Title

Robust Parametric Statistical Methods and Software for Multi-Species and Multi-Phylum Age Prediction, and Quantification of Human HIV-Induced and HAART-Mitigated Age Acceleration

Author

Zoller, Joseph Alan

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Robust Parametric Statistical Methods and Software for Multi-Species and Multi-Phylum Age Prediction, and Quantification of Human HIV-Induced and HAART-Mitigated Age Acceleration

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Biostatistics

by

Joseph A. Zoller

2024

ABSTRACT OF THE DISSERTATION

Robust Parametric Statistical Methods and Software for
Multi-Species and Multi-Phylum Age Prediction, and
Quantification of Human HIV-Induced and HAART-Mitigated
Age Acceleration

by

Joseph A. Zoller
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2024
Professor Christina Ramirez, Chair

Epigenetic clocks, DNA methylation-based biomarkers, accurately measure age within specific species and tissues but face challenges in multi-species and non-mammalian contexts. This study aims to enhance the construction of biologically meaningful, multi-species epigenetic clocks, particularly applicable to both humans and animals. Utilizing supervised machine learning on DNA methylation data, this research incorporates biological information, such as genome mapping and life history traits, to improve clock accuracy and interpretability. Findings include the development of diverse epigenetic clocks for various mammalian species, enabled by robust statistical methods and a reproducible software pipeline. An R package, MammalMethylClock, was introduced to facilitate the construction, assessment, and application of these clocks, incorporating existing models from the Mammalian Methylation Consortium. This package supports translational biomedical research by enhancing the study of age-related chronic diseases across different model organisms.

Moreover, the remainder of this study examines the research problem of accelerated aging in people living with HIV (PLWH) and the mitigating effects of highly active antiretroviral therapy (HAART). Previous research indicates that HIV-related age acceleration begins shortly after seroconversion. This longitudinal study within the Multicenter AIDS Cohort Study (MACS) examines 60 PLWH and 60 matched controls across four specific time points: pre-seroconversion, closest to seroconversion, pre-HAART, and post-1-3 years of HAART. The study uses comprehensive epigenetic clock data and flow cytometry measures, including naive, senescent, activated, and total CD4/CD8 counts. The method involves analyzing epigenetic aging through five different aging clocks at these time points to assess the impact of HIV on aging and the partial restoration effects of HAART on immune function and aging patterns. By tracking these changes, the study aims to identify specific aging patterns associated with HIV and guide potential interventions. The findings reveal that HIV infection significantly accelerates epigenetic aging in most clocks, and this acceleration is only partially mitigated by HAART. This suggests that HIV contributes to the premature onset of clinical aging through its profound effects on the immune system, even when viral replication is controlled. The study underscores the need for interventions to enhance the healthspan of PLWH and offers insights into the broader aging process.

The dissertation of Joseph A. Zoller is approved.

Zhe Fei

David Elashoff

Jason Ernst

Christina Ramirez, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS OF THE DISSERTATION

FIGURES

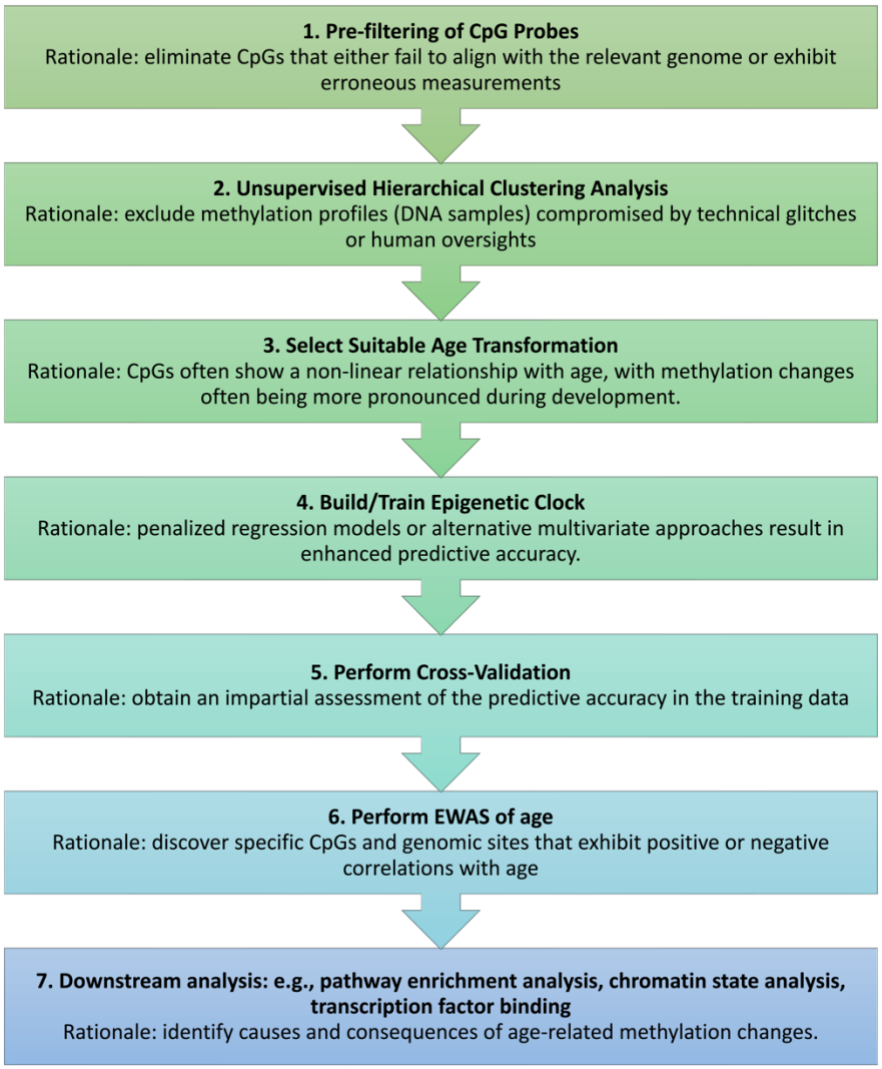

Figure 4-1. Flowchart for Constructing and Evaluating Epigenetic Clocks. This flowchart outlines the foundational stages of creating and appraising epigenetic clocks, alongside the typical procedures and their underlying logic. For a deeper understanding of the causes and effects of epigenetic clocks, one might employ functional enrichment tools and various other computational resources.

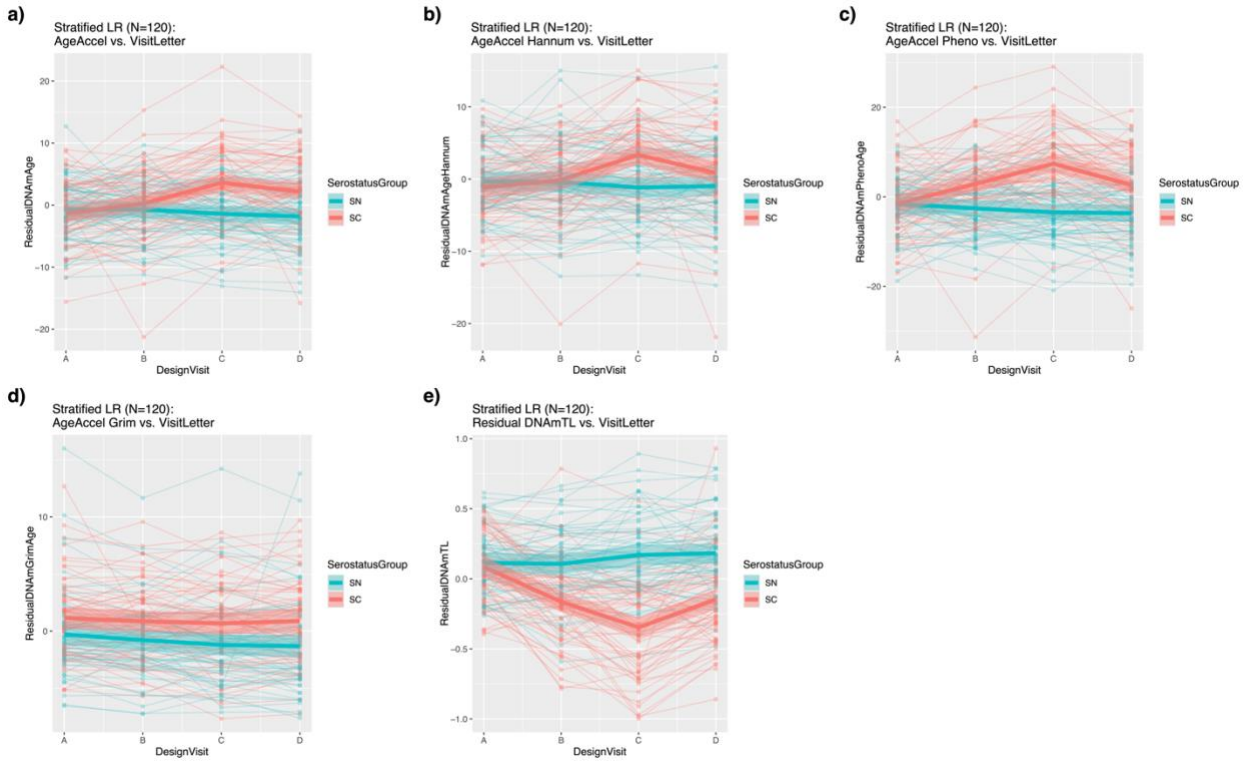

Figure 5-1: Multiple epigenetic measures in peripheral blood mononuclear cells (PBMC) demonstrate significant changes in biologic aging after initial HIV infection, with partial reversal after initiation of HAART, compared to age-matched HIV-uninfected persons. Longitudinal PBMC samples from men before (visit A) and after (visit B) documented HIV infection and seroconversion (SC) and before (visit C) and after (visit D) initiation of HAART, and from matched (chronologic age, Hepatitis C status, and time interval) persistently HIV seronegative men (SN), were evaluated for biologic aging by five different age adjusted epigenetic clocks: a) Age Acceleration Residual (AAR), b) Extrinsic Epigenetic Age Acceleration (EEAA), c) Phenotypic Epigenetic Age Acceleration (PEAA), d) Grim Epigenetic Age Acceleration (GEAA), and e) age-adjusted DNA methylation-based estimate of telomere length (aaDNAmTL). Each panel shows spaghetti plots (heavy line = fitted LMER line, translucent band = 95th percentile confidence band) for SC (red) and SN (cyan) participants at visits A through D; light lines are trajectories for individual subjects, with same coloring. 60 SC and 60 SN subjects were evaluated.

