest that epigenetic aging reflects a multifactorial process encompassing cell composition (granulocytes), systemic signaling (plasma biomarkers), sex, and race. Additionally in analyzing transcriptomic data from both mixed and isolated blood leukocytes, we find that epigenetic age acceleration is significantly associated with expression levels of the gene desmocollin 2 and genes involved in interferon signaling. Overall, this study elucidates the interferon pathway as a potential intermediary between inflammation, cellular senescence, and aging of the epigenome.

Chapter 4: Towards a universal molecular assay

ABSTRACT

Global measurements using nucleic acid based technologies have been widely adopted by the research community, substantially accelerating scientific progress. In contrast, untargeted measurement of other types of biochemicals remains relatively unpopular. This may in part be due to highly specialized nature of these analyses, with the measurement of each class often requiring a separate laboratory methods.

To simplify this process, I developed a new analytical method using liquid chromatography coupled to high resolution mass spectrometry to jointly quantify proteins, lipids, metabolites, and electrolytes. I use the assay to qualitatively analyze a diverse range of samples including plasma, urine, cells, muscle, adipose, bone marrow, blood vessel, and tendon samples. I demonstrate that this method has quantitative reproducibility with smaller intra-sample versus inter-sample variation and is able to distinguish between the measurements of different samples for representative analytes. I also report detection of short oligonucleotides after treatment with of cell samples with an endoribonuclease, suggesting the possibility of incorporating transcriptomics into the assay; with further development the integration of glycomics and even genomics might be possible using similar bottom-up strategies.

Overall, this work demonstrates that contrary to conventional wisdom, the analysis of diverse chemical species using a single method is not only feasible but practical. The capability of integrating multiple complementary sources of bioinformation presents a variety of unique opportunities in biomedical research and practice.

INTRODUCTION

A common shortcoming in global -omics studies is the lack of sample annotations. These labels are often the crux of molecular studies and are particularly important for abstract data types which are not readily interpretable; without being linked to phenotypic observations or prior knowledge, such data exists in a vacuum. For example, the identification of mutations in a putative gene has extremely limited scientific value without characterization of that gene and/or the establishment of the biological significance of the mutations. Though high quality sample annotations such as quantitative measurements are desirable, they are typically expensive and/or difficult to collect, thus the availability of this type of data is usually limited. Sample annotations can be acquired using lower quality instruments such as self-report however it is well-established that these types of methods entail compromises in accuracy and reliability. For example, in the study presented in Chapter 1, reported intake of fruits and vegetables was not significantly associated with epigenetic aging whereas measured plasma carotenoids levels were, despite the latter being a measure of the former. In addition, missing sample annotations can also lead to confounding, invalidating the entire studies if left unaddressed.

One potential solution to the lack of high quality sample labeling may lie in the global quantification of metabolites and proteins. Metabolite and protein measurements make up the majority of clinical laboratory tests conducted [158], and are used in the diagnosis of a large number of clinical conditions including vitamin deficiencies, organ function (liver, kidney, thyroid), hormonal abnormalities, cardiometabolic risk, and cancers. Additionally metabolite biomarkers have been established and compiled for a number of lifestyle and environmental exposures such as food [159], tobacco [160], alcohol [161], drugs, medications, pesticides [162], and various pollutants [163]. The ability to capture even a modest fraction of these

variables using a single procedure could substantially strengthen our ability to conduct rigorous and informative studies of genomics and other abstract data types.

Mass spectrometry stands out as a preferred analytical technology because it can be highly sensitive and specific. Liquid chromatography-mass spectrometry (**LC-MS**) has become the primary tool in both the fields of metabolomics and proteomics; however current methods are specialized with separate laboratory procedures for classes and subclasses of analytes. This hyper-specialization trend continues with new methods being developed for increasingly specific analyses, e.g. lipidomics, phosphoproteomics, glycomics, etc. An ideal LC-MS method would capture as much biochemical information as possible for as little cost as possible; to address this perceived missing capability, I asked whether it would be possible to develop a method to analyze all classes of molecules simultaneously.

LC-MS operates via two major principles: separation of compounds based on liquid-phase chromatographic chemistry and separation in a vacuum based on molecular weight and charge. In liquid chromatography, a mixture of molecules is separated based on their differences in attraction to the stationary phase—as molecules flow through a chromatography column they separate based how much time they spend in the flowing solution versus how much time they spend adhered to the stationary material in the column. In electrospray ionization, after the molecules are washed out of the column they are aerosolized in a high voltage electric field and a proportion of them become charged by gaining or losing charged particles such as protons. These charged molecules (ions) are then separated based on their mass-to-charge ratio by the mass spectrometer using electric and/or magnetic fields and detected by an ion sensor. The mass resolution of modern mass spectrometers allows us to provide near-unequivocal identification of ions based on their exact mass and fragmentation patterns. In the remainder of this chapter, I

describe my exploration of the feasibility of using this powerful technology to develop an all-inclusive assay.

METHODS

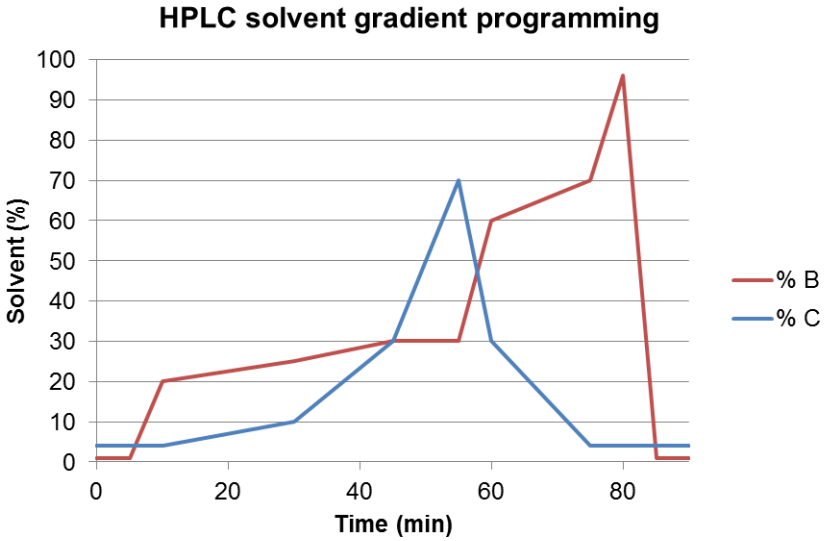
Sample preparation

Samples are prepared by first adding methanol (1 sample volume) to denature proteins and permeabilize membranes, and then a digestion solution (1.5 sample volumes) is added to cleave proteins into peptide fragments which are more amenable to LC-MS analysis. The final concentrations of the additives in the digestion are 1:20 trypsin:protein by mass, 5mM EDTA, and 50 mM ammonium bicarbonate, pH 7.8. Following the proteolytic digestion at 37°C for 2 to 12 hours with gentle agitation, the sample is treated acetonitrile and acetone (1 sample volume each) to precipitate undigested protein and other cell debris components [164]. The samples are allowed to come to solubility equilibration for 1 hour at room temperature, centrifuged at 15,000 x g for 5 minutes to pellet the insoluble material, and the supernatant is transferred to a new vial for LC-MS analysis. Alternatively the supernatant can be concentrated by evaporation or lyophilization and reconstituted a small volume of 50% methanol prior to analysis [165].

Liquid chromatography coupled mass spectrometry

The sample is injected onto a high pressure liquid chromatography (HPLC) system for the separation of analytes using a mixed mode HPLC column (2.1mm x 150mm Dionex Trinity P1 featuring reverse phase, and cation/anion exchange properties). A three solvent gradient is composed of 20mM formic acid in water (Solvent A), 20mM formic acid in acetonitrile (Solvent B), and 200mM ammonium acetate and formic acid (Solvent C) which was programmed using a Thermo Surveyor MS Pump Plus. Combinations of organic and salt concentration gradients are used for elution at a flow rate of 250 µL per min at 30°C (**Figure 4-1**).

Figure 4-1. Three-solvent high pressure liquid chromatography solvent gradient programming. The solvent composition is plotted in terms of percent solvents B and C with the remaining percentage being allocated to solvent A and the system flowing at 250 uL/min throughout the 90-minute run. Solvent A is composed of water with 0.1% formic acid, solvent B is composed of acetonitrile with 0.1% formic acid, and solvent C is composed of 250mM ammonium formate and acetic acid in water. The exact gradient percentages are presented in the right sub-panel. The column is washed and equilibrated after every sample injection.



A

Time	% B	% C
0	1	4
5	1	4
10	20	4
30	25	10
45	30	30
55	30	70
60	60	30
75	70	4
80	96	4
85	1	4
90	1	4

B

The HPLC is coupled to a Thermo LTQ Orbitrap XL through an electrospray ionization source (ESI). The mass spectrometer is set to do high resolution full scans from 110 to 2000 units mass-to-charge (m/z) with data-dependent selection of precursor ions for fragmentation scans. The data is collected in both positive and negative ion modes.

Data analysis

Manual analysis using Thermo Qual Browser software was used to quantify representative compounds which are selected based on their chemical diversity and perceived biomedical importance. Thermo Proteome Discoverer software was used to match fragmentation spectra to proteins based on theoretical tryptic peptide sequences. Metabolites are identified

based on matching exact monoisotopic mass, relative isotope abundances, and major fragmentation ions reported in the Human Metabolome Database [166].