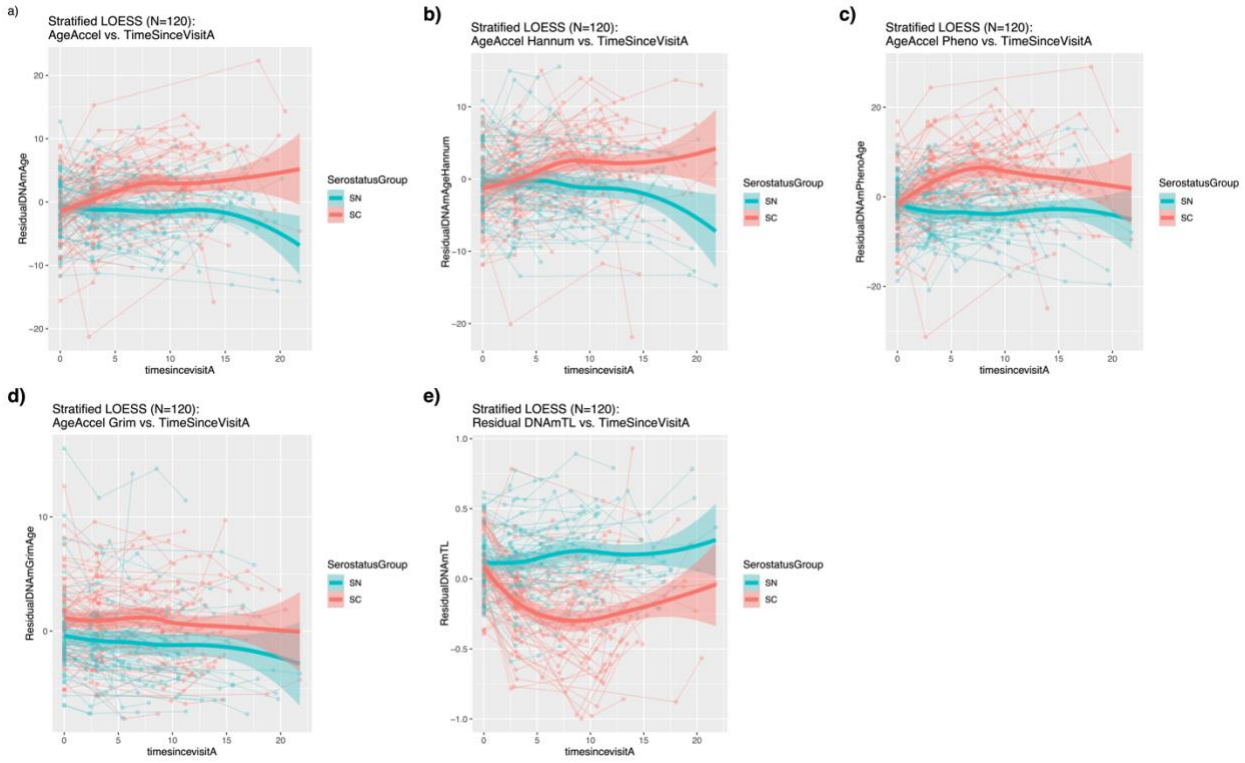

Figure 5-**2: Multiple epigenetic measures in peripheral blood mononuclear cells (PBMC) demonstrate significant changes in biologic aging over continuous time after initial HIV infection, with partial reversal after initiation of HAART, compared to age-matched HIV-uninfected persons.** Longitudinal PBMC samples from men before (visit A) and after (visit B) documented HIV infection and seroconversion (SC) and before (visit C) and after (visit D) initiation of HAART, and from matched (chronologic age, Hepatitis C status, and time interval) persistently HIV seronegative men (SN), were evaluated for biologic aging by five different age adjusted epigenetic clocks: a) Age Acceleration Residual (AAR), b) Extrinsic Epigenetic Age Acceleration (EEAA), c) Phenotypic Epigenetic Age Acceleration (PEAA), d) Grim Epigenetic Age Acceleration (GEAA), and e) age-adjusted DNA methylation-based estimate of telomere length (aaDNAmTL). Each panel shows spaghetti plots (heavy line = fitted LOESS line, translucent band = 95th percentile confidence band) for SC (red) and SN (cyan) participants at visits A through D, as functions of time since visit A; light lines are trajectories for individual subjects, with same coloring. 60 SC and 60 SN subjects were evaluated.

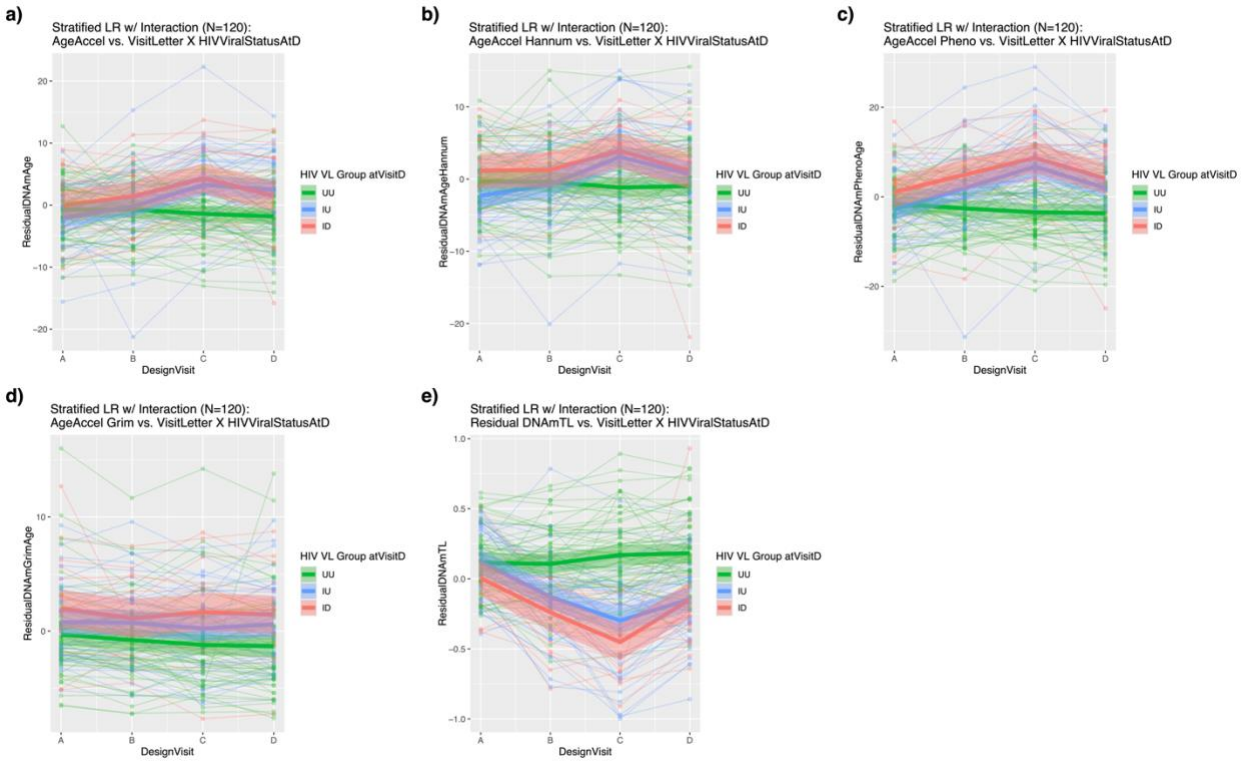

Figure 5-3: Multiple epigenetic measures demonstrate similar changes in aging at all post-HIV infection visits and at the post-HAART visit in SC, regardless of final plasma HIV viral load (HIV VL) at visit D, compared to age-matched HIV-uninfected persons. The changes in epigenetic measures, grouped by HIV detectability at post-HAART visit (visit D) are shown for: a) AAR, b) EEAA, c) PEAA, d) GEAA, and e) aaDNAmTL. HIV detectability group is a trichotomous variable (UU = uninfected and undetectable, IU = infected and undetectable, ID = infected and detectable). Each panel shows spaghetti plots (heavy line = fitted LMER line, translucent band = 95th percentile confidence band) for SC HIV-detectable at visit D (red), SC HIV-undetectable at visit D (blue), and SN uninfected (green) participants at visits A through D; light lines are trajectories for individual subjects, with same coloring. 19 SC in ID, 41 SC in IU, and 60 SN in UU subjects were evaluated.

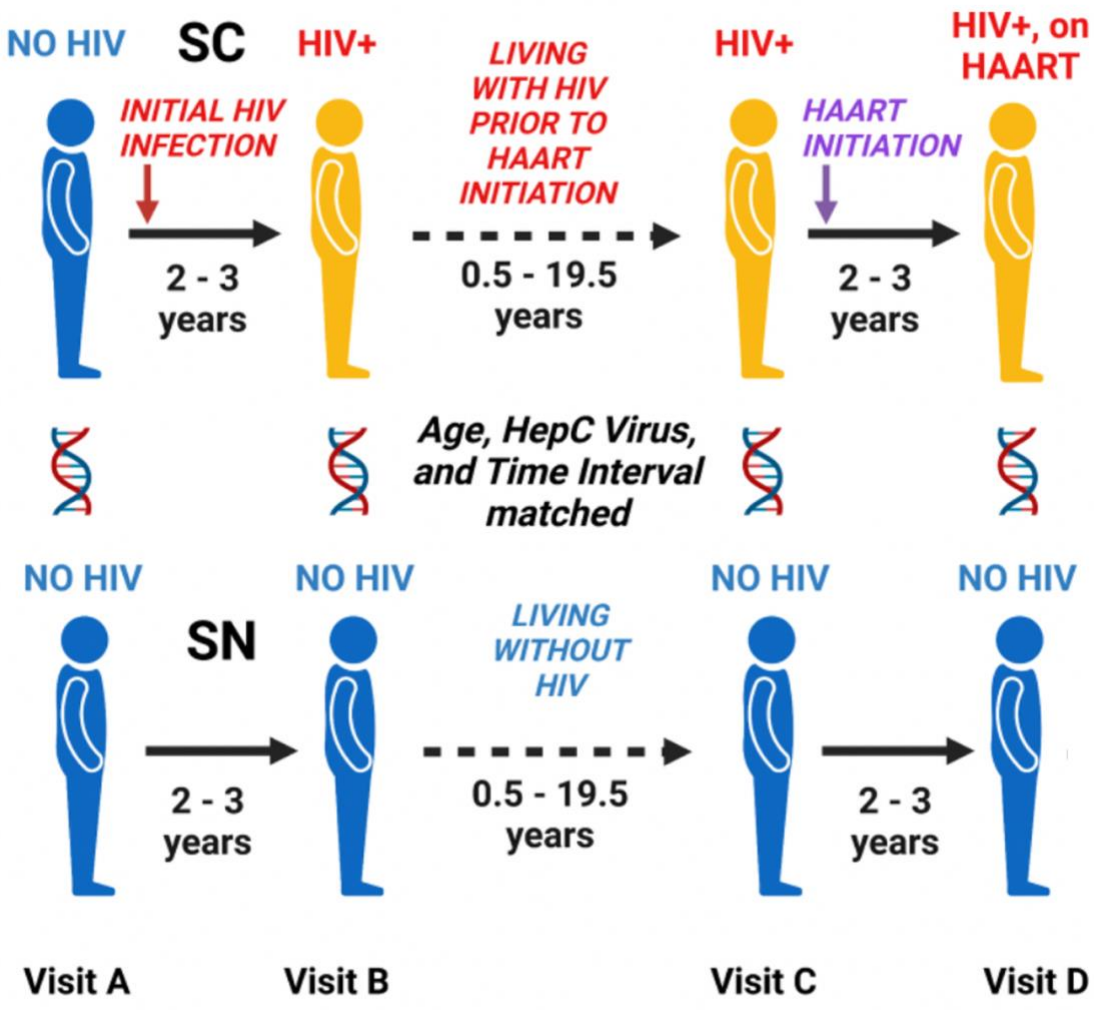

Figure 5-4: Participant Demographics and Characteristics

TABLES

Table 5-1: Demographics and characteristics of HIV seroconverter (SC) and matched HIV seronegative (SN) participants from the Multicenter AIDS Cohort Study (MACS)

Participants	**SN** [a], n (%) or mean (SD)				**SC** [b], n (%) or mean (SD)			
White Race	47 (78.3%)				53 (88.3%)			
1+ year of College Education	53 (88.3%)				56 (93.3%)			
Visit A to Visit B, years	2.8 (0.6)				3 (0.4)			
Visit A to Visit C, years	8.3 (4.2)				9.3 (3.4)			
Visit A to Visit D, years	11 (4.4)				12.1 (3.5)			
At each visit	**Visit A**	**Visit B**	**Visit C**	**Visit D**	**Visit A**	**Visit B**	**Visit C**	**Visit D**
Age, years	36.4 (9)	39.2 (8.9)	44.7 (8.7)	47.4 (8.7)	35.5 (8.7)	38.5 (8.6)	44.9 (8.6)	47.7 (8.7)
Hepatitis B Virus surface antigen-positive	1 (1.7%) n=58	2 (3.4%) n=58	2 (3.3%)	2 (3.3%)	1 (1.7%)	0 (0%)	0 (0%)	1 (1.7%) n=59
Hepatitis C Virus RNA-positive	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Body Mass Index, kg/m^2	25.4 (5.1) n=59	25.6 (4) n=57	27.4 (5) n=57	27.8 (4.9) n=59	24.4 (3.5)	25.3 (3.7) n=57	25.9 (4) n=55	26 (3.9) n=55
Smoking, cumulative pack years	8 (14.8)	8.5 (15.3)	9.1 (16.1)	9.5 (16.3)	10.7(13.9) n=59	11.6 (15) n=56	12.9(16.7) n=57	13.2(17.1) n=59
Plasma HIV Viral Load [c], copies/mL	N/A	N/A	N/A	N/A	N/A	29,994 (40,301)	58,893 (119,006) n=56	5,884 (15,569)
Estimated time since HIV infection [d], years	N/A	N/A	N/A	N/A	N/A	2.3 (0.4)	8.6 (3.4)	11.5 (3.5)
Estimated time since HAART initiation [e], years	N/A	N/A	N/A	N/A	N/A	N/A	N/A	2.1 (0.4)

a: n=60 SN at Visits A through D, unless indicated otherwise
b: n=60 SC at Visits A through D, unless indicated otherwise
c: For 10 SC missing HIV VL at Visit B, HIV VL from the closest MACS study visit 3-6 months prior to Visit B was used
d: Date of HIV infection estimated as midpoint between the last MACS study visit that was HIV seronegative and HIV VL undetectable (if VL data were available) and the first MACS study visit with either HIV-positive serostatus or detectable HIV VL, whichever came first
e: Date of HAART initiation estimated as midpoint between the last MACS study visit that was before HAART initiation and the first MACS study visit that was after HAART initiation