RESULTS AND DISCUSSION

The methods for global metabolomic, lipidomic, and proteomic analysis are well-established. These procedures consist of four main steps: sample preparation, chromatography, mass spectrometry, and data analysis. In metabolomics and lipidomics, the sample preparation typically consists of deproteinization, extraction, and concentration steps, e.g. methanolic protein precipitation and lyophilization, or solid phase extraction. In bottom-up proteomics, proteins are isolated and cleaved with disulfide reducing agents and proteases in order to generate smaller polypeptides which are more amenable to detection and analysis compared to intact proteins. These peptides are then fragmented in the mass spectrometer and their fragmentation signatures are mapped to reference protein sequences. Reverse phase liquid chromatography and electrospray ionization mass spectrometry are most commonly used across these fields, though a range of other separation and ionization modalities are also employed. The design of the mass spectrometry and data analysis is generally finely tailored to the objectives of the work and types of molecules being analyzed.

Most alterations to these standard procedures are motivated by increased sensitivity towards a subset of analytes. These protocols typically modify the steps prior to introduction into the mass spectrometer, tuning the chromatographic and ionization chemistry to suit the target compounds. For example, changes to the extraction solvents, the chromatographic media, mobile phase additives, and the ion source parameters can be implemented depending on the analytes of interest. To my knowledge there have been no reported attempts at combining proteomics,

lipidomics, and polar metabolomics analyses into a single global assay despite sustained interest in their integration at the data analysis level [167, 168].

Conventional wisdom in analytical chemistry states that no single method can be used to analyze all chemical diversity [165]; consequently many perceive the existence of a universal analytical method to be infeasible. While this principle is technically true, it is possible that compromises are not as severe and unavoidable as previously thought. Given the potential of a universal assay method to collect large amounts of biochemical information, I sought to address this missing capability by developing a general procedure to analyze metabolites, lipids, and proteins. A significant amount of time was dedicated to exploring different approaches to address the obstacles of such a procedure and a working prototype method is described in the Methods section above.

The sample preparation process has three main objectives: generating the analytes, extracting LC-MS compatible compounds from the sample, and removing LC-MS incompatible components from the extract. As previously stated, disulfide and tryptic cleavage are standard for protein analysis, however, considering the technical complications associated with disulfide reduction and alkylation, and with using detergent-aided denaturation of proteins, these steps were omitted in favor of more robust and reproducible the method. Computational analysis of the human proteome reveals that 95% of proteins are made up of more than 50% non-cysteine containing tryptic peptides. Thus in theory the majority of proteins should produce tryptic fragments without disulfide bridge cleavage, given that steric hindrance due to protein structure is ignored. In order to avoid detergent-based facilitation of trypsin digestion, I considered using organic solvents to serve as denaturants. Previous work has shown that trypsin enzymatic activity is maintained or improved in the presence methanol and other organic solvents [169, 170]. I

found that methanol was suitable for this role and was also compatible with previously identified methods for extracting both polar metabolites and nonpolar lipids [171] and for deproteinizing these extracts [164] using 1:1:1 methanol:acetonitrile:acetone. Taken together, omitting the reduction and alkylation of disulfides, denaturing proteins using methanol during digestion, and undigested protein with organic solvents provides as simple procedure for producing LC-MS compatible sample preparations.

Separation of a diverse range of compounds requires multi-modal chromatographic retention. Two commercially available mixed mode columns were considered: the Scherzo SS C-18 and the Acclaim Trinity P1. Both of these columns have reverse phase and cation/anion exchange properties, however the strong reverse phase character of the Scherzo column required much longer elution times and stronger organic solvents in order to the elute lipids. Thus for the purposes of this work we focused on the Acclaim column. Ammonium formate and acetic acid solution was used as the ion exchange eluent as these additives have been found to yield moderate ESI-MS sensitivity across most classes of lipids [172]. Both organic-to-aqueous and aqueous-to-organic solvent gradient programming were found to retain the majority of compounds as long as the column was properly equilibrated with acidified water prior to sample injection. An aqueous-to-organic solvent was chosen in order to maintain a conventional retention ordering compared to conventional reverse-phase gradients.

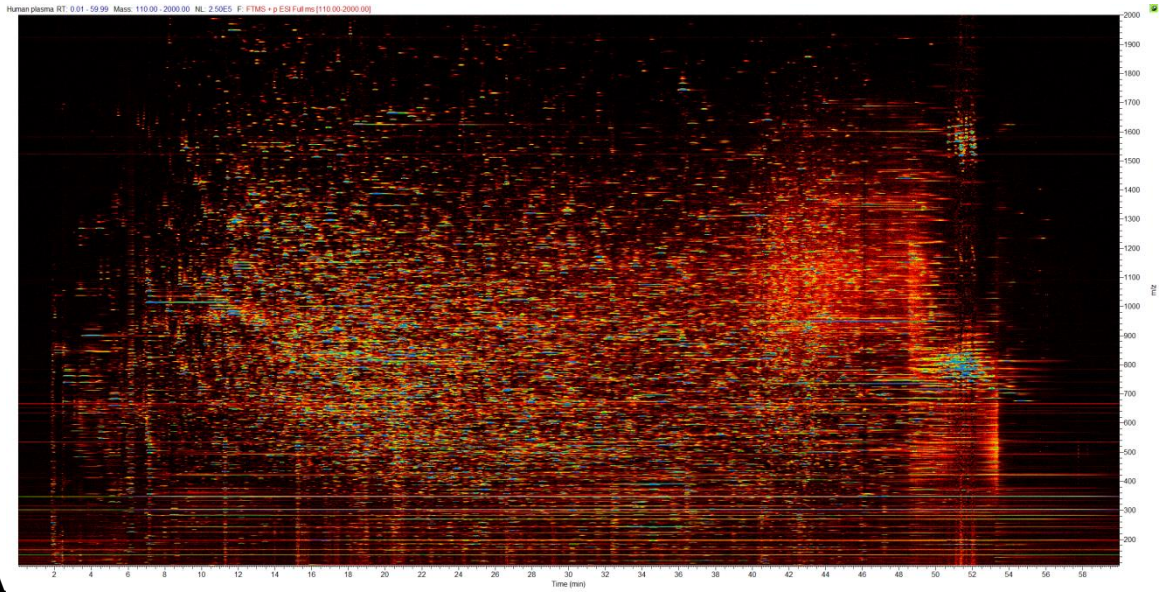
Complexity of the plasma metabolome and proteome

Blood is among the most accessible and informative tissues, containing a wide range of proteins, lipids, and metabolites. As a proof-of-concept, blood plasma was analyzed using the developed method. The molecular composition of plasma using this method is quite complex, including signals from sugars, amino acids, peptides, and polar and nonpolar lipids (**Figure 4-2**).

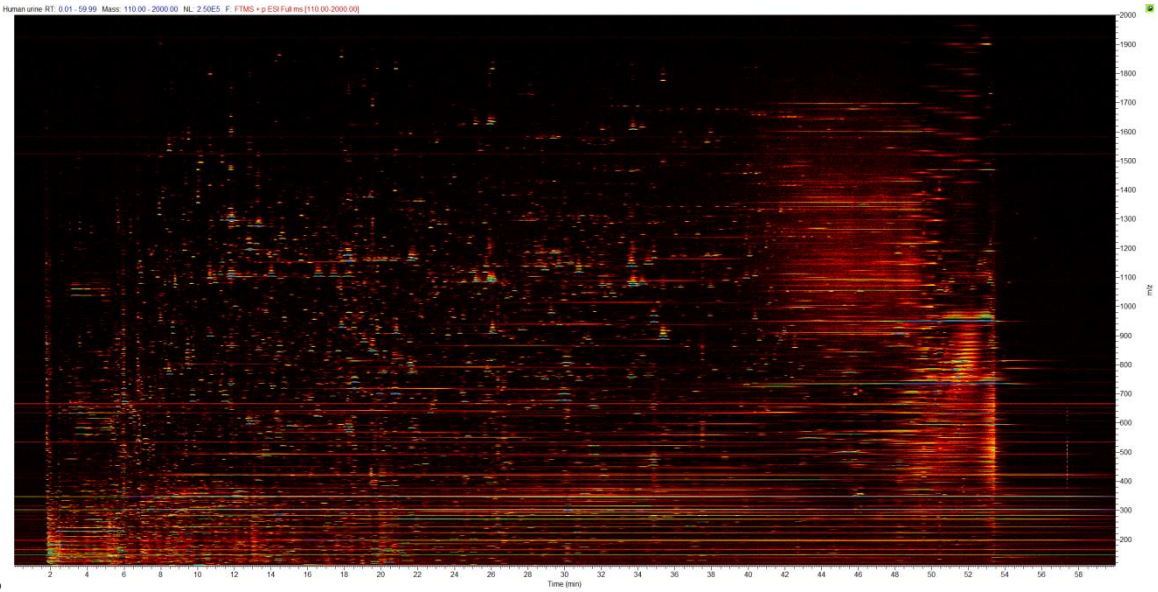
Unexpectedly electrolytes such as sodium, potassium, chloride, phosphate and also metals such as iron and copper (as EDTA chelates) were also measurable by this method. Profiling of samples not treated with trypsin revealed showed a much sparser heatmap, indicating that the majority of observed signals are derived from tryptic protein fragments.

Figure 4-2. Ion heatmap of various specimens. Time and m/z are represented along the x and y axes respectively, where ion log-intensity is represented by a color scale ranging from red to blue. Each of the specks on in the image represents a unique ionized molecule from the processed plasma sample. The ion heatmaps for human plasma (**A**), urine (**B**), and bovine adipose tissue (**C**) are presented.

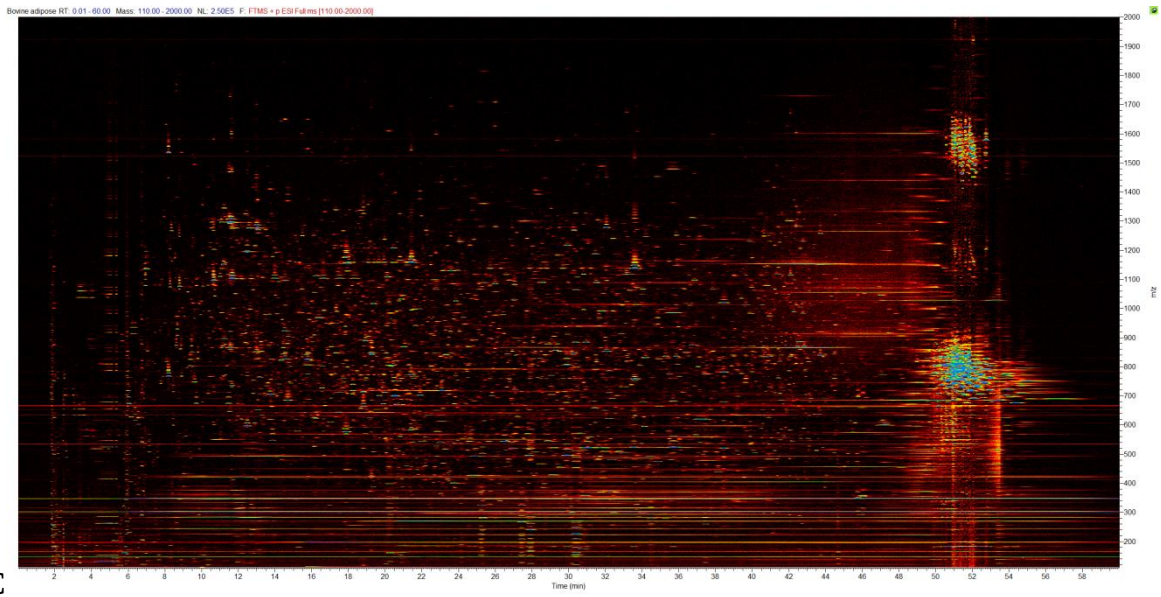
A



B



C



To assess the proteomic coverage of the method, we matched ion fragmentation spectra to all known human protein sequences using Thermo ProteomeDiscoverer software. This analysis reported the identification peptide fragments from 1975 protein families including albumin, immunoglobulins, apolipoproteins, fibrinogens, and C-reactive protein. To assess the metabolomic coverage, we used Thermo CompoundDiscoverer which was able to tentatively identify over 273 of metabolites. We suspect the true number of metabolites to be much greater

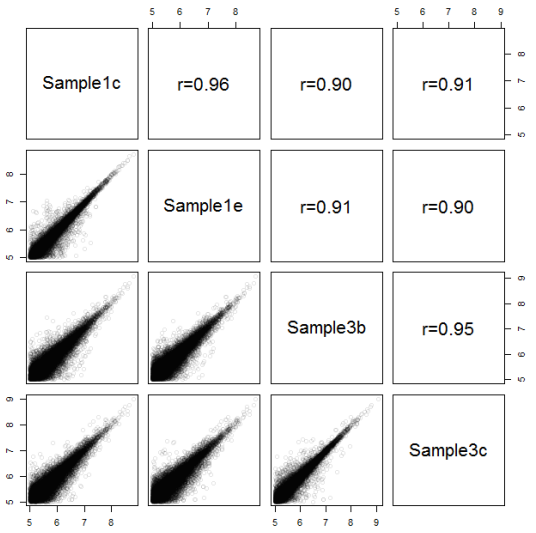
because CompoundDiscoverer is designed for newer instruments and the report only included one metabolite with a molecular weight under 400 Daltons which is conflicting with our manual identification of multiple metabolites within that range (e.g. sugars and amino acids).

The method was also applied to various specimens including urine, cell lines, muscle, bone marrow, adipose tissue, tendon, and blood vessel. Qualitative examination of these data indicates that they have vastly different biochemical profiles. Urine has a high relative abundance of compounds below 500 Daltons whereas adipose and bone marrow tissues have relatively high lipid content; cells, muscle, tendon, and artery have relatively high protein content as expected (**Figure 4-2**). Overall, the method appears to be generally applicable to wet biological specimens.

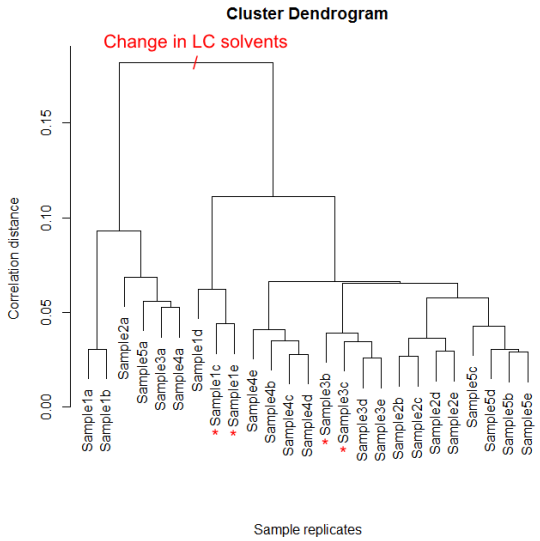
Reproducibility of quantitation

To assess the reproducibility of the method, the pooled plasma was processed and analyzed in replicate (5 sample processing replicates of 5 different samples = 25 total samples). Global agreement between was stronger between technical replicates than between samples (**Figure 4-3**). Quantitative reproducibility was also assessed manually using a set of abundant metabolites and peptides selected based on their chemical diversity and their perceived clinical importance. The chromatographic retention of these compounds appears to be highly reproducible (**Figure 4-4**). The assay was also able to distinguish between the different levels of most analytes between samples ($0.0005 < \text{Kruskal-Wallace } p < 0.074$), though some of these measurements did not appear to be highly reproducible (**Figure 4-5**).

Figure 4-3. Reproducibility of measurements. Representative inter-replicate and inter-sample reproducibility are presented as pairwise scatterplots and correlations (**A**). Ion intensity counts are summed into 0.01 m/z and 2 minute bins generating approximately 20,000 raw features which are then log10 transformed. Correlation-based hierarchical clustering of samples is also presented to demonstrate the agreement across replicate preparation of all five samples (**B**). Representative profiles found in the scatterplots are denoted with asterisks. Notably refilling solvents had a dramatic effect on the similarity between intensity profiles resulting in a separate sample clusters pre and post-refill.



A



B

Figure 4-4. Extracted ion chromatograms of selected molecules from plasma sample analyses. Chromatograms are generated by extracting the exact monoisotopic masses of the following analytes (from top to bottom): creatine, taurine, uric acid, hemoglobin (tryptic peptide), serum albumin (tryptic peptide), cholesterol, phosphatidylcholines 36:2, and triglycerides 48:0. The five sample processing replicates of the five samples are overlaid with partial transparency.