Table 5-2: Potential contribution of demographic and behavioral covariates to epigenetic measures over time, using mixed effects models

Potential Contributors to Epigenetic Measures	F-value *(p-value)* [a]				
	AAR	**EEAA**	**PEAA**	**GEAA**	**aaDNAmTL**
Study Visit, Visits A-D	6.56 (***<0.001***)	5.19 (***0.0016***)	8.56 (***<0.001***)	3.83 (***0.01***)	19.21 (***<0.001***)
HIV Serostatus Group, SC vs SN [b]	7.8 (***0.0063***)	2.34 (0.13)	26.14 (***<0.001***)	11.27 (***0.0011***)	28.7 (***<0.001***)
Study Visit * HIV Serostatus Group	16.03 (***<0.001***)	14.4 (***<0.001***)	19.24 (***<0.001***)	0.5 (0.68)	34.51 (***<0.001***)
Race, white vs non-white	0.66 (0.42)	0.81 (0.37)	0.05 (0.82)	6.48 (***0.013***)	0.3 (0.59)
BMI, kg/m^2	2.19 (0.14)	2.16 (0.14)	0.08 (0.77)	0.61 (0.43)	0.44 (0.51)
Smoking, cumulative pack years	0.84 (0.36)	3.3 (0.072)	1.46 (0.23)	32.87 (***<0.001***)	1.34 (0.25)

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length, BMI = Body Mass Index

a: Type III F-values and (Pr >F) p-values (p-values in italics, bold if <0.05) from mixed models incorporating all potential covariates for all participants at all visits (n=400 out of 480 total observations due to missing data for some covariates, 54 SN and 46 SC) in a single model

b: HIV serostatus groups classified as SC (became HIV infected and seroconverted between visits A and B) vs SN (persistently HIV uninfected and seronegative at visits A through D)

Table 5-3: Mean absolute T cell counts of the SC and SN groups, at Visits A through D

T cell population	Visit A [a], Mean (SE) n		Visit B [b], Mean (SE) n		Visit C [c], Mean (SE) n		Visit D [d], Mean (SE) n	
	SN	**SC**	**SN**	**SC**	**SN**	**SC**	**SN**	**SC**
CD4 T cells, cells/mm^3	991 (463) n=58	1117 (419) n=53	1042 (396) n=56	658 (240) n=59	1005 (353) n=60	362 (215) n=59	983 (322) n=60	537 (260) n=60
CD8 T cells, cells/mm^3	625 (302) n=58	627 (249) n=53	684 (335) n=56	892 (417) n=59	614 (295) n=60	1026 (463) n=59	596 (316) n=60	994 (471) n=60
Naïve (CD45RA+CCR7+) CD4 T cells, cells/mm^3	401 (332) n=57	402 (205) n=53	385 (239) n=55	249 (141) n=59	363 (246) n=59	128 (109) n=57	350 (231) n=60	209 (142) n=59
Naïve (CD45RA+CCR7+) CD8 T cells, cells/mm^3	208 (143) n=57	207 (114) n=53	215 (127) n=55	135 (76) n=59	182 (130) n=59	108 (79) n=57	168 (123) n=60	178 (119) n=59
Senescent (CD28-CD57+) CD4 T cells, cells/mm^3	34 (30) n=57	51 (63) n=53	43 (41) n=55	48 (59) n=59	48 (56) n=59	36 (42) n=57	45 (52) n=60	42 (48) n=59
Senescent (CD28-CD57+) CD8 T cells, cells/mm^3	117 (93) n=57	114 (93) n=53	129 (120) n=55	153 (134) n=59	119 (102) n=59	233 (217) n=57	127 (134) n=60	229 (181) n=59
Activated (HLA DR+CD38+) CD4 T cells, cells/mm^3	29 (20) n=56	29 (13) n=51	32 (20) n=53	33 (17) n=57	28 (16) n=59	27 (17) n=57	27 (14) n=58	23 (10) n=59
Activated (HLA DR+CD38+) CD8 T cells, cells/mm^3	23 (17) n=56	26 (18) n=51	30 (34) n=53	163 (119) n=57	22 (19) n=59	198 (135) n=57	21 (14) n=58	86 (79) n=59

a: all participants HIV-uninfected at Visit A, matched on age and hepatitis C status
b: SC recently HIV-infected, SN persistently HIV-uninfected at matched time intervals at Visit B
c: SC shortly before beginning HAART, SN persistently HIV-uninfected at matched time intervals at Visit C
d: SC recently on HAART, SN persistently HIV-uninfected at matched time intervals at Visit D

Table 5-4: Mean percentages of T cell subsets within live lymphocytes of the SC and SN groups, at Visits A through D

	Visit A [a], Mean (SE) n		**Visit B [b]**, Mean (SE) n		**Visit C [c]**, Mean (SE) n		**Visit D [d]**, Mean (SE) n	
T cell population	**SN**	**SC**	**SN**	**SC**	**SN**	**SC**	**SN**	**SC**
CD4 T cells, %	0.593 (0.0926) n=49	0.618 (0.0926) n=53	0.595 (0.109) n=53	0.427 (0.143) n=52	0.629 (0.0976) n=53	0.286 (0.146) n=52	0.631 (0.0979) n=49	0.378 (0.159) n=52
CD8 T cells, %	0.344 (0.0866) n=49	0.327 (0.0805) n=53	0.345 (0.104) n=53	0.511 (0.135) n=52	0.319 (0.088) n=53	0.647 (0.145) n=52	0.313 (0.0902) n=49	0.559 (0.153) n=52
Naïve (CD45RA+CCR7+) CD4 T cells, %	0.16 (0.0786) n=50	0.149 (0.0587) n=56	0.15 (0.0728) n=55	0.113 (0.0643) n=54	0.153 (0.0844) n=53	0.0689 (0.0559) n=52	0.142 (0.0739) n=49	0.096 (0.0689) n=52
Naïve (CD45RA+CCR7+) CD8 T cells, %	0.0761 (0.0377) n=50	0.073 (0.0364) n=56	0.07 (0.0349) n=55	0.0558 (0.0328) n=54	0.0635 (0.0378) n=53	0.0447 (0.031) n=52	0.0548 (0.0341) n=49	0.0645 (0.0437) n=52
Senescent (CD28-CD57+) CD4 T cells, %	0.0133 (0.0112) n=50	0.0167 (0.0202) n=56	0.0155 (0.0132) n=55	0.0183 (0.0186) n=54	0.0181 (0.0165) n=53	0.0167 (0.0174) n=52	0.0209 (0.0206) n=49	0.0178 (0.0203) n=52
Senescent (CD28-CD57+) CD8 T cells, %	0.0445 (0.0325) n=50	0.038 (0.0279) n=56	0.0471 (0.0411) n=55	0.0619 (0.042) n=54	0.0415 (0.0339) n=53	0.0929 (0.07) n=52	0.0468 (0.0438) n=49	0.0831 (0.0538) n=52
Activated (HLA DR+CD38+) CD4 T cells, %	0.0108 (0.00357) n=49	0.0107 (0.00414) n=53	0.0116 (0.00513) n=53	0.0139 (0.00569) n=52	0.0118 (0.00474) n=53	0.0136 (0.0073) n=52	0.0112 (0.00364) n=49	0.0102 (0.00462) n=52
Activated (HLA DR+CD38+) CD8 T cells, %	0.00857 (0.00565) n=49	0.00943 (0.00695) n=53	0.0102 (0.0126) n=53	0.0609 (0.0396) n=52	0.0082 (0.00685) n=53	0.0847 (0.0501) n=52	0.00817 (0.0071) n=49	0.0288 (0.0227) n=52

a: all participants HIV-uninfected at Visit A, matched on age and hepatitis C status
b: SC recently HIV-infected, SN persistently HIV-uninfected at matched time intervals at Visit B
c: SC shortly before beginning HAART, SN persistently HIV-uninfected at matched time intervals at Visit C
d: SC recently on HAART, SN persistently HIV-uninfected at matched time intervals at Visit D

Table 5-5: Potential contribution of absolute T cell counts to epigenetic measures over time, using mixed effects models

Potential Contributors to Epigenetic Measures	F-value *(p-value)* [a]				
	AAR	**EEAA**	**PEAA**	**GEAA**	**aaDNAmTL**
Study Visit, Visits A-D	0.53 (0.66)	1.2 (0.31)	1.34 (0.26)	3.79 (***0.011***)	3.82 (***0.011***)
HIV Serostatus Group, SC vs SN [b]	1.44 (0.23)	1.2 (0.28)	1.28 (0.26)	3.95 (0.05)	3.36 (0.07)
Study Visit * HIV Serostatus Group	2.15 (0.09)	2.75 (***0.044***)	3.87 (***0.01***)	0.64 (0.59)	2.89 (***0.036***)
Absolute Count CD8 T cells [c], log(cells/mm^3)	0.61 (0.43)	1.32 (0.25)	0.65 (0.42)	0.34 (0.56)	3.8 (0.052)
% Naïve (CD45RA+CCR7+) CD4 T cells, log(%)	16.5 (***<0.001***)	9.02 (***0.0029***)	21.87 (***<0.001***)	3.66 (0.057)	30.78 (***<0.001***)
% Naïve (CD45RA+CCR7+) CD8 T cells, log(%)	0.09 (0.76)	33.48 (***<0.001***)	11.99 (***<0.001***)	0.93 (0.33)	20.73 (***<0.001***)
% Senescent (CD28-CD57+) CD8 T cells, log(%)	13.72 (***<0.001***)	17.16 (***<0.001***)	3.19 (0.075)	0.03 (0.86)	33.64 (***<0.001***)
% Activated (HLA DR+CD38+) CD8 T cells, log(%)	0.003 (0.96)	3.56 (0.06)	3.65 (0.057)	0.58 (0.45)	3.65 (0.057)

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Type III F-values and (Pr >F) p-values (p-values in italics, bold if <0.05) from mixed models incorporating all potential covariates for all participants at all visits (n=288 out of 480 total observations due to missing data for some covariates, 34 SN and 38 SC) in a single model

b: HIV serostatus groups classified as SC (became HIV infected and seroconverted between visits A and B) vs SN (persistently HIV uninfected and seronegative at visits A through D)

c: Absolute counts of T cell subsets as described in Methods and Table S6; all cell counts and percentages were natural log-transformed (log) for analyses

Table 5-6: Regression summary of epigenetic measures over time by HIV detectability status at Visit D, using mixed effects models