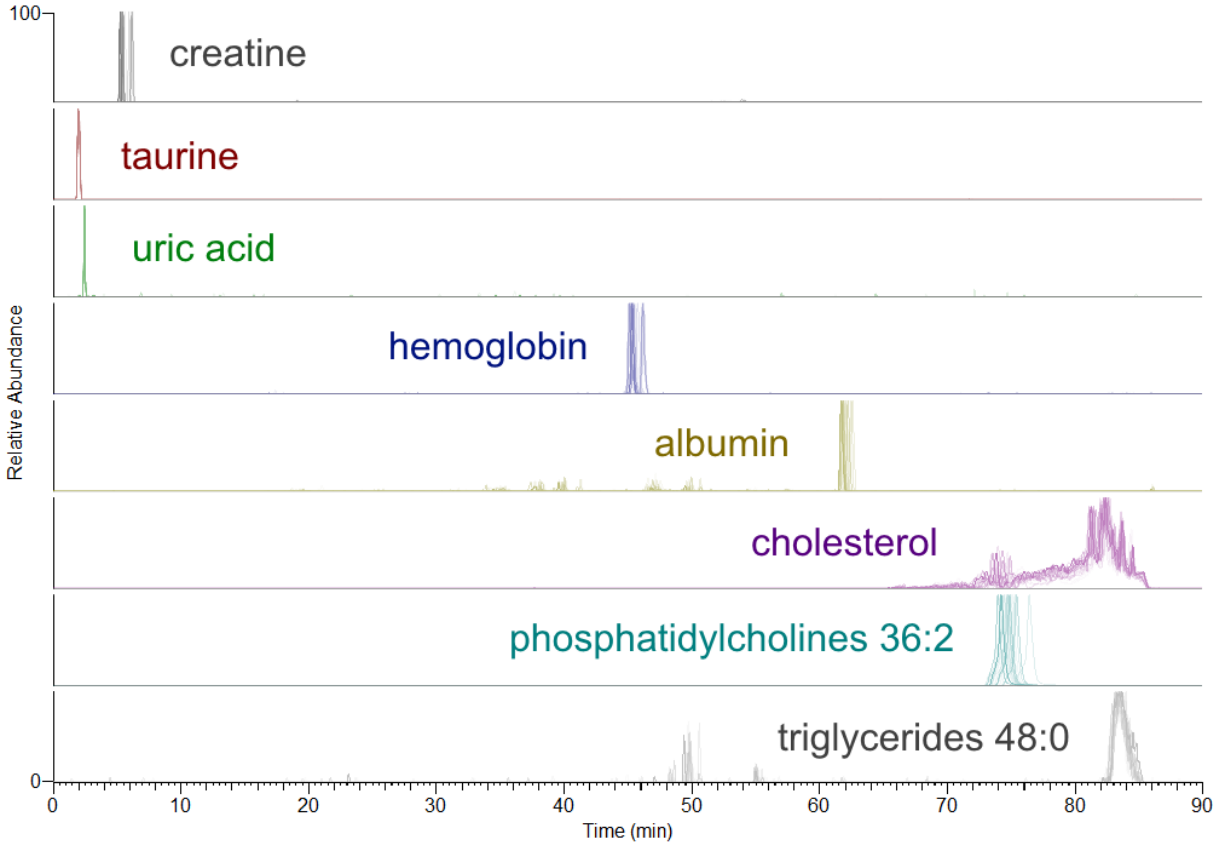
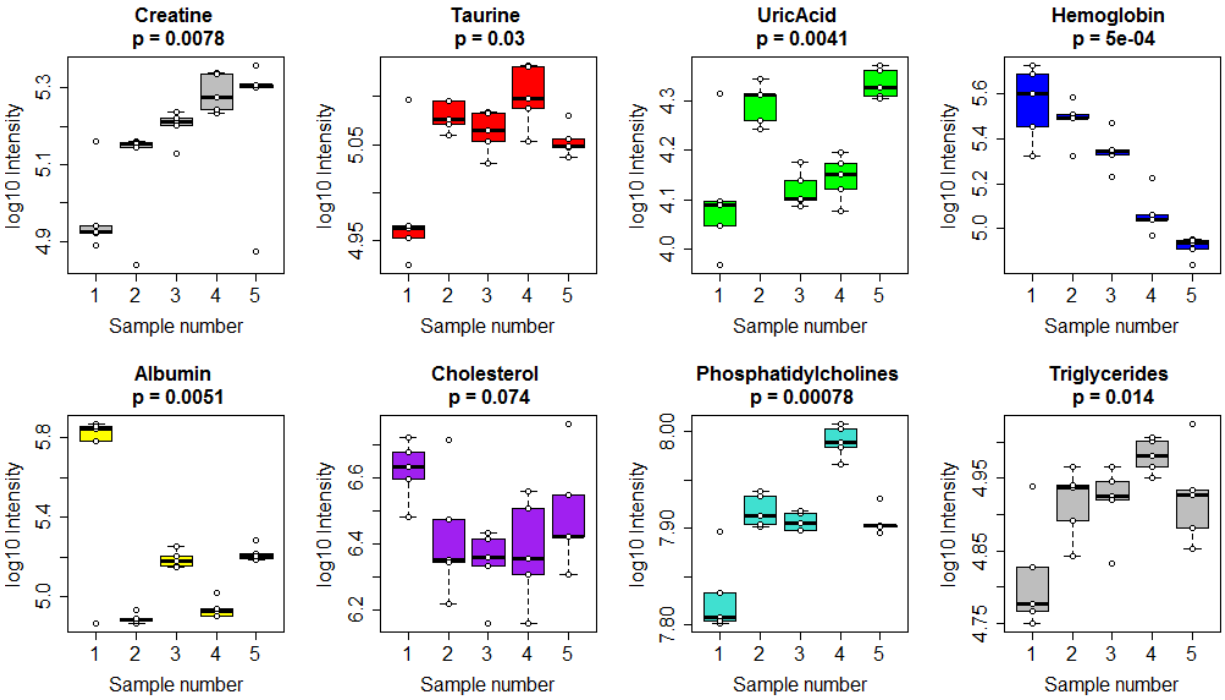


Figure 4-5. Quantitation of eight selected analytes from five different plasma samples. Five sample processing replicates were measured for 5 different plasma samples. The peak areas for creatine, taurine, uric acid, hemoglobin (tryptic peptide), serum albumin (tryptic peptide), cholesterol, phosphatidylcholines 36:2, triglycerides 48:0 were manually collected using Thermo Qual Browser software. Peak areas were log10 transformed and each sample was normalized by the average peak areas in order to adjust for variation in injection volume and mass spectrometer response (which was found to decrease across LC-MS runs). Boxplots presenting quartiles for each analyte are displayed with log10 intensity and sample number represented on the y and x axes respectively. Individual replicate measurements are shown as small white circles outlined in black.



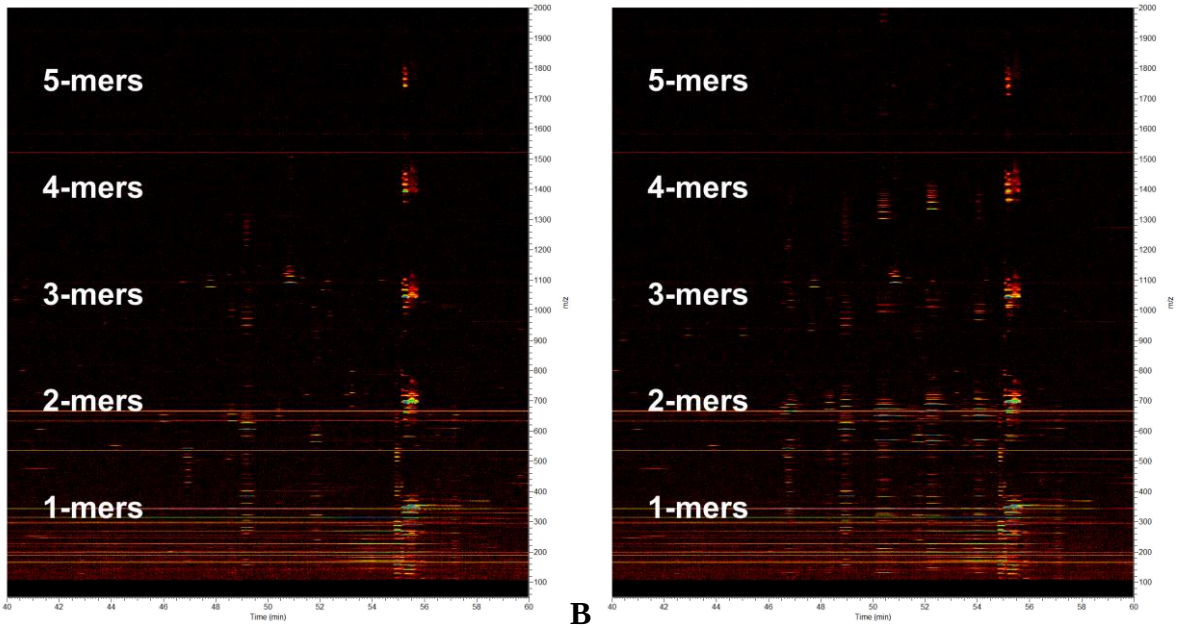
Importantly, in attempting to verify notable analytes it appears that well-studied hormones are below the limit of detection using this method. This is expected as these molecules are present at low concentrations in plasma and usually require specialized highly sensitive analytical methods for quantitation. These results demonstrate that although the method is indeed capable of globally capturing a wide range of molecules, the sensitivity levels will preclude

compounds which have low abundance and weak ESI-MS responses. Future improvements on the assay and/or in instrumentation may increase analytical sensitivity to levels which are capable of detecting low concentration analytes such as hormones or tissue leakage molecules.

Bottom-up transcriptomics using RNA endonucleases

Since RNA and DNA capture different types of bioinformation which may be complementary, the analysis of these types of macromolecules was also explored. Huh7 human hepatocellular carcinoma cell line samples consisting of approximately 50 million cells were processed using the method developed here except with and without RNase A in the digestion step. RNA fragments were found to be separated by chromatographic retention and detectable by negative ion mode MS as singly charged ions less than 5 nucleotides long (**Figure 4-6**). The low specificity of RNase A, cleaving after cytidine and uridine residues, produces mostly monomers with a decreasing proportion of 2- to 4-mers. Though fragmentation of these oligonucleotides yields characteristic ions which allow for the oligonucleotide sequencing, their short length sequences do not specifically map to the transcriptome. Usage of more specific endonucleases, such as RNase T1 with guanosine specificity, will produce longer oligonucleotides which may be able to capture some mapping specificity. In a computational analysis of all of the theoretical fragments (<20 bases long) produced by RNase T1 from the human transcriptome, 95% of these oligonucleotides map to the transcriptome less than 22 times and 50% of them map to the transcriptome less than 2 times. These results suggest that a substantial proportion of RNase T1 products would have some degree of transcriptomic mapping specificity.

Figure 4-6. Detection of small RNA oligonucleotides in Huh7 cells. Negative ion heatmaps are presented of cells untreated (**A**) and treated with RNase A (**B**). The average molecular weight of nucleotide residues is approximately 350 Daltons; k-mers are labeled according to length in white toward the left of each heatmap. Samples appear to have endogenous oligomers as indicated by the ladder present in the RNase-untreated sample. The addition of RNase A results in the appearance of new oligonucleotide species which separate in along the x-axis (retention time).



Whether or not such oligonucleotides will be present at high enough concentrations to detect remains unconfirmed. It may be advantageous to produce fragments with degenerate sequences in order to bring molar concentrations above instrumental limits of detection. Though this would necessitate the development of deconvolutional algorithms, having a method to capture global RNA sequence information using LC-MS would be of scientific interest. Overall, these results suggest that a bottom-up strategy using site-specific endoribonucleases has some potential to enable the measurement of specific RNA sequences and suggests that applying similar bottom-up strategies to other biological polymers such as the polysaccharides,

proteoglycans, and lipopolysaccharides would bring the concept of a universal molecular assay into the realm of possibility.

Limitations

Compared to specialized methods there is reduced coverage of low abundance species due to suboptimal extraction and ionization conditions. For example fractionated bottom-up proteomics experiments regularly show detection of >4,000 protein families in plasma [173] where we only detect about 2,000 protein families. Likewise, low abundance metabolites are better approached with specialized methods which can enrich for specific molecules such as solid phase or liquid-liquid extraction.

The vast majority of ions detected using this method remains unidentified. Though software exists to analysis specific classes of molecules such as peptides, lipids, or metabolites, none of these existing solutions are well-suited for the analyses of all of these types of compounds. In order to realize any potential that this method has, biochemical-agnostic computational algorithms must be developed which can quantify and identify features in this complex data.

Future directions

There are still a number of steps before the technology is ready for general adoption. The establishment of software for data normalization, quantification, and identification of ions is a high priority. In the short-term, substantial improvements can be made by using state-of-the-art technologies such as nanospray ionization and modern mass spectrometers. Over the long term, technological advancements in LC-MS sensitivity are expected to continue (about 50-fold every decade). If so, measurement of trace sample components will become increasingly probably even in the face of poor extraction and ionization yields, allowing for the detection of low abundance

protein modifications, hormones, toxic exposures, heavy metals [174], infectious agents [175], and oligonucleotides [176, 177].

This work has demonstrated that it is possible to measure a wide range of clinical biomarkers including proteins, lipids, metabolites, and electrolytes simultaneously, establishing potential value in biomedical research and practice. Being able to quantify such molecules could be invaluable to large-scale studies as it would provide surrogate data on an array of biochemical traits, clinical phenotypes, and confounding factors (e.g. tobacco usage) at dramatically reduced cost. These types of data are typically collected one at a time, incurring additional cost with each extra data point; this method could be used to multiplex the measurement of many of these variables at no additional cost (e.g. glucose, cholesterol, and drug/medication metabolites). This concept is not a new as similar strategies have been proposed for genomics, metabolomics, and proteomics [158], however this assay is uniquely suited for this role as it covers a relatively large proportion of clinical laboratory tests and is relatively cost-efficient (<$50 per assay instrument costs included).

Conclusions

In this work I demonstrate that contrary to prevailing assumptions, it is possible to combine the analysis of proteins, lipids, polar metabolites and electrolytes in a single universal untargeted assay. The procedure integrates several prerequisite technologies including mixed mode chromatography and high resolution mass spectrometry to provide a method that is sufficiently sensitive to measure a number of clinically relevant of molecules in plasma. This work also provides a starting point for the development of a more universal molecular assay which may come to integrate polynucleotide and polysaccharide bioinformation.

Chapter 5: Overarching conclusions

Over the course of my doctoral studies, I encountered the challenges of analyzing transcriptomic and epigenomic data. In this dissertation, I have described the relationships between epigenetic aging and lifestyle factors, the relationships between HIV-associated neurocognitive disorders and gene expression in peripheral monocytes, and the relationships between epigenetic aging and gene expression in PBMCs and monocytes. These studies have elucidated that the epigenetic clock phenomenon is multi-factorial, being related to healthy diet, alcohol, BMI, education, cardiometabolic health, plasma biomarkers, sex, cell composition, inflammation, and interferon signaling. These studies have led me to propose interferon-mediated induction of senescence as a major determinant of the epigenetic clock phenomenon. Overall, this theory ties together the cell-intrinsic, non-replicative process, and pro-inflammatory aspects of the epigenetic clock.

Limitations

All of these analyses are based on microarray profiles which entail some limitations. For example, the Illumina methylation arrays only measure methylation at a subset of all CpG sites, and measurements on this platform can be affected by genetic polymorphisms which affect probe function. This technology also makes key assumptions during the estimation of methylation levels. For example, these arrays are unable to distinguish between unmethylated cytosines and spontaneous 5-methylcytosine deaminations because the assay converts unmethylated cytidines to thymidines during the bisulfite conversion and amplification process; it is possible that apparent hypomethylation that occurs with age may be a partial reflection of the accumulation of 5-methylcytosine to thymidine mutations over time. In contrast, deep sequencing of bisulfite

treated and untreated samples should allow for the direct measurement of C to T mutations in the untreated samples.

Though the conclusions of these studies on epigenetic aging were intuitive and consistent with prior knowledge, they were largely unsurprising. Upon reflection, I gathered that insightful findings tend to rely on new types of rich sample annotation data. For example, the study of plasma biomarkers allowed us to establish the connection between epigenetic aging and fruit and vegetable intake—these results would have been insignificant had we relied on self-reported intake levels. If a larger number of quantitative surrogate biomarkers had been available, it is likely we would have found many other associations. In order to address this common deficiency across genomics and other studies, I developed an assay to integrate the measurement of diverse biochemicals in hopes of simplifying the process of acquiring informative sample annotations.

Future directions

There are many factors that have been found to be associated with epigenetic aging yet a coherent theory has yet to be confirmed that unifies these observations. Future work may be directed at experimental validation of leading theories on the causes of epigenetic aging and testing the therapeutic potential of intervening on this surrogate measure of aging.

BIBLIOGRAPHY

1. Boehm, A.-M., et al., *FoxO is a critical regulator of stem cell maintenance in immortal Hydra*. Proceedings of the National Academy of Sciences, 2012. **109**(48): p. 19697-19702.

2. Devarapalli, P., et al., *The conserved mitochondrial gene distribution in relatives of Turritopsis nutricula, an immortal jellyfish.* Bioinformation, 2014. **10**(9): p. 586-591.

3. López-Otín, C., et al., *The Hallmarks of Aging.* Cell, 2013. **153**(6): p. 1194-1217.

4. Longo, V.D. and L. Fontana, *Calorie restriction and cancer prevention: metabolic and molecular mechanisms.* Trends in pharmacological sciences, 2010. **31**(2): p. 89-98.

5. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome biology, 2013. **14**(10): p. R115.

6. Saxonov, S., P. Berg, and D.L. Brutlag, *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.* Proceedings of the National Academy of Sciences, 2006. **103**(5): p. 1412-1417.

7. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nature Reviews Genetics, 2012. **13**: p. 484.

8. Christensen, B., et al., *Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context.* PLoS Genet, 2009. **5**(8): p. e1000602.

9. Bollati, V., et al., *Decline in genomic DNA methylation through aging in a cohort of elderly subjects.* Mech Ageing Dev, 2009. **130**(4): p. 234-239.

10. Rakyan, V.K., et al., *Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains.* Genome Res, 2010. **20**(4): p. 434-9.

11. Teschendorff, A.E., et al., *Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer.* Genome Res, 2010. **20**(4): p. 440-6.

12. Horvath, S., et al., *Aging effects on DNA methylation modules in human brain and blood tissue.* Genome Biol., 2012. **13**: p. R97.

13. Numata, S., et al., *DNA Methylation Signatures in Development and Aging of the Human Prefrontal Cortex.* Am J Hum Genet., 2012. **90**(2): p. 260-272.

14. Alisch, R.S., et al., *Age-associated DNA methylation in pediatric populations.* Genome Res., 2012. **22**(4): p. 623-632.

15. Johansson, A., S. Enroth, and U. Gyllensten, *Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan.* PLoS One, 2013. **8**(6): p. e67378.

16. Day, K., et al., *Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape.* Genome Biol., 2013. **14**(9): p. R102.

17. Bocklandt, S., et al., *Epigenetic predictor of age.* PLoS ONE, 2011. **6**(6): p. e14821.

18. Garagnani, P., et al., *Methylation of ELOVL2 gene as a new epigenetic marker of age.* Aging Cell, 2012. **11**(6): p. 1132-1134.

19. Hannum, G., et al., *Genome-wide methylation profiles reveal quantitative views of human aging rates.* Mol Cell, 2013. **49**(2): p. 359-67.

20. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biol, 2013. **14**(R115).

21. Lin, Q., et al., *DNA methylation levels at individual age-associated CpG sites can be indicative for life expectancy.* Aging (Albany NY), 2016. **8**(2): p. 394-401.

22. Marioni, R., et al., *DNA methylation age of blood predicts all-cause mortality in later life.* Genome Biol., 2015. **16**(1): p. 25.

23. Christiansen, L., et al., *DNA methylation age is associated with mortality in a longitudinal Danish twin study.* Aging Cell, 2015.

24. Perna, L., et al., *Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort.* Clinical Epigenetics, 2016. **8**(1): p. 1-7.

25. Chen, B.H., et al., *DNA methylation-based measures of biological age: meta-analysis predicting time to death.* Aging (Albany NY), 2016. **8**(9): p. 1844-1865.