Regression Fixed Effects	Coefficient Estimates (Standard Errors) [a]				
	AAR	**EEAA**	**PEAA**	**GEAA**	**aaDNAmTL**
(Intercept)	-0.7 (0.69)	-0.31 (0.7)	-1.64 (0.97)	-0.29 (0.48)	**0.11 (0.04)**
Study Visit, Visit B vs A	-0.03 (0.58)	-0.09 (0.58)	-0.93 (0.84)	-0.49 (0.3)	-0.01 (0.03)
Study Visit, Visit C vs A	-0.71 (0.58)	-0.86 (0.58)	-1.82 (0.84)	**-0.91 (0.3)**	0.06 (0.03)
Study Visit, Visit D vs A	-1.12 (0.58)	-0.67 (0.58)	-2.02 (0.84)	**-1.03 (0.3)**	0.07 (0.03)
HIV Detectability Group, ID vs UU [b]	0.65 (1.41)	1.49 (1.42)	2.71 (1.99)	2.26 (0.97)	-0.11 (0.07)
HIV Detectability Group, IU vs UU [b]	-1.34 (1.08)	-2.01 (1.1)	-1.16 (1.53)	1.08 (0.75)	0.01 (0.06)
Visit B vs A * ID vs UU	1.46 (1.19)	0.16 (1.19)	**4.8 (1.71)**	-0.36 (0.61)	**-0.24 (0.06)**
Visit B vs A * IU vs UU	1.86 (0.91)	1.64 (0.91)	**5.66 (1.31)**	0.44 (0.47)	**-0.24 (0.05)**
Visit C vs A * ID vs UU	**5.08 (1.19)**	**3.61 (1.19)**	**9.46 (1.71)**	0.6 (0.61)	**-0.51 (0.06)**
Visit C vs A * IU vs UU	**6.02 (0.91)**	**6.23 (0.91)**	**11.4 (1.31)**	0.35 (0.47)	**-0.48 (0.05)**
Visit D vs A * ID vs UU	2.65 (1.19)	0.6 (1.19)	**5.05 (1.71)**	0.51 (0.61)	**-0.22 (0.06)**
Visit D vs A * IU vs UU	**5.62 (0.91)**	**3.71 (0.91)**	**6.78 (1.31)**	0.85 (0.47)	**-0.34 (0.05)**

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Coefficient values and standard errors for fixed effects (bold if Wald test p-value <0.01) from mixed models (n=480 all observations) in a single model

b: HIV detectability groups classified as ID (“infected and detectable”, became HIV infected, seroconverted, and still detectable (VL > 50) after HAART) vs IU (“infected and undetectable”, became HIV infected, seroconverted, but now undetectable (VL < 50) after HAART) vs UU (“uninfected and undetectable”, persistently HIV uninfected and seronegative at visits A through D)

VITA

Education:

- **Carnegie Mellon University**
  Mellon College of Science, Pittsburgh, PA
  M.S. and B.S. in Mathematical Science, Aug 2014 – May 2018

Research Employment:

1. University of California, Los Angeles — Fielding School of Public Health
   Graduate Student Researcher for Horvath Lab, Dept. of Biostatistics & Dept. of Human Genetics
   (*Jan 2019 – Jun 2022*)
   Performed preliminary and project-specific data analyses for many collaborative research projects, as well as several lab research projects. Built DNA methylation-based biomarkers for age, called "epigenetic clocks" (based on original work of Dr. Steve Horvath), using high-quality DNA methylation data sets coupled with age, species, and tissue information, coming from a variety of different mammalian and non-mammalian species. Analyses during collaborative research included building accurate biomarkers (sparse regression models) for one or multiple species, and using leave-one-out or leave-one-group-out analysis to identify outliers and evaluate performance of model building steps, like probe pre-selection and outcome transforming. Resultant in co-authorship on many different research publications for a variety of animal models. Also assisted with the initial development of a third-generation clock, the "Universal Mammalian Clock", that is trained to predict age accurately in all mammals and all tissue types. Expectant to be author on at least two papers due to contributions.

Selected Publications:

Chiavellini, P., Lehmann, M., Canatelli Mallat, M., **Zoller, J. A**., Herenu, C. B., Morel, G. R., . . . Goya, R. G. (2022). Hippocampal DNA Methylation, Epigenetic Age, and Spatial Memory Performance in Young and Old Rats. *The Journals of Gerontology: Series A*, *77*(12), 2387-2394. https://doi.org/10.1093/gerona/glac153

Chiavellini, P., Lehmann, M., Gallardo, M. D., Canatelli Mallat, M., Pasquini, D. C., **Zoller, J. A.**, . . . Goya, R. G. (2024). Young Plasma Rejuvenates Blood Dna Methylation Profile, Extends Mean Lifespan And Improves Physical Appearance In Old Rats. *The Journals of Gerontology: Series A*. https://doi.org/10.1093/gerona/glae071

Horvath, S., Haghani, A., Macoretta, N., Ablaeva, J., **Zoller, J. A.**, Li, C. Z., . . . Gorbunova, V. (2021). DNA methylation clocks tick in naked mole rats but queens age more slowly than nonbreeders. *Nature Aging*, *2*(1), 46-59. https://doi.org/10.1038/s43587-021-00152-1

Horvath, S., Singh, K., Raj, K., Khairnar, S. I., Sanghavi, A., Shrivastava, A., **Zoller, J. A.**, . . . Katcher, H. L. (2023). Reversal of biological age in multiple rat organs by young porcine plasma fraction. *GeroScience*. https://doi.org/10.1007/s11357-023-00980-6

Kordowitzki, P., Haghani, A., **Zoller, J. A.**, Li, C. Z., Raj, K., Spangler, M. L., & Horvath, S. (2021). Epigenetic clock and methylation study of oocytes from a bovine model of reproductive aging. *Aging Cell*, *20*(5), e13349. https://doi.org/10.1111/acel.13349

Lu, A. T., Fei, Z., Haghani, A., Robeck, T. R., **Zoller, J. A.**, Li, C. Z., . . . Horvath, S. (2023). Universal DNA methylation age across mammalian tissues. *Nature Aging*. https://doi.org/10.1038/s43587-023-00462-6

Sugrue, V. J., **Zoller, J. A.**, Narayan, P., Lu, A. T., Ortega-Recalde, O. J., Grant, M. J., . . . Horvath, S. (2021). Castration delays epigenetic aging and feminizes DNA methylation at androgen-regulated loci. *eLife*, *10*, e64932. https://doi.org/10.7554/eLife.64932

Watson, G. L., Xiong, D., Zhang, L., **Zoller, J. A.**, Shamshoian, J., Sundin, P., . . . Ramirez, C. M. (2021). Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. *PLOS Computational Biology*, *17*(3), e1008837. https://doi.org/10.1371/journal.pcbi.1008837

Wilkinson, G. S., Adams, D. M., Haghani, A., Lu, A. T., **Zoller, J.**, Breeze, C. E., . . . Flores-Martínez, J. J. (2021). Author Correction: DNA methylation predicts age and provides insight into exceptional longevity of bats. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-23129-5

Xiong, D., Zhang, L., Watson, G. L., Sundin, P., Bufford, T., **Zoller, J. A.**, . . . Ramirez, C. M. (2020). Pseudo-likelihood based logistic regression for estimating COVID-19 infection and case fatality rates by gender, race, and age in California. *Epidemics*, *33*, 100418. https://doi.org/10.1016/j.epidem.2020.100418

Zoller, J., & Horvath, S. (2024). MammalMethylClock R package: Software for DNA Methylation-Based epigenetic clocks in mammals. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btae280

Zoller, J. A., Parasyraki, E., Lu, A. T., Haghani, A., Niehrs, C., & Horvath, S. (2023). DNA methylation clocks for clawed frogs reveal evolutionary conservation of epigenetic aging. *GeroScience*. https://doi.org/10.1007/s11357-023-00840-3

Conference Presentations:

1. Schachtschneider, K. M., Schook, L. B., Meudt, J. J., Shanmuganayagam, D., Zoller, J. A.*, Haghani, A., Li, C. Z., Zhang, J., Yang, A., Raj, K., Horvath, S.
   "Porcine epigenetic clocks: Tools for investigating the epigenetics of aging and disease" (*June 12, 2022*)
   Swine in Biomedical Research 2022 — Plenary Session Speaker, Oral Presentation
   University of Wisconsin, Madison

Honors and Awards:

1. The Celia and Joseph Blann Fellowship, 2023-2024 (*for academic excellence within the School of Public Health*)
   The Celia and Joseph Blann Fellowship, which includes a $10,000 stipend, is awarded annually to doctoral students in the UCLA School of Public Health on the basis of academic excellence. It was established by Annette Blann to honor the memory of her parents. A 1954 UCLA graduate, she was committed to the enhancement of health and well-being.
2. Student Employee of the Year for 2017-2018 (*for work as EXCEL Leader & SI/EXCEL Student Supervisor*)
   Nominated and Won for dedicated and diligent work as an EXCEL Leader and SI/EXCEL Student Supervisor, supporting 2 different foundational math courses while simultaneously working as a supervisor for other SI/EXCEL leaders.

CHAPTER 1: INTRODUCTION AND MOTIVATION

Epigenetic clocks are broadly defined as DNA methylation-based biomarkers for age, and are accurate within the set of species and tissue types that they are trained on. The construction of epigenetic clocks involves the careful use of supervised machine learning applied to DNA methylation data along CpG islands, along with the manual incorporation or consideration of biological information like array probe mapping to the genome and life history traits. The ability to accurate quantify and measure biological aging in various mammalian (and non-mammalian) model organisms is an important part of improving translational biomedical research, when these model organisms are used for studying chronic diseases that are also known to be age-related diseases in humans. However, there are ongoing challenges with the task of building multi-species and non-mammalian clocks, especially regarding accuracy and biological interpretability. This research aims to identify strategies for building accurate multi-species clocks, especially those that work on humans with other animals, and to modify the approach to building epigenetic clocks to be more innately biologically meaningful. This chapter will provide an introduction to this research, by discussing the background, context, and motivation of developing these biomarkers for age.

Since the 1960's, long before the development of the first epigenetic clocks in humans, a large body of literature demonstrated that DNA methylation levels at millions of CpG dinucleotide sites genome wide were seen to consistently change with age (Ahuja, et al., 1998; Bell, et al., 2012; Berdyshev, et al., 1967; Christensen, et al., 2009; Hernandez, et al., 2011; Maegawa, et al., 2010; Rakyan, et al., 2010; Teschendorff, et al., 2010). With the development of sophisticated statistical models for high-dimensional data and supervised machine learning, the computational power to train these models extremely quickly, and the advancement in array technology to be able to measure DNA methylation levels at hundreds of thousands of sites simultaneously, came the ability to "reconstruct" an organism's age. In other words, given enough samples across the lifespan of an animal, and given enough sites that consistently gain

or lose methylation with age, one could construct a model that accurately estimates age using DNA methylation levels at select sites. The first notable example of this was the development of the first multi-tissue epigenetic clock for humans, which measures intrinsic biological age using only 353 CpG sites, and was dubbed the “Horvath clock” (Horvath, 2013).

As it would turn out, the most consequential and age-associated changes in DNA methylation occur on special loci, known as “CpG islands”, where CpG dinucleotides occur in high frequency over the span of several hundred base-pairs. CpG islands are known to contain or be adjacent to transcriptional regulatory elements, especially promoter regions, in the DNA of most mammals and many other animals. In addition, a large number of CpG islands were found to be evolutionarily conserved across most mammalian species. Given this knowledge, a DNA methylation array chip was developed, dubbed the HorvathMammalMethylChip40, and commonly referred to as the “mammalian array”; this chip captures methylation levels at over 37,000 CpG island sites, and it was designed such that all of these sites are evolutionarily conserved across most mammalian species (Arneson, et al., 2022). With the construction of this array, it provides a uniform set of data points on which to build all intrinsic epigenetic clocks, and thus allows us to combine methylation data from different species, and cross-apply clocks trained in one species to methylation data in another species.

While these scientific developments and insights may be interesting to some, one might ask what the significance is of being able to estimate age without looking at a date of birth. Besides being a statistically and biologically fascinating tool to study, biomarkers for age are of significant interest to clinical and biomedical researchers, and development of epigenetic clocks for measuring intrinsic aging has resulted in significant insight into the molecular mechanistic relationship between aging and many different chronic diseases. Specifically, the quantity of intrinsic “age acceleration”, the difference between the biomarker’s estimate of a person’s age and the person’s chronological age, is clinically significant, as positive intrinsic age acceleration is linked with Alzheimer’s disease (Levine, et al., 2015; Lu, et al., 2017), Cancer (Ambatipudi, et

al., 2017; Dugué, et al., 2018; Levine, et al., 2015; Zheng, et al., 2016), Obesity (Levine, et al., 2018; Quach, et al., 2017), Osteoarthritis (Vidal, et al., 2016), Parkinson disease (Horvath and Ritz, 2015), and many more chronic disease and conditions.

Moreover, certain diseases are thought to or know to rapidly accelerate epigenetic aging, because these diseases do substantial and/or organism-wide damage, often by attacking the immune system. One such prominent disease is HIV, which is known to cause significant epigenetic age acceleration when left untreated, even within the first few years of infection (Breen, et al., 2022; Sehl, et al., 2022). This has led to an area of open research, with ongoing research to understand the molecular mechanisms which cause this rapid age acceleration, and whether modern treatment for HIV through anti-retroviral therapy (ART) reverses this acceleration once someone becomes undetectable (i.e., they no longer have active HIV circulating).