26. Breitling, L.P., et al., *Frailty is associated with the epigenetic clock but not with telomere length in a German cohort.* Clin Epigenetics, 2016. **8**.

27. Levine, M.E., et al., *DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative.* Aging (Albany NY), 2015. **7**(9): p. 690-700.

28. Marioni, R.E., et al., *The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936.* Int J Epidemiol, 2015. **44**.

29. Horvath, S., et al., *Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring.* Aging (Albany NY), 2015. **7**(12): p. 1159-70.

30. Levine, M.E., et al., *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): p. 1198-211.

31. Horvath, S., et al., *The cerebellum ages slowly according to the epigenetic clock.* Aging (Albany NY), 2015. **7**(5): p. 294-306.

32. Walker, R.F., et al., *Epigenetic age analysis of children who seem to evade aging.* Aging (Albany NY), 2015. **7**(5): p. 334-9.

33. Horvath, S., et al., *Accelerated Epigenetic Aging in Down Syndrome.* Aging Cell, 2015. **14**(1).

34. Horvath, S. and A.J. Levine, *HIV-1 infection accelerates age according to the epigenetic clock.* J Infect Dis, 2015.

35. Horvath, S., et al., *Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels.* Aging (Albany NY), 2016. **8**(7): p. 1485-512.

36. Horvath, S., et al., *Obesity accelerates epigenetic aging of human liver.* Proc Natl Acad Sci U S A, 2014. **111**(43): p. 15538-43.

37. Zannas, A., et al., *Lifetime stress accelerates epigenetic aging in an urban, African American cohort: relevance of glucocorticoid signaling.* Genome Biology, 2015. **16**(1): p. 266.

38. Levine, M.E., et al., *Menopause accelerates biological aging.* Proc Natl Acad Sci U S A, 2016. **113**(33): p. 9327-32.

39. Vidal, L., et al., *Specific increase of methylation age in osteoarthritis cartilage.* Osteoarthritis and Cartilage, 2016. **24**: p. S63.

40. Horvath, S. and B.R. Ritz, *Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients.* Aging (Albany NY), 2015. **7**(12): p. 1130-42.

41. Kaczmarczyk, M.M., M.J. Miller, and G.G. Freund, *The health benefits of dietary fiber: beyond the usual suspects of type 2 diabetes mellitus, cardiovascular disease and colon cancer.* Metabolism, 2012. **61**(8): p. 1058-66.

42. Kris-Etherton, P.M., W.S. Harris, and L.J. Appel, *Fish consumption, fish oil, omega-3 fatty acids, and cardiovascular disease.* Circulation, 2002. **106**(21): p. 2747-57.

43. van't Veer, P., et al., *Fruits and vegetables in the prevention of cancer and cardiovascular disease.* Public health nutrition, 2000. **3**(01): p. 103-107.

44. Giugliano, D., A. Ceriello, and K. Esposito, *The effects of diet on inflammation: emphasis on the metabolic syndrome.* Journal of the American College of Cardiology, 2006. **48**(4): p. 677-685.

45. Mozaffarian, D., *Dietary and Policy Priorities for Cardiovascular Disease, Diabetes, and Obesity.* A Comprehensive Review, 2016. **133**(2): p. 187-225.

46. Alberti, K.G., et al., *Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity.* Circulation, 2009. **120**(16): p. 1640-5.

47. Grundy, S.M., et al., *Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement.* Circulation, 2005. **112**(17): p. 2735-52.

48. Wilson, P.W., et al., *Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus.* Circulation, 2005. **112**(20): p. 3066-72.

49. Thun, M.J., et al., *Alcohol Consumption and Mortality among Middle-Aged and Elderly U.S. Adults.* New England Journal of Medicine, 1997. **337**(24): p. 1705-1714.

50. Rimm, E.B., et al., *Moderate alcohol intake and lower risk of coronary heart disease: meta-analysis of effects on lipids and haemostatic factors.* BMJ, 1999. **319**(7224): p. 1523.

51. Warburton, D.E., C.W. Nicol, and S.S. Bredin, *Health benefits of physical activity: the evidence.* Cmaj, 2006. **174**(6): p. 801-9.

52. Gonzalez, M.A., F. Rodriguez Artalejo, and J.R. Calero, *Relationship between socioeconomic status and ischaemic heart disease in cohort and case-control studies: 1960-1993.* Int J Epidemiol, 1998. **27**(3): p. 350-8.

53. Madsen, M., et al., *Does educational status impact adult mortality in Denmark? A twin approach.* Am J Epidemiol, 2010. **172**(2): p. 225-34.

54. Conti, G., J. Heckman, and S. Urzua, *THE EDUCATION-HEALTH GRADIENT.* The American economic review, 2010. **100**(2): p. 234-238.

55. Currie, J., *Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development.* National Bureau of Economic Research Working Paper Series, 2008. **No. 13987**.

56. Cutler, D.M. and A. Lleras-Muney, *Understanding Differences in Health Behaviors by Education.* Journal of health economics, 2010. **29**(1): p. 1-28.

57. Mirabello, L., et al., *The association between leukocyte telomere length and cigarette smoking, dietary and physical variables, and risk of prostate cancer.* Aging cell, 2009. **8**(4): p. 405-413.

58. Boccardi, V., et al., *Mediterranean diet, telomere maintenance and health status among elderly.* PLoS One, 2013. **8**(4): p. e62781.

59. García-Calzón, S., et al., *Longitudinal association of telomere length and obesity indices in an intervention study with a Mediterranean diet: the PREDIMED-NAVARRA trial.* International journal of obesity, 2014. **38**(2): p. 177-182.

60. Brouilette, S.W., et al., *Telomere length, risk of coronary heart disease, and statin treatment in the West of Scotland Primary Prevention Study: a nested case-control study.* The Lancet, 2007. **369**(9556): p. 107-114.

61. Kiecolt-Glaser, J.K., et al., *Omega-3 fatty acids, oxidative stress, and leukocyte telomere length: a randomized controlled trial.* Brain, behavior, and immunity, 2013. **28**: p. 16-24.

62. Cassidy, A., et al., *Associations between diet, lifestyle factors, and telomere length in women.* Am J Clin Nutr, 2010. **91**(5): p. 1273-80.

63. Latifovic, L., et al., *The Influence of Alcohol Consumption, Cigarette Smoking, and Physical Activity on Leukocyte Telomere Length.* Cancer Epidemiol Biomarkers Prev, 2016. **25**(2): p. 374-80.

64. Cherkas, L.F., et al., *The association between physical activity in leisure time and leukocyte telomere length.* Archives of Internal Medicine, 2008. **168**(2): p. 154-158.

65. Adler, N., et al., *Educational attainment and late life telomere length in the Health, Aging and Body Composition Study.* Brain Behav Immun, 2013. **27**(1): p. 15-21.

66. Horvath, S., et al., *An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease.* Genome Biol, 2016. **17**(1): p. 171.

67. McGuire, S., *US Department of Agriculture and US Department of Health and Human Services, Dietary Guidelines for Americans, 2010. Washington, DC: US Government Printing Office, January 2011.* Advances in Nutrition: An International Review Journal, 2011. **2**(3): p. 293-294.

68. Committee, D.G.A., *Scientific Report of the 2015 Dietary Guidelines Advisory Committee.* Washington (DC): USDA and US Department of Health and Human Services, 2015.

69. Lu, A.T., et al., *Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum.* Nat Commun, 2016. **7**: p. 10561.

70. Castelo-Branco, C. and I. Soveral, *The immune system and aging: a review.* Gynecological Endocrinology, 2014. **30**(1): p. 16-22.

71. Finch, C.E., *Evolution of the human lifespan and diseases of aging: Roles of infection, inflammation, and nutrition.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**: p. 1718-1724.

72. Chowdhury, R., et al., *Association between fish consumption, long chain omega 3 fatty acids, and risk of cerebrovascular disease: systematic review and meta-analysis.* Bmj, 2012. **345**: p. e6698.

73. Zheng, J.-S., et al., *Intake of fish and marine n-3 polyunsaturated fatty acids and risk of breast cancer: meta-analysis of data from 21 independent prospective cohort studies.* Bmj, 2013. **346**.

74. Farina, E.K., et al., *Protective effects of fish intake and interactive effects of long-chain polyunsaturated fatty acid intakes on hip bone mineral density in older adults: the Framingham Osteoporosis Study.* The American journal of clinical nutrition, 2011. **93**(5): p. 1142-1151.

75. Bannenberg, G.L., et al., *Molecular circuits of resolution: formation and actions of resolvins and protectins.* The Journal of Immunology, 2005. **174**(7): p. 4345-4355.

76. Ridker, P.M., et al., *C-reactive protein, the metabolic syndrome, and risk of incident cardiovascular events an 8-year follow-up of 14 719 initially healthy American women.* Circulation, 2003. **107**(3): p. 391-397.

77. De Labry, L.O., et al., *Alcohol consumption and mortality in an American male population: recovering the U-shaped curve--findings from the normative Aging Study.* Journal of studies on alcohol, 1992. **53**(1): p. 25-32.

78. Knott, C.S., et al., *All cause mortality and the case for age specific alcohol consumption guidelines: pooled analyses of up to 10 population based cohorts.* bmj, 2015. **350**: p. h384.

79. Beach, S.R., et al., *Methylomic Aging as a Window onto the Influence of Lifestyle: Tobacco and Alcohol Use Alter the Rate of Biological Aging.* Journal of the American Geriatrics Society, 2015. **63**(12): p. 2519-2525.

80. Grønbæk, M., et al., *Type of alcohol consumed and mortality from all causes, coronary heart disease, and cancer.* Annals of internal medicine, 2000. **133**(6): p. 411-419.

81. Volpato, S., et al., *Relationship of Alcohol Intake With Inflammatory Markers and Plasminogen Activator Inhibitior-1 in Well-Functioning Older Adults The Health, Aging, and Body Composition Study.* Circulation, 2004. **109**(5): p. 607-612.

82. Fillmore, K.M., et al., *Moderate alcohol use and reduced mortality risk: Systematic error in prospective studies.* Addiction Research & Theory, 2006. **14**(2): p. 101-132.

83. Brien, S.E., et al., *Effect of alcohol consumption on biological markers associated with risk of coronary heart disease: systematic review and meta-analysis of interventional studies.* Bmj, 2011. **342**: p. d636.

84. Dauchet, L., et al., *Fruit and vegetable consumption and risk of coronary heart disease: a meta-analysis of cohort studies.* The Journal of nutrition, 2006. **136**(10): p. 2588-2593.

85. He, F., et al., *Increased consumption of fruit and vegetables is related to a reduced risk of coronary heart disease: meta-analysis of cohort studies.* Journal of human hypertension, 2007. **21**(9): p. 717-728.

86. He, F.J., C.A. Nowson, and G.A. MacGregor, *Fruit and vegetable consumption and stroke: meta-analysis of cohort studies.* The Lancet, 2006. **367**(9507): p. 320-326.

87. Carter, P., et al., *Fruit and vegetable intake and incidence of type 2 diabetes mellitus: systematic review and meta-analysis.* Bmj, 2010. **341**: p. c4229.

88. Gandini, S., et al., *Meta-analysis of studies on breast cancer risk and diet: the role of fruit and vegetable consumption and the intake of associated micronutrients.* European journal of cancer, 2000. **36**(5): p. 636-646.

89. Wang, X., et al., *Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies.* 2014.

90. Holt, E.M., et al., *Fruit and vegetable consumption and its relation to markers of inflammation and oxidative stress in adolescents.* Journal of the American Dietetic Association, 2009. **109**(3): p. 414-421.

91. Esmaillzadeh, A., et al., *Fruit and vegetable intakes, C-reactive protein, and the metabolic syndrome.* The American journal of clinical nutrition, 2006. **84**(6): p. 1489-1497.

92. Ornish, D., et al., *Effect of comprehensive lifestyle changes on telomerase activity and telomere length in men with biopsy-proven low-risk prostate cancer: 5-year follow-up of a descriptive pilot study.* The Lancet Oncology, 2013. **14**(11): p. 1112-1120.

93. Farzaneh-Far, R., et al., *Association of marine omega-3 fatty acid levels with telomeric aging in patients with coronary heart disease.* Jama, 2010. **303**(3): p. 250-257.

94. Klemera, P. and S. Doubal, *A new approach to the concept and computation of biological age.* Mech Ageing Dev, 2006. **127**(3): p. 240-8.

95. WHI, *Design of the Women's Health Initiative clinical trial and observational study.* Controlled clinical trials, 1998. **19**(1): p. 61-109.

96. Patterson, R.E., et al., *Measurement characteristics of the Women's Health Initiative food frequency questionnaire.* Annals of epidemiology, 1999. **9**(3): p. 178-187.

97. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC bioinformatics, 2012. **13**(1): p. 86.

98. Horvath, S. and e. al, *An epigenetic age analysis of race/ethnicity, sex, and coronary heart disease.* Genome Biol, 2016.

99. Anonymous, A., *Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group.* Control Clin Trials, 1998. **19**(1): p. 61-109.

100. Johnson, W.E., A. Rabinovic, and C. Li, *Adjusting batch effects in microarray expression data using Empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-127.

101. Pisani, P., et al., *Relative validity and reproducibility of a food frequency dietary questionnaire for use in the Italian EPIC centres.* Int J Epidemiol, 1997. **26 Suppl 1**: p. S152-60.

102. Bartali, B., et al., *Low nutrient intake is an essential component of frailty in older persons.* The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 2006. **61**(6): p. 589-593.

103. Ferrucci, L., et al., *Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study.* J Am Geriatr Soc, 2000. **48**(12): p. 1618-25.

104. Heaton, R.K., et al., *HIV-associated neurocognitive disorders before and during the era of combination antiretroviral therapy: differences in rates, nature, and predictors.* J Neurovirol, 2011. **17**(1): p. 3-16.

105. Sacktor, N., et al., *Prevalence of HIV-associated neurocognitive disorders in the Multicenter AIDS Cohort Study.* Neurology, 2016. **86**(4): p. 334-40.

106. Pulliam, L., et al., *Unique monocyte subset in patients with AIDS dementia.* Lancet, 1997. **349**(9053): p. 692-5.

107. Ellery, P.J., et al., *The CD16+ monocyte subset is more permissive to infection and preferentially harbors HIV-1 in vivo.* J Immunol, 2007. **178**(10): p. 6581-9.

108. Peluso, R., et al., *A Trojan Horse mechanism for the spread of visna virus in monocytes.* Virology, 1985. **147**(1): p. 231-6.

109. Ancuta, P., A. Moses, and D. Gabuzda, *Transendothelial migration of CD16+ monocytes in response to fractalkine under constitutive and inflammatory conditions.* Immunobiology, 2004. **209**(1-2): p. 11-20.

110. Kraft-Terry, S.D., et al., *A coat of many colors: neuroimmune crosstalk in human immunodeficiency virus infection.* Neuron, 2009. **64**(1): p. 133-45.

111. Kedzierska, K. and S.M. Crowe, *The role of monocytes and macrophages in the pathogenesis of HIV-1 infection.* Curr Med Chem, 2002. **9**(21): p. 1893-903.

112. Glass, J.D., et al., *Immunocytochemical quantitation of human immunodeficiency virus in the brain: correlations with dementia.* Ann Neurol, 1995. **38**(5): p. 755-62.

113. Adle-Biassette, H., et al., *Neuronal apoptosis does not correlate with dementia in HIV infection but is related to microglial activation and axonal damage.* Neuropathol Appl Neurobiol, 1999. **25**(2): p. 123-33.

114. Lindl, K.A., et al., *Expression of the endoplasmic reticulum stress response marker, BiP, in the central nervous system of HIV-positive individuals.* Neuropathol Appl Neurobiol, 2007. **33**(6): p. 658-69.

115. Kaul, M. and S.A. Lipton, *Mechanisms of neuronal injury and death in HIV-1 associated dementia.* Curr HIV Res, 2006. **4**(3): p. 307-18.

116. Boven, L.A., *Macrophages and HIV-1-associated dementia.* Arch Immunol Ther Exp (Warsz), 2000. **48**(4): p. 273-9.

117. Kusdra, L., D. McGuire, and L. Pulliam, *Changes in monocyte/macrophage neurotoxicity in the era of HAART: implications for HIV-associated dementia.* AIDS, 2002. **16**(1): p. 31-8.

118. Levine, A.J., et al., *Transcriptome analysis of HIV-infected peripheral blood monocytes: gene transcripts and networks associated with neurocognitive functioning.* J Neuroimmunol, 2013. **265**(1-2): p. 96-105.

119. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.

120. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.

121. Salazar-Gonzalez, J.F., et al., *Relationship of plasma HIV-RNA levels and levels of TNF-alpha and immune activation products in HIV infection.* Clin Immunol Immunopathol, 1997. **84**(1): p. 36-45.

122. Shay, A.H., et al., *Impairment of antimicrobial activity and nitric oxide production in alveolar macrophages from smokers of marijuana and cocaine.* J Infect Dis, 2003. **187**(4): p. 700-4.

123. Levine, A.J., et al., *The longitudinal and interactive effects of HIV status, stimulant use, and host genotype upon neurocognitive functioning.* J Neurovirol, 2014. **20**(3): p. 243-57.

124. Lawton, M.P. and E.M. Brody, *Assessment of older people: self-maintaining and instrumental activities of daily living.* Gerontologist, 1969. **9**(3): p. 179-86.

125. Antinori, A., et al., *Updated research nosology for HIV-associated neurocognitive disorders.* Neurology, 2007. **69**(18): p. 1789-99.

126. Woods, S.P., et al., *Interrater reliability of clinical ratings and neurocognitive diagnoses in HIV.* J Clin Exp Neuropsychol, 2004. **26**(6): p. 759-78.

127. Letendre, S., *Central nervous system complications in HIV disease: HIV-associated neurocognitive disorder.* Top Antivir Med, 2011. **19**(4): p. 137-42.

128. Radloff, L.S., *The CES-D scale: A self-report depression scale for research in the general population.* Applied Psychological Measurement, 1977. **1**(1): p. 385-401.

129. Horvath, S. and J. Dong, *Geometric interpretation of gene coexpression network analysis.* PLoS Comput Biol, 2008. **4**(8): p. e1000117.

130. Langfelder, P., B. Zhang, and S. Horvath, *Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R.* Bioinformatics, 2008. **24**(5): p. 719-20.

131. Langfelder, P. and S. Horvath, *Eigengene networks for studying the relationships between co-expression modules.* BMC Syst Biol, 2007. **1**: p. 54.