Even with the significant accuracy of and resulting insights from epigenetic clocks in humans, one may be curious about building epigenetic clocks for non-human mammalian species, or for multiple mammalian species at once, given the evolutionary conserved nature of age-related methylation changes, and the careful construction of the mammalian array. As will be shown in this study, it has become abundantly clear that one can build epigenetic clocks for other mammalian species, for entire families of mammalian species, for humans and relatively distantly related mammalian species, and more recently, even for some non-mammalian species, although the framework for building these clocks is non-uniform out of necessity, due to the complexity of identifying shared epigenetic aging features across distantly-related species.

The first portion of this study will present a whole host of successfully built epigenetic clocks for mammalian species in general, made possible by a variety of statistical strategies, and the creation of a software pipeline for reproducibility for building new epigenetic clocks when new data is provided. The remainder of this study will shift focus towards the application of epigenetic clocks into humans, and aim to further quantify the rejuvenating (age decelerating)

effect ART has on individuals with active HIV infection once they become partially treated or completely undetectable. The research aims of this study are:

1) To expand the epigenetic clock-building pipeline with additional steps, to allow for non-human and multi-species data,
2) To demonstrate the ability to train epigenetic clocks for single mammalian species, for families of mammalian species, and for human-animal species pairs,
3) To examine the effects of both HIV and HAART (highly active ART) on epigenetic age acceleration and T cell subset composition.

In Chapter One, the background and motivation of the study have been introduced, and the research aims have been identified. In Chapter Two, the existing literature will be reviewed to explain the current framework for building epigenetic clocks, and to present the landscape of trained epigenetic clocks. In addition, the existing literature on HIV-induced age acceleration will be reviewed. In Chapter Three, the techniques for building non-human epigenetic clocks will be presented, and the myriads of next-generation epigenetic clocks will be discussed and examined. In Chapter Four, the newly-published R software package containing a pipeline for reproducibility for building novel mammalian epigenetic clocks will be presented, including an overview of features. In Chapter Five, novel research on data from the Multicenter AIDS Cohort Study (MACS), and the insights gained about the relationship between aging, HIV, ART, and T cell composition will be discussed.

CHAPTER 2: LITERATURE REVIEW

This literature review begins with a brief overview of the biological framework of epigenetic clocks, followed by the current standard statistical approach to building epigenetic clocks, and then by an overview of existing and published epigenetic clocks for mammals. To start with, is a review of the definition of what is an epigenetic clock. When interpreted as a biomarker, an epigenetic clock is any highly accurate estimator of age that is based on DNA methylation levels, although it may potentially incorporate life history traits during final computation. Moreover, an intrinsic epigenetic clock is an estimator of intrinsic age, meaning that it is unperturbed by changes in cell-type composition and independent of tissue type, which will be the case for all of the clocks presented during this literature review. An epigenetic clock can also be interpreted biologically/conceptually, and refer to a collection of innate biological mechanisms which give rise to these underlying age-related changes in DNA methylation levels (Horvath and Raj, 2018).

When presenting the problem of building an intrinsic epigenetic clock as an estimation problem, it is specified as follows:

$$F(y) = x^T\beta + \varepsilon, \qquad \beta \in \mathbb{R}^P, \mathbb{E}[\varepsilon] = 0$$

where y is Age (in years), and $x \in [0,1]^P$ is a vector of DNA methylation beta values (normalized values, such that 1 means consistently hypermethylated and 0 means consistently hypomethylated). Furthermore, $F: (-1, +\infty) \to \mathbb{R}$ is any continuously differentiable, concave, strictly increasing function, and is referred to as the "age transformation" of the epigenetic clock (although it may simply the identity function in many cases). The model could be fitted to find $\hat{\beta}$, given F, using any supervised machine learning method designed for extremely high-dimensional data. The estimated age $\hat{y}$ is referred to as "DNA methylation age" or "epigenetic age", or simply "DNAm age". The estimation residual $\hat{y} - y$ is referred to as "age acceleration", and is an important clinical measure that captures abnormal aging effects. It is important to

note here that tissue type is NOT incorporated into the model specification, which is the main reason why such an epigenetic clock is labeled as “intrinsic” (Horvath and Raj, 2018). The fact that building such an estimator is possible (as shall be seen) is an interesting discovery all its own, given the biology behind DNA methylation and cell type definition and differentiation, but for this study we will simply move past it. Much like with any statistical estimator, an epigenetic clock should not be used for extrapolation; it is only expected to be accurate over the range of ages and set of tissue types on which it was trained.

While the definition of an epigenetic clock is broad, and the corresponding estimation problem is broadly specified, there is a specific and standard methodological approach used when building intrinsic epigenetic clocks for animal species or sets of animal species. First, regarding the data being used to train the clock, it is typically generated using the Horvath mammalian array (Arneson, et al., 2022), when building clocks for non-human mammals. Second, regarding the model fitting method, it is standard (assuming normal errors) to use elastic net regularization (Friedman, et al., 2010) on $F(y) \sim x$ with an equal weighting between the L_1 and L_2, and 10-fold internal cross-validation for selecting the penalty parameter. Then, $\hat{y}$ is computed by applying the inverse function F^{-1} to the predicted values. This regression method captures common features of both lasso and ridge regularization, namely that of sparsity and shrinkage. Given that there are over 37,000 probes in the mammalian array, and that there are hundreds of thousands of probes in other arrays, the model fitting method must be sparse in order to generate a flexible and parsimonious estimator; the clock should not only be accurate, but biologically interpretable, by using a small fraction of the sites on the genome. Regarding the interpretation, the regression model selects the ‘most informative’ probes for the final clock, based purely on the ability to collectively explain a sufficient amount of the variance in the outcome, but without selecting too many probes. Therefore, single selected CpG sites cannot be interpreted as biologically meaningful; they must be interpreted as a collection.

Regarding how epigenetic clocks are assessed for accuracy, aside from considering the age range and sample size of the training data, there are two importance measures for assessing accuracy. Rather than looking at MSE, we instead use "age correlation", which is the Pearson's correlation between DNAm age and age, $r = cor(\hat{y}, y)$. As a second measurement, we calculate "median error", which is the median absolute error between DNAm age and age, $\sum|\hat{y} - y|$. These measurements are both typically assessed not from the final model, but from a cross-validated variation, using leave-one-out (LOO) estimation; this is done to maximize the sample size of the training data. Specifically, for every training sample, fit an estimator (with the exact same specifications) using all of the other training samples, and apply that estimator to the sample that was left out; we call this value "DNAm age LOO", and the resulting correlation "LOO age correlation". An age estimator is deemed to be a proper and accurate epigenetic clock if the LOO age correlation exceeds 0.8 with a significant sample size (Horvath and Raj, 2018).

Having defined intrinsic epigenetic clocks, both conceptually and statistically, and having reviewed the standard statistical framework for training epigenetic clocks, it now stands to review the literature of published intrinsic epigenetic clocks, and examine their accuracy, methodological differences, and biomedical applications. The first major intrinsic epigenetic clock to be published, the "Horvath clock", was published in 2013, and it is a highly accurate multi-tissue clock (age correlation = 0.96, and MAE = 3.6, in independent testing data set) designed to measure intrinsic age in humans (Horvath, 2013). This clock was trained on over 7,000 tissue samples from 51 different tissue and cell types, making it extremely flexible. This clock, and other clocks like it, that are designed to work on various tissue types from a single species, are often referred to as "1st-generation" clocks. The Horvath clock was a significant development, because it was a biomarker that only used DNA methylation levels at 353 sites on the genome. It should be noted that the Horvath clock used a distinctive non-linear age transformation, with logarithmic behavior at pre-juvenile ages, and linear behavior at post-

juvenile ages. The idea of using this kind of age transformation, to try to capture the non-linear nature of mammalian development and maturity, is one of the main reasons why it has been possible to build future epigenetic clocks.

The demonstration that it was possible to build a highly accurate predictor of intrinsic age using an extremely small number of CpG islands, motivated the research work to build 1st-gen clocks in other mammalian species. Significantly many other 1st-gen epigenetic clocks have since been successfully trained (and are highly accurate) for other non-human mammalian species as well (some using a large diversity of tissue types), including: rhesus macaque (pan-tissue) (Horvath, et al., 2021), naked mole rat (pan-tissue) (Horvath, et al., 2021), brown rat (pan-tissue) (Chiavellini, et al., 2022), pig (pan-tissue) (Schachtschneider, et al., 2021), cattle (blood-oocyte) (Kordowitzki, et al., 2021), sheep (blood-ear) (Sugrue, et al., 2021), common bottlenose dolphin (blood-skin) (Barratclough, et al., 2021; Robeck, et al., 2021), Indo-Pacific bottlenose dolphin (skin) (Peters, et al., 2023), vervet monkey (blood-brain-liver) (Jasinska, et al., 2022), plains zebra (blood-skin) (Larison, et al., 2021), horse (blood-liver) (Horvath, et al., 2022), opossum (ear-liver-tail) (Horvath, et al., 2022), masked shrew (fetal-liver-tail) (Cossette, et al., 2023), common marmoset (blood) (Horvath, et al., 2021), roe deer (blood) (Lemaître, et al., 2022), yellow-bellied marmot (blood) (Pinho, et al., 2022), Tasmanian devil (blood) (Horvath, et al., 2022), red kangaroo (blood) (Horvath, et al., 2022), killer whales (skin) (Parsons, et al., 2023), bowhead whales (skin) (Parsons, et al., 2023), and olive baboon (pan-tissue) (Horvath, et al., 2023).

With the development of these clocks, the next step was to see if it were possible to build more advanced epigenetic clocks, which would work in multiple species simultaneously, and potentially capture underlying aging processes that were shared across taxonomic families rather than being species-specific. It quickly became apparent that these clocks were possible to build (and still highly accurate), and this led to the emergence of 3rd-gen clocks (clocks that are accurate and designed to work for multiple species simultaneously). 3rd-gen clocks have

since been developed for many mammalian families, sub-orders, and highly heterogeneous species, including: bats (skin) (Wilkinson, et al., 2021), toothed whales (blood-skin) (Robeck, et al., 2021), elephants (blood) (Prado, et al., 2021), cats (blood) (Raj, et al., 2021), spiny mice (pan-tissue) (Horvath, et al., 2022), *Macropus* (kangaroos and wallabies) (blood) (Horvath, et al., 2022), dogs (blood) (Horvath, et al., 2022), killer-bowhead whales (skin) (Parsons, et al., 2023), pinnipeds (blood-skin) (Robeck, et al., 2023), lemurs (blood-skin) (Horvath, et al., 2023), and primates (pan-tissue) (Horvath, et al., 2023). Incredibly, it was also possible to build 3rd-gen clocks for sets of species that were NOT close relatives of each other, as many "human-animal" clocks have also been developed, including: human-rhesus macaque (pan-tissue) (Horvath, et al., 2021), human-naked mole rat (pan-tissue) (Horvath, et al., 2021), human-brown rat (pan-tissue) (Chiavellini, et al., 2022), human-cattle (blood-oocyte) (Kordowitzki, et al., 2021), human-sheep (blood-ear) (Sugrue, et al., 2021), human-vervet monkey (blood-brain-liver) (Jasinska, et al., 2022), human-horse (blood-liver) (Horvath, et al., 2022), human-common marmoset (blood) (Horvath, et al., 2021), human-elephants (blood) (Prado, et al., 2021), human-cats (blood) (Raj, et al., 2021), human-pig (pan-tissue) (Schachtschneider, et al., 2021), human-spiny mice (pan-tissue) (Horvath, et al., 2022), human-opossum (ear-liver-tail) (Horvath, et al., 2022), human-dogs (blood) (Horvath, et al., 2022), human-olive baboon (pan-tissue) (Horvath, et al., 2023), and human-clawed frog (pan-tissue) (Zoller, et al., 2023). As the development of new epigenetic clocks emerge, there are early results about successful clocks for non-mammalian species, and even human-amphibian clocks, including axolotl, clawed frog, and human-clawed frog clocks (all under editor review). The most significant development in 3rd-gen clocks is the "universal mammalian clock", which is designed to work in all mammalian species and all tissue types (and is currently under editor review) (Lu, et al., 2023). The universal clock, like most of the 3rd-gen clocks, uses an age transformation, and the age transformation chosen for each of these clocks can be quite variable. For example, the universal clock uses a more sophisticated transformation, a "log-linear" age transformation, that is similar in idea to the original Horvath

clock (but parameterized by the age of sexual maturity of a given animal). The motivation and details of these age transformations will be a primary point of discussion in Chapter 3.

The development of most of these 3rd-gen clocks has been possible due to statistical strategies for re-parameterizing the clock model, and assisting the machine learning algorithm by manually “pre-filtering” the probes before passing them in to the machine learning algorithm (the details of which shall be a primary point of discussion in Chapter 3). However, the techniques necessary to achieve these successful developments highlight a significant short-coming in the current approach to building epigenetic clocks, which is the complete reliance on the supervised machine learning algorithm to identify the most predictive probes, and do so in a way that maximizes LOO correlation, and thus largely guarantees consistent accuracy in new external data sets. Another, secondary, short-coming that is highlighted, is that the current approach does not utilize any information about how these thousands of CpG sites interact with each during transcription, and how they may or may not regulate or co-regulate gene expression. Furthermore, in some of these new 3rd-gen clocks, the best performance is achieved through data-driven pre-filtering of probes, and results in epigenetic clocks where some of the selected probes do not map to a CpG site on the specific animal’s genome. These short-comings motivate the development of a more biologically-informed modeling framework for building epigenetic clocks, so that the machine learning algorithm might be able to leverage biological associations between CpG sites on the genomes.

This literature review concludes with an overview of HIV-induced age acceleration. In recent years, significant strides have been made in the treatment of HIV, notably through the use of antiretrovirals, which have markedly extended life expectancy for those living with the virus. Despite these advancements, people living with HIV (PLWH), even those maintaining undetectable viral loads, continue to face elevated risks of aging-related diseases such as frailty, heart disease, and neurocognitive degeneration. This phenomenon has spurred interest in measuring biological aging beginning at HIV acquisition, and using metrics like the "age

acceleration residual" (AAR) from the Horvath multi-tissue clock. Furthermore, other epigenetic tools like the Grim Epigenetic Age Acceleration (GEAA) and the Phenotypic Epigenetic Age Acceleration (PEAA) have emerged, offering insights into life and health spans, and DNAmTL which measures telomere length, an important marker of cellular aging. In the context of HIV, epigenetic tools have revealed significant accelerated aging in individuals with HIV compared to age matched uninfected controls (Breen, et al., 2022; Sehl, et al., 2021). However, most studies have lacked longitudinal data from the onset of HIV infection through its progression and treatment.

CHAPTER 3: PARAMETRIC STATISTICAL METHODS FOR MULTI-SPECIES AND MULTI-PHYLUM AGE PREDICTION

The purpose of this chapter is to present the initial methodological developments of this study; specifically, expanding the epigenetic clock-building pipeline to allow for non-human and multi-species data. These developments can be categorized into two set of strategies: (1) utilizing intuitive age-transformations that account for different "speeds" of aging in different animal species, and (2) strategies for sensibly "pre-filtering" probes using species genome annotations or using data itself. This chapter will conclude with a presentation of an R package containing all of the functions for the clock-building pipeline, and containing options to allow the user to use these various strategies.

The principal challenge (a statistical challenge) of expanding the epigenetic clock-building framework to allow for multi-species data is the issue of time scales. Specifically, if one wants to build an age-predictor that operates by assigning linear weights on methylation levels at CpG sites, then the magnitude of these weights is influenced by the lifespan of the animal, such that the magnitudes will be smaller if the lifespan is larger. As a result, it becomes infeasible and statistically inappropriate to pool samples from different species, if those species have significantly different lifespans; it will result in huge variation, in order to keep the bias down. Therefore, a natural solution is to introduce the concept of "scaling the age space", so that the weights on CpG sites no longer have a uniform effect on the predicted outcome across all species. In this approach, one wants to reparametrize the outcome variable, age, so that in all species, this variable has roughly the same value for every species at every key life history trait (e.g., conception, birth, sexual maturity, adulthood, and death). By doing so, the weights on CpG sites would correspond to progression along the life course, rather than unit increments to age in years. This is the motivation behind what will be referred to as "species-parametric age transformations".

When using a species-parametric age transformation, the new statistical model for the estimation problem becomes:

$$F(y|\theta) = x^T\beta + \varepsilon, \qquad \beta \in \mathbb{R}^P, \mathbb{E}[\varepsilon] = 0$$

where θ is a vector containing one or more life history traits (which change depending on species). The model could be then fitted in the same ways to find $\hat{\beta}$, given F, and given θ for every species in the training data, using any supervised machine learning method designed for extremely high-dimensional data. In our studies on building 3rd-gen epigenetic clocks for various sets of species, including human-animal clocks, we find that one particular class/functional form of F is both effective for prediction, and intuitive to interpret, which is a "log-linear" age transformation. This type of smooth age transformation is defined as a logarithmic function for values less than some significant developmental stage (usually age at sexual maturity), and is defined as a linear function for values greater than that stage.

Within the many human-animal clock papers that I have worked on (presented in Chapter 2), there is one age transformation that I developed that we have found to be effective in all of these papers, and which I named the "LLin3" transformation (standing for "Log-Linear version 3", because it was the 3rd transformation that I developed and tested). Learning from previous iterations, I found that a transformation based on a single life-history trait was more effective in improving prediction, because it was "simpler" and less rigid. I used a transformation parameterized by Age of Sexual Maturity (A). The LLin3 transformation is $F(y|A)$, given by

$$F(y|A) = g\left(\frac{y+1.5}{A+1.5}\right) \text{ where } g(t) = \begin{cases} \log(t), & for\ t \le 1 \\ t-1, & for\ 1 \le t' \end{cases}$$

where y denotes Age. This transformation ensures continuity and smoothness at the change point y = A. Explicit, $F(y|A)$ is given by

$$F(y|A) = \begin{cases} \log\left(\frac{y+1.5}{A+1.5}\right), & for\ 0 \le y \le A \\ \frac{y-A}{A+1.5}, & for\ A \le y \end{cases}.$$

This transformation was used in all of the human-animal clocks, the primate clock, and the bat clock, and we found that it was highly necessary for ensuring accurate prediction.

This transformation later formed the basis for a more sophisticated transformation, one that is used for the universal mammalian clock (Lu, et al., 2023). This age transformation, which I will refer to as the "LLinR" transformation (or "Log-linear Relative" transformation), is also parametrized by Age of Sexual Maturity (A), but is also parametrized by Gestation Time (G), and by a species-specific scaling/change point parameter, given by m. The LLinR transformation is motivated by the functional form $F(y|A,G,m)$, given by

$$u(y|A,G) = \frac{y+G}{A+G},$$

$$F(y|A,G,m) = g\left(\frac{u(y|A,G)}{m}\right) \text{ where } g(t) = \begin{cases} \log(t), & for\ t \le 1 \\ t-1, & for\ 1 \le t' \end{cases}$$

where y denotes Age. This transformation ensures continuity and smoothness at the change point u = m. Explicit, $F(y|A,G,m)$ is given by

$$F(y|A,G,m) = \begin{cases} \log\left(\frac{y+G}{m(A+G)}\right), & for\ 0 \le y \le m(A+G)-G \\ \frac{y+G-m(A+G)}{m(A+G)}, & for\ m(A+G)-G \le y \end{cases}$$

The parameter m is a substitute for the ideal parameter value m*;

$$m^* = c_1 * \left(\frac{MaxLifespan+G}{A+G}\right),$$

where $c_1 < 1$ is a proportionality constant that would control the distribution of $F(y|A,G,m^*)$, such that $F(y|A,G,m^*)$ is approximately normal. The reason that we substitute m for m* is because we want to define our transformation *without* the use of maximum lifespan, since it is an unreliable quantity for understudied species (i.e., maximum lifespan is something that must be observed over a long period of time, and is occasionally updated and increased as new

longevity records are broken). We noticed that the ratio (Max Lifespan/Age of Sexual Maturity) is moderately constant across species, and that when examining the logarithm of m*/c_1,

$$\log(m^*/c_1) = \log(MaxLifespan + G) - \log(A + G)$$

$$= \log\left(\frac{MaxLifespan + G}{A}\right) - \log\left(\frac{A + G}{A}\right)$$

$$= \log\left(\frac{G}{A} + \frac{MaxLifespan}{A}\right) - \log\left(\frac{G}{A} + 1\right).$$

Therefore, we decided to construct an approximation of m* (for each species) based on the quantity G/A. Indeed, the $\log\left(\frac{MaxLifespan+G}{A+G}\right)$ and the $\log\left(\frac{G}{A}\right)$ have moderate correlation across all sampled mammalian species in the universal mammalian clock (r=0.5), and the specific approximation was

$$\log\left(\frac{MaxLifespan + G}{A + G}\right) \approx 2.92 + 0.38 \log\left(\frac{G}{A}\right)$$

$$\Rightarrow \frac{MaxLifespan + G}{A + G} \approx e^{2.92}\left(\frac{G}{A}\right)^{0.38}$$

$$\Rightarrow m^* \approx c_1 e^{2.92}\left(\frac{G}{A}\right)^{0.38} = c_2\left(\frac{G}{A}\right)^{0.38},$$

where $c_2 = c_1 e^{2.92}$. Therefore, we defined our substitution approximation, $\widehat{m}(A, G)$, as

$$\widehat{m}(A, G) = 5.0 * \left(\frac{G}{A}\right)^{0.38},$$

where we chose $c_2 = 5.0$ because it led to $F(y|A, G, \widehat{m}(A, G)))$ being approximately normal and mean 0 in our data, with a median of 0.0009 and a skewness of -0.02. In summary, the LLinR transformation is $F(y|A, G)$ given by

$$u(y|A, G) = \frac{y + G}{A + G},$$

$$F(y|A, G, m) = g\left(\frac{u(y|A, G)}{5 * (G/A)^{0.38}}\right) \text{ where } g(t) = \begin{cases} \log(t), & for\ t \le 1 \\ t - 1, & for\ 1 \le t \end{cases}.$$

The secondary challenge (a biological challenge) of expanding the epigenetic clock-building framework to allow for non-human array data, generated by the mammalian array or by

a generic human-specific array, is the issue of interpretability and signal. Specifically, if one wants to build age-predictors for many different mammalian species, starting from the same set of CpG probes, then there is the concern that the probes used for prediction in the final model may not all map to the genome of the given species. Although we hope that the machine learning algorithm would be good at selecting the probes that map to CpG sites on the genome of a non-human species, this is not usually the case. This is because, when running an array like this on a non-human tissue sample, the probes that do not map to any CpG sites on the genome do not necessarily return with an ambiguous signal (i.e., a beta value close to 0.5). One theory is that these probes may instead be capturing technical background noise that is important or useful for more accurate prediction, but at the cost of interpretability. Often, the machine learning algorithm alone is not biologically informed enough to find a strong signal and generate a strong prediction using only the probes that map to the animal's genome. Therefore, a natural and necessary solution is to "pre-filter" the methylation data before passing it into the algorithm. In this approach, one wants to simply reference the most up-to-date genome annotation for that animal, and cross reference it with the list of probes that are present in the methylation data, removing any probes that do not have a corresponding match. By doing so, one guarantees that the probes used for prediction in the final model will all map to the animal's genome. Using mappability to genome annotation as a pre-filtering step has been effective in all of the non-human and human-animal clocks that I have worked on that are published (see Chapter 2); it led to interpretable clocks, and it also almost always resulted in a stronger accuracy of the clocks (when using the cross-validation analyses to estimate unbiased performance).