132. Langfelder, P., et al., *Is My Network Module Preserved and Reproducible?* PLoS Comput Biol, 2011. **7**(1): p. e1001057.

133. Sun, B., et al., *Peripheral biomarkers do not correlate with cognitive impairment in highly active antiretroviral therapy-treated subjects with human immunodeficiency virus type 1 infection.* J Neurovirol, 2010. **16**(2): p. 115-24.

134. Kallianpur, A.R. and A.J. Levine, *Host Genetic Factors Predisposing to HIV-Associated Neurocognitive Disorder.* Curr HIV/AIDS Rep, 2014.

135. Levine, A.J., S.E. Panos, and S. Horvath, *Genetic, transcriptomic, and epigenetic studies of HIV-associated neurocognitive disorder.* J Acquir Immune Defic Syndr, 2014. **65**(4): p. 481-503.

136. Devlin, K.N., et al., *Neurocognitive effects of HIV, hepatitis C, and substance use history.* J Int Neuropsychol Soc, 2012. **18**(1): p. 68-78.

137. Levine, A.J., et al., *Predictors and impact of self-reported suboptimal effort on estimates of prevalence of HIV-associated neurocognitive disorders.* J Acquir Immune Defic Syndr, 2017.

138. Quach, A., et al., *Epigenetic clock analysis of diet, exercise, education, and lifestyle factors.* Aging, 2017.

139. Reas, E.T., et al., *Moderate, Regular Alcohol Consumption is Associated with Higher Cognitive Function in Older Community-Dwelling Adults.* The journal of prevention of Alzheimer's disease, 2016. **3**(2): p. 105-113.

140. Nguyen, T.P., V.M. Soukup, and B.B. Gelman, *Persistent hijacking of brain proteasomes in HIV-associated dementia.* Am J Pathol, 2010. **176**(2): p. 893-902.

141. Reddy, P.V., et al., *Inhibition of nuclear factor erythroid 2-related factor 2 exacerbates HIV-1 gp120-induced oxidative and inflammatory response: role in HIV associated neurocognitive disorder.* Neurochem Res, 2012. **37**(8): p. 1697-706.

142. Al-Harthi, L., *Interplay between Wnt/beta-catenin signaling and HIV: virologic and biologic consequences in the CNS.* J Neuroimmune Pharmacol, 2012. **7**(4): p. 731-9.

143. Lu, A.T., et al., *Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum.* Nature communications, 2016. **7**: p. 10561.

144. Lu, A.T., et al., *GWAS of epigenetic aging rates in blood reveals a critical role for TERT.* Nature Communications, 2018. **9**(1): p. 387.

145. Marioni, R.E., et al., *The epigenetic clock and telomere length are independently associated with chronological age and mortality.* International Journal of Epidemiology, 2016. **45**(2): p. 424-432.

146. Lowe, D., S. Horvath, and K. Raj, *Epigenetic clock analyses of cellular senescence and ageing.* Oncotarget, 2016. **7**(8): p. 8524.

147. Dawber, T.R., G.F. Meadors, and F.E. Moore Jr, *Epidemiological approaches to heart disease: the Framingham Study.* American Journal of Public Health and the Nations Health, 1951. **41**(3): p. 279-286.

148. Reynolds, L.M., et al., *Age-related variations in the methylome associated with gene expression in human monocytes and T cells.* Nature Communications, 2014. **5**: p. 5366.

149. Kawano, S., et al., *Gender-related alterations in plasma adrenomedullin level and its correlation with body weight gain.* Endocrine Connections, 2015. **4**(1): p. 43-49.

150. Kennedy, A., et al., *The Metabolic Significance of Leptin in Humans: Gender-Based Differences in Relationship to Adiposity, Insulin Sensitivity, and Energy Expenditure*.* The Journal of Clinical Endocrinology & Metabolism, 1997. **82**(4): p. 1293-1300.

151. Lin, Q. and W. Wagner, *Epigenetic Aging Signatures Are Coherently Modified in Cancer.* PLOS Genetics, 2015. **11**(6): p. e1005334.

152. Gierman, H.J., et al., *Whole-Genome Sequencing of the World's Oldest People.* PLOS ONE, 2014. **9**(11): p. e112430.

153. Yu, Q., et al., *DNA-Damage-Induced Type I Interferon Promotes Senescence and Inhibits Stem Cell Function.* Cell Reports, 2015. **11**(5): p. 785-797.

154. Kreienkamp, R., et al., *A Cell-Intrinsic Interferon-like Response Links Replication Stress to Cellular Aging Caused by Progerin.* Cell Reports, 2018. **22**(8): p. 2006-2015.

155. Baruch, K., et al., *Aging-induced type I interferon response at the choroid plexus negatively affects brain function.* Science, 2014. **346**(6205): p. 89-93.

156. Nishimura, K., et al., *Perturbation of Ribosome Biogenesis Drives Cells into Senescence through 5S RNP-Mediated p53 Activation.* Cell Reports, 2015. **10**(8): p. 1310-1323.

157. Dalet, A., E. Gatti, and P. Pierre, *Integration of PKR-dependent translation inhibition with innate immunity is required for a coordinated anti-viral response.* FEBS Letters, 2015. **589**(14): p. 1539-1545.

158. Geyer, P.E., et al., *Revisiting biomarker discovery by plasma proteomics.* Molecular systems biology, 2017. **13**(9): p. 942.

159. Scalbert, A., et al., *The food metabolome: a window over dietary exposure.* The American Journal of Clinical Nutrition, 2014. **99**(6): p. 1286-1308.

160. Chang, C.M., et al., *Biomarkers of Tobacco Exposure: Summary of an FDA-Sponsored Public Workshop.* Cancer Epidemiology Biomarkers & Prevention, 2017. **26**(3): p. 291-302.

161. Kechagias, S., et al., *Phosphatidylethanol Compared with Other Blood Tests as a Biomarker of Moderate Alcohol Consumption in Healthy Volunteers: A Prospective Randomized Study.* Alcohol and Alcoholism, 2015. **50**(4): p. 399-406.

162. Yusa, V., et al., *Occurrence of biomarkers of pesticide exposure in non-invasive human specimens.* Chemosphere, 2015. **139**: p. 91-108.

163. Neveu, V., et al., *Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors.* Nucleic Acids Research, 2017. **45**(D1): p. D979-D984.

164. Bruce, S.J., et al., *Investigation of Human Blood Plasma Sample Preparation for Performing Metabolomics Using Ultrahigh Performance Liquid Chromatography/Mass Spectrometry.* Analytical Chemistry, 2009. **81**(9): p. 3285-3296.

165. Lindahl, A., et al., *Tuning Metabolome Coverage in Reversed Phase LC–MS Metabolomics of MeOH Extracted Samples Using the Reconstitution Solvent Composition.* Analytical Chemistry, 2017. **89**(14): p. 7356-7364.

166. Wishart, D.S., et al., *HMDB 4.0: the human metabolome database for 2018.* Nucleic Acids Research, 2018. **46**(D1): p. D608-D617.

167. Mayr, M., et al., *Vascular proteomics: linking proteomic and metabolomic changes.* Proteomics, 2004. **4**(12): p. 3751-3761.

168. Oberbach, A., et al., *Combined Proteomic and Metabolomic Profiling of Serum Reveals Association of the Complement System with Obesity and Identifies Novel Markers of Body Fat Mass Changes.* Journal of Proteome Research, 2011. **10**(10): p. 4769-4788.

169. Strader, M.B., et al., *Efficient and Specific Trypsin Digestion of Microgram to Nanogram Quantities of Proteins in Organic−Aqueous Solvent Systems.* Analytical Chemistry, 2006. **78**(1): p. 125-134.

170. Fumin, L., S.C. M., and J.Q. C., *Accelerated tryptic digestion of proteins in plasma for absolute quantitation using a protein internal standard by liquid chromatography/tandem mass spectrometry.* Rapid Communications in Mass Spectrometry, 2009. **23**(5): p. 729-732.

171. Cai, X. and R. Li, *Concurrent profiling of polar metabolites and lipids in human plasma using HILIC-FTMS.* Scientific Reports, 2016. **6**: p. 36490.

172. Cajka, T. and O. Fiehn, *Increasing lipidomic coverage by selecting optimal mobile-phase modifiers in LC–MS of blood plasma.* Metabolomics, 2016. **12**(2): p. 34.

173. Keshishian, H., et al., *Multiplexed, quantitative workflow for sensitive biomarker discovery in plasma yields novel candidates for early myocardial injury.* Molecular & Cellular Proteomics, 2015: p. mcp. M114. 046813.

174. Bortey-Sam, N., et al., *Association between human exposure to heavy metals/metalloid and occurrences of respiratory diseases, lipid peroxidation and DNA damage in Kumasi, Ghana.* Environmental Pollution, 2018. **235**: p. 163-170.

175. Yao, Z.-P., P.A. Demirev, and C. Fenselau, *Mass Spectrometry-Based Proteolytic Mapping for Rapid Virus Identification.* Analytical Chemistry, 2002. **74**(11): p. 2529-2534.

176. Biba, M., et al., *Factors influencing the separation of oligonucleotides using reversed-phase/ion-exchange mixed-mode high performance liquid chromatography columns.* Journal of Chromatography A, 2013. **1304**: p. 69-77.

177. Studzińska, S., *Review on investigations of antisense oligonucleotides with the use of mass spectrometry.* Talanta, 2018. **176**: p. 329-343.