As a final remark, more recent research work has been on building epigenetic clocks in *non-mammalian* species, while still using the mammalian array to collect DNA methylation. In this more extreme modeling scenario, it turns out that it is still possible to build moderately or highly accurate epigenetic clocks, depending on the animal; however, it becomes much more

challenging from a prediction perspective. There is a dilemma about how to choose a pre-filtering step, because while in mammalian species clocks, pre-filtering based on genome mappability is sufficient to achieve both biological interpretability and higher predictive accuracy, in these non-mammalian species clocks, there is sometimes a trade-off that must be evaluated between statistical accuracy and biological interpretability. While it is not surprising that on average, a small fraction (10%-15%) of the probes on the mammalian array map to these non-mammalian species of interest (like axolotls, which are amphibians, and clawed frogs, which are reptiles) (Zoller, et al., 2023), it is somewhat surprising that one can still construct accurate epigenetic clocks if one keeps probes that do not map to the animal's genome, and instead pre-filter probes using data-driven methods. Currently, the most successful and promising strategy, which I have called the "middle filter", is to simply remove probes that have a mean methylation (across samples across all ages) "near" to 0.5. Just as what has been discussed above, a significant fraction of the probes that remain do still map to the genome of the animal, but a significant fraction of the probes that remain also do not map, and sometimes removing these probes hurts substantially hurts prediction.

CHAPTER 4: SOFTWARE FOR DNA METHYLATION-BASED EPIGENETIC CLOCKS IN MAMMALS

The purpose of this chapter is to present the R package "MammalMethylClock" and the accompanying statistical techniques pivotal for the development, assessment, and application of these epigenetic clocks. I will describe the steps to use this software package, including how to easily apply the aforementioned published non-human and multi-species clocks (notably including the Universal Pan-Mammalian clocks (Haghani, et al., 2023)).

DISCUSSION

Overview of the software package

The MammalMethylClock R package serves as a comprehensive suite tailored to the demands of constructing, assessing, and deploying new epigenetic clocks. An overview of the typical analysis sequence can be found in **Figure 4-1**. To begin, the **Building Clocks** module empowers users to fashion new epigenetic clocks. Our package also offers an **Applying Clocks** module. If pertinent, an inverse age transformation should also be specified. Recognizing the criticality of rigorous model evaluation, the package integrates two distinct cross-validation schemes: **Leave-One-Out (LOO)** and **Leave-One-Species-Out (LOSO)**. For researchers keen on examining the relationship between DNA methylation and age, the **Epigenome-Wide Association Studies (EWAS)** feature has been incorporated, including **Meta EWAS** utility.

Epigenetic clocks as penalized regression models

Formally, epigenetic clocks are conceptualized as regression models, given mathematically as:

$$F(y) = x^T\beta + \varepsilon, \qquad \beta \in \mathbb{R}^P, \mathbb{E}[\varepsilon] = 0$$

Here, y is Age (in years), and $x \in [0,1]^P$ signifies a vector of DNA methylation beta values from *P* CpG probes. These beta values render a spectrum wherein '1' epitomizes consistent hypermethylation and '0' characterizes consistent hypomethylation. The “age transformation” function F is integral to the epigenetic clock, although in numerous applications, it is simply the identity function.

Central to the construction of most epigenetic clocks is a sparse, penalized regression method such as elastic net regression or LASSO (Tibshirani, 1996; Zou and Hastie, 2005). The clock building tools employ R functions from the "glmnet" package (Friedman, et al., 2010).

Overview of DNA methylation data handling

Our membership of the Mammalian Methylation Consortium used the HorvathMammalMethylChip40 Infinium array platform, also known as the mammalian methylation array, which profiles oligonucleotide probes that pinpoint sites conserved across a vast majority of mammalian species (Arneson, et al., 2022). Every published clock integrated into our software package owes its genesis to this mammalian methylation array, most notably the Universal Pan-Mammalian clocks (Lu, et al., 2023).

Raw array data (idat files resulting from the iScan machine) require normalization to arrive at DNA methylation values (beta values). Our membership of the Mammalian Methylation Consortium used SeSaMe normalization, a trusted method for all clocks embedded in this software, but other approaches can be adopted (Arneson, et al., 2022; Zhou, et al., 2018).

CpG probe pre-filtering

For a given array, the beta values should typically follow a bimodal distribution, with most values clustering near 0 or 1. However, in many species, one often observes a trimodal distribution, characterized by an additional intermediate peak around 0.5. This mid-point peak,

which represents a technical artifact, signifies the presence of CpGs on the array that fail to align with the underlying genome.

Annotation Mappability Filtering: The mammalian methylation array's probes aren't universally applicable to every species. For example, several thousand CpGs aren't applicable for mice, but are to humans (Arneson, et al., 2022). I advise data analysts to exclude CpG probes from their analysis when the associated oligonucleotide sequence isn't found in the target genome. CpG genome annotations for hundreds of species (including non-mammalian) are available on the Mammalian Methylation Consortium's GitHub page (Haghani, et al., 2023; Lu, et al., 2023) at https://github.com/shorvath/MammalianMethylationConsortium/tree/v2.0.0.

Middle Filtering: It can be advisable to exclude CpG probes with an average DNA methylation value approximating 0.5 across all (training) samples. This simple procedure can be accessed within the “selectProbes.middleFilter()” function.

Sesame Detection P-value Filtering: It can be advisable to discard CpG probes with detection p-values exceeding 0.05. This filtering method is done via supplemental data generated during the SeSaMe normalization (Zhou, et al., 2018).

Unsupervised hierarchical clustering

To filter out technical outliers and identify inherent batch effects in the data, unsupervised hierarchical clustering is applied to the normalized DNA methylation data via the standard R function, “hclust()” (Murtagh, 1985). Each branch of the hierarchical clustering dendrogram represents a cluster (a group of closely correlated DNA samples). These clusters can be identified from various branch cutting techniques, using “cutreeStatic()” from the

WGCNA package (Langfelder and Horvath, 2008). Hierarchical clustering effectively highlights outlier samples based on a large height value (y-axis) in the cluster tree.

To evaluate the pairwise resemblance in DNA methylation array readings, I employ the inter-array Pearson correlation coefficient across CpGs. Generally, anticipated pairwise correlations of the same tissue exceed 0.9. For discerning intergroup disparities, I usually choose "average linkage", but other intergroup dissimilarities can be used as well.

Transforming age in the clock model

Age transformations play a pivotal role in bolstering the predictive accuracy of epigenetic clocks in external datasets. This software suite offers predefined transformation functions (along with their inverse functions) tailored to the specific pre-constructed clock in use.

Existing clocks included in the package

The MammalMethylClock package is equipped with built-in clocks designed for a diverse range of mammalian species and groups, spanning from rats, primates, and cetaceans to marsupials. To ensure accurate results, data should be obtained from the mammalian methylation array, and the “predictAge()” function should be employed. A comprehensive list of built-in clocks and other pertinent information can be found in the “getClockDatabase()” function, and also on the software's GitHub repository (https://github.com/jazoller96/mammalian-methyl-clocks/tree/main).

EWAS

The Epigenome-Wide Association Study (EWAS) module is used to identify individual CpGs and genomic locations that highly relate to age. During an EWAS, correlations between every CpG probe and a specified outcome variable (e.g., age) are ascertained across all (training) samples.

The results of an EWAS act as precursors for downstream analyses including eFORGE (Breeze, et al., 2019), Genomic Regions Enrichment of Annotations Tool (GREAT) analysis (McLean, et al., 2010), and universal chromatin state analysis (Vu and Ernst, 2023). These functionalities are not currently embedded in our software package.

METHODS

Training, Validation, and Downstream Analysis Pipelines

The MammalMethylClock package delivers a comprehensive toolkit for researchers venturing into the domain of epigenetic clock studies. This package is designed to incorporate a suite of functionalities tailored for the development, assessment, and utilization of epigenetic clocks. Among its main features is the ability to create new epigenetic clocks/biomarkers using training datasets provided by users. These datasets should include age values coupled with normalized DNA methylation data. It's worth noting that the output of this feature might be influenced by inherent algorithmic randomness. To ensure reproducibility, users are strongly advised to use the “set.seed()” function in their R scripts before executing these functions.

Another primary feature of this package is the capacity to predict DNAm age for specific samples. This feature only requires normalized DNA methylation data, coupled with any pre-built epigenetic clock, which is itself characterized by a coefficients table and, if relevant, an inverse age transformation.

For model evaluation, the package offers the Leave-One-Out (LOO) cross-validation technique, in which each sample is systematically excluded to assess the accuracy and robustness of the model. As with the clock construction, outputs from this analysis might exhibit some randomness. Thus, the insertion of the “set.seed()” function is strongly advised. Furthermore, the package includes a variant of this approach, called Leave-One-Species-Out (LOSO) cross-validation. Here, instead of single samples, all data from a distinct species are left

out during the model training step. Consistency and reproducibility in results can be maintained with the “set.seed()” function.

Beyond these features, the MammalMethylClock package is equipped to conduct Epigenome-Wide Association Studies (EWAS). This allows researchers to probe the relationship between DNA methylation patterns and specific phenotypes, especially age. Complementing this, the package also supports a simple Meta EWAS feature, which refines the study by taking stratifying variables such as tissue type or species into account.

Implementing Cross-validation Techniques

To ascertain the accuracy of a clock, I advocate for users delve into cross-validation statistics. The software package boasts two distinct methodologies which hinge on the nature of the data at hand: LOO (Leave-One-Out) and LOSO (Leave-One-Species-Out). Each approach partitions the dataset in a unique manner for validation purposes.

In the LOO analysis, employed via the “saveLOOEstimation()” function, the process unfolds for every sample within the dataset as follows: Initially, a provisional replica of the datasheet and associated DNA methylation data are constructed, deliberately excluding the chosen sample from both. Then, a clock model is fit utilizing this provisional dataset. The specifications for this step are parallel to those presented in the “saveBuildClock()” function. Following this, the coefficient table of this LOO clock is archived into a list of matrices. Concluding the process, this provisional LOO clock is applied to the omitted sample, and the forecasted value is stored in the output.

For computational efficiency, it's crucial to understand that this function avoids conducting an internal cross-validation at each iteration to pinpoint a potentially varied optimal value for the penalty parameter lambda for every LOO clock. Instead, recognizing that these optimal lambda values typically converge closely, the function designates a singular optimal

lambda value based on the starting dataset. This identical lambda value is then harnessed to configure every LOO clock.

Species characteristics from AnAge

Characteristics of species, including maximum lifespan, age of sexual maturity, and gestational period, which serve as parameters for certain clocks within this package, were sourced from a revised edition of the Animal Aging and Longevity Database (AnAge, https://genomics.senescence.info/species/index.html) (de Magalhães, et al., 2007). I employ the same enhanced version of AnAge that was presented in a paper I contributed to by Lu, et al (Lu, et al., 2023).

CONCLUSION

The MammalMethylClock R package offers specialized statistical and bioinformatics tools dedicated to developing epigenetic clocks for single and multiple species. The R functions can be used to both devise new epigenetic clocks and deploy pre-existing ones. While our primary focus has been on mammalian species, we've recently extended the software's application to amphibians, like *Xenopus* frogs (Zoller, et al., 2023).

While the MammalMethylClock package is optimized for data from the mammalian methylation array, most of package’s functions apply to data from other measurement platforms as well.

CHAPTER 5: LONGITUDINAL ANALYSIS OF EPIGENETICS FROM SEROCONVERSION UNTIL FIRST ART INITIATION

The purpose of this chapter is to present and results and discussion from a longitudinal study of already examined subjects in the MACS (Multicenter AIDS Cohort Study), which spans 4 specific time points, and which examines 60 PLWH and 60 age-matched, hepatitis C-matched and time-interval matched individuals living without HIV as controls. The study captures various important timepoints: 2-3 years prior to seroconversion (Visit A), the period closest to seroconversion (Visit B), the phase of living with HIV before ART initiation (Visit C) and after 1-3 years of being on ART (Visit D). All participants had complete observations on all 5 epigenetic clocks examined. Comprehensive data on flow cytometry measures was also utilized, including naive CD4/CD8 counts, senescent CD4/CD8, activated CD4/CD8, as well as total CD4/CD8, in both absolute counts and as percentages.

It has previously been shown that acquisition of HIV impacts the immune system and can accelerate aging as measured with epigenetic clocks. Antiretroviral therapy has been shown to slow these aging effects and can partially restore the immune function in patients who achieve durable virologic suppression. By tracking epigenetic changes from the time of seroconversion to ART initiation, this study seeks to unravel the specific aging patterns associated with HIV, potentially guiding the development of interventions to enhance the healthspan of PLWH and also lend insights into aging. In this chapter, I analyze epigenetic aging using five different aging clocks across 4 four time points, running from pre-seroconversion to the initiation of ART in 60 seroconverters and 60 age-matched controls. I incorporate multivariate adaptive regression splines, a machine learning model to aid in model selection and mixed models to explain the data. My findings indicate that, across most aging clocks, HIV infection significantly and prematurely alters the epigenetic aging process and this is only partially restored when the viral replication is controlled. This suggests that HIV itself

contributes to the earlier onset of clinical aging through its profound effects on the immune system.

RESULTS

Demographics

Participants from the Multicenter AIDS Cohort Study (MACS) who were included in the current sub-study of initial HIV infection were predominantly white, college-educated men who have sex with men (Table 5-1), similar to the overall MACS demographics at the initiation of the study in the 1980s (Kaslow, et al., 1987). The proportion of non-white men among the persistently HIV-seronegative (SN) participants (13/60) compared to the participants who became HIV-infected and seroconverted was not significantly different (SC; 7/60, p=0.14). Mean age in SC and SN at Visit B (post-HIV infection or equivalent visit) was 39.2 and 38.5 years, at Visit C was 44.7 and 44.9, and at Visit D was 47.4 and 47.7, respectively, reflecting the matching criteria (range 21-75 years across all participants and visits in this study). In this substudy neither patients nor controls had Hepatitis C infection. The SC and SN groups also had similarly low rates of active Hepatitis B Virus (HBV) infection at all four visits (1-2%). SC trended towards greater cumulative pack-years of smoking compared to SN at Visit D, but was not significant (p= 0.22).

By design, the mean time intervals between peripheral blood mononuclear cell (PBMC) samples at Visits A to B, A to C, and A to D were very similar in SC (3.0 years, 9.3 years, 12.1 years), and SN (2.8 years, 8.3 years, 11.0 years). Among SC, the mean time interval between estimated date of HIV infection and the post-HIV infection PBMC sample at Visit B was 2.3 years (sd 0.4 years). In addition, the mean time interval between estimated date of HAART initiation and the post-HAART initiation PBMC sample at Visit D was 2.1 years (sd 0.4 years). The calendar dates of estimated HIV infection ranged from 1985-2004, with 93% of the infection dates occurring prior to 1995. All SC post-HIV infection samples (i.e., Visit B samples) were

prior to the initiation of highly active antiretroviral therapy (HAART) (Castillo-Mancilla, et al., 2016). Mean plasma HIV viral load (VL) in SC at or immediately preceding Visit B was 29,994 copies/mL (sd 40,301 copies/mL), ranging from a single individual with <50 copies/mL up to 192,000 copies/mL. For Visit C, the mean was 58,893 copies/mL (sd 119,006 copies/mL), ranging from three individuals with <50 copies/mL up to 815,000 copies/mL. Mean plasma HIV VL in SC at or immediately preceding Visit D (and after HAART initiation) was 5,884 copies/mL (sd 15,569 copies/mL), ranging from 68% of individuals with <50 copies/mL up to 81,000 copies/mL.

Multiple epigenetic age acceleration measures differ significantly after initial HIV infection

At visit A, the pre seroconversion visit, all patients were (by design) HIV-seronegative. When comparing the SC and matched SN at Visit A, there were no statistically-significant differences between the two groups in Age Acceleration Residual (AAR), Extrinsic Epigenetic Age Acceleration (EEAA), Phenotypic Epigenetic Age Acceleration (PEAA), and Age-adjusted DNA methylation-based estimates of telomere length (aaDNAmTL) (Figure 5-1a-c,1e). The Grim Epigenetic Age Acceleration (GEAA) demonstrated a small difference bordering on statistical significance (Figure 5-1d), with the SC group being slightly epigenetically older with respect to Grim Age than the SN group.

At Visits B and/or C, following initial HIV infection in the SC, SC showed dramatic differences in some epigenetic age metrics (greater EEAA, PEAA) and estimated telomere length (shorter aaDNAmTL) that were significant, and moderately significant differences in AAR, compared to matched HIV-uninfected SN, as well as continuing to be slightly more accelerated in GEAA (Figure 5-1a-e).

Following initial ART initiation (Visit D), SC showed deceleration in epigenetic aging however there remained significant differences in some epigenetic age metrics (greater AAR, PEAA) and estimated telomere length (shorter aaDNAmTL) compared to matched HIV-

uninfected SN, as well as continuing to be slightly more accelerated in GEAA (Figure 5-1a, 5-1c-e). Average EEAA at Visit D was still greater in SC compared to SN, but the difference was not statistically significant (Figure 5-1b).

Out of concerns about the significant variation between time points across subjects (namely, Visits B and C), we created models using continuous time (time since Visit A) rather than design visit (visit letter). However, the results remain largely unchanged, as the changes in epigenetic measures follow similar trends (Figure 5-2a-e).

Initial HIV infection remains associated with older EEAA, and PEAA ages, and shorter estimated aaDNAmTL, even after taking demographic factors and T cell type composition into account

Mixed model analyses were conducted on all participants across all visits, accounting for demographic characteristics. This model, consistently applied across all 5 epigenetic measures, assessed the potential influences of demographic factors, time, HIV status, time*HIV status, and ART usage on the results of each epigenetic measure (Table 5-2, Table S2). The analyses showed that seroconversion to HIV at Visit B significantly affected epigenetic aging as measured by AAR, PEAA, GEAA and aaDNAmTL (all $p< 0.005$). Conversely, factors such as race (white vs. not white), body mass index (BMI), or smoking did not generally impact these measures (all $p> 0.10$), except for a highly significant contribution to GEAA made by smoking ($p< 0.001$), and a mildly significant contribution to GEAA made by race ($p= 0.013$).

The analysis also highlighted that the Study visit (A, B, C, or D) alone was a strong determinant of the chronologic age-adjusted epigenetic values over time for all five epigenetic measures (all $p< 0.025$, Table 5-2). However, when considering the interaction of the study visit and HIV serostatus group (i.e., SC or SN), which reflects the transition from HIV-uninfected to HIV-infected and the subsequent initiation of HAART in the SC group compared to the persistently HIV uninfected SN group, it is apparent that the initial HIV acquisition was the predominant factor affecting the epigenetic values seen at the four visits for AAR, EEAA, PEAA,

and aaDNAmTL. Although GEAA showed significant variation with respect to study visit and HIV serostatus group, it was not linked with most other covariates, including the study visit∗HIV serostatus group interaction (all p≥ 0.05). As expected, smoking history was strongly associated with GEAA, due to the inclusion of smoking-related biomarkers into this specific epigenetic clock (Table S1).

As noted in a previous study (Breen, et al., 2022), it was necessary to develop another set of mixed models that could account for potential differences in the cell composition of each sample. This study used the already collected flow cytometry data that was available from this particular subset. A description of these measures is given in for participants at all visits (Table 5-3, Table 5-4). As expected, there were no differences in any of the mean absolute T cell counts or mean percentages of T cell subsets between the SC and SN group at visit A (pre-seroconversion). Significant differences between the groups were found (for both absolute counts and percentages) at visits B and C in 6 out of the 8 T cell subsets (namely, CD4 and CD8, naïve CD4 and CD8, senescent CD8, and activated CD8) (Table 5-3, Table 5-4), with the exception of senescent CD8 T cells at Visit B. Furthermore, there were still significant differences (for both absolute counts and percentages) at visit D (post ART visit) in 5 out of those 6 T cell subsets (namely, CD4 and CD8, naïve CD4, senescent CD8, and activated CD8) (Table 5-3, Table 5-4). Correlations between absolute counts of different T cell subsets, and between each T cell subset and the five epigenetic measures, are shown in Tables S3A-E. Similarly, correlations between percentages of T cell subsets within live lymphocytes are shown in Tables S4A-E.

Utilizing various combinations of T cell subsets as described in the Methods, a consensus model was identified for this mixed model, showing the best fit across all five of the epigenetic measures. This model consisted of 5 T cell subsets, as the following covariates: natural log-transformed T cell percentages of live lymphocytes for naïve CD4, naïve CD8, senescent CD8, and activated CD8, and natural log-transformed absolute T cell count for total

CD8. This model analyzed the possible contributions of changes in T cell numbers over time, as well as changes in HIV infection and ART utilization status, to the observed values for each epigenetic measure (Table 5-5, Table S5). Significant associations were observed between T cell subset percentages and one or more of the epigenetic measures; however, the absolute total CD8 T cell count was not significantly associated with any epigenetic measures, although aaDNAmTL was marginally significant (p= 0.084). Notably, when these five T cell subsets were considered, the three of the four epigenetic measures that showed significant relationships in our initial mixed model with demographics (namely, EEAA, PEAA, and aaDNAmTL), still exhibited a statistically significant relationship to the interaction of study visit*HIV serostatus group; in addition, AAR was still marginally significant (p=0.09).

Efficacy of HAART in achieving undetectable status in short-term does not associate with complete reversal of HIV-induced age acceleration

Given the persistent significant associations between four of the epigenetic measures and the study visit*HIV serostatus group interaction, representing the HIV infection event at Visit B in the SC group, a natural question would be to ask whether the effectiveness of ART in controlling viral replication is associated with the age acceleration induced by HIV infection. PLWH were stratified by whether or not they were undetectable (< 50 copies/ml) at Visit D.

To assess if being undetectable at Visit D was associated with reversing the age accelerating effects of HIV, regression analyses were performed. Figure 5-3 illustrates the stratified regression by Visit Number. As expected, PLWH who had detectable viremia had higher age acceleration versus their age-matched controls. However, the age acceleration was not significantly different from PLWH who were undetectable, although the trend was that people who had undetectable levels of HIV at Visit D showed greater age rejuvenation versus those with uncontrolled viremia. Table 5-6 lists the parameters of the model.

DISCUSSION

While other longitudinal studies have previously explored accelerated aging in HIV seroconverters (Breen, et al., 2022; Leung, et al., 2017; Sehl, et al., 2021), and in perinatally HIV-infected youth years after HIV infection (Shiau, et al., 2021), and cross-sectional studies suggest that persons living with HIV (PLWH) may be aging at a faster rate (Gross, et al., 2016; Horvath and Levine, 2015; Rickabaugh, et al., 2015), the study described here stands out as one of the largest studies of seroconverters with four well-described longitudinal time points. This study follows individuals from before they contract HIV through initiation of ART, documenting epigenetic changes that align with accelerated biologic aging. We utilized five well-validated epigenetic measures based on methylation patterns of genomic DNA, each estimating biological age acceleration relative to chronologic age alongside a DNA methylation-based proxy for telomere length, which typically shortens with age and cell division.

Over three time periods during which age-matched HIV-uninfected men showed stable changes in epigenetic measures of aging, men who became infected with HIV at Visit B showed highly significant age acceleration in three out of four of the epigenetic clocks (AAR, EEAA, and PEAA) as well as accelerated estimated telomere shortening (Figure 5-1, Figure 5-2), and this association became stronger by Visit C. In mixed models that include demographic characteristics, age acceleration in specific epigenetic measures (AAR, EEAA, PEAA, and aaDNAmTL) continues to show a strong and significant association with initial HIV infection (see Table 5-2). Furthermore, when accounting for T-cell counts and percentages (referenced in Table 5-5), these associations persist, though the link with AAR becomes only marginally significant. This suggests that changes in T-cell metrics during the initial phase of HIV infection have a notable impact on epigenetic aging. The results suggest that there is an early and substantial impact of HIV infection on the epigenetic aging process, which also continues if viremia remains uncontrolled, linked to changes in the T-cell subtypes and overall immune system dynamics. The results seen in the different epigenetic measures are not just a reflection

of the correlations observed among them. Each clock was independently developed to assess distinct aspects of aging. Therefore, it is anticipated that multiple epigenetic indicators would point to age acceleration as a consequence of HIV infection.

Notably, PEAA indicated significant age acceleration associated with initial HIV infection, with an expected increase in biological age acceleration of 5.39 years in SC relative to SN, when comparing Visits B vs. A, after adjusting for chronologic age (Figure 5-1). PEAA is a recently-developed epigenetic clock that predicts lifespan by measuring mortality from 513 CpGs (Levine, et al., 2018). In populations not infected with HIV, a one-year increase in PEAA age acceleration is associated with a 4.5% increase in all-cause mortality risk (Levine, et al., 2018). The significant HIV-associated acceleration that was priorly stated suggests that being infected with HIV and living with uncontrolled HIV for approximately three years or less is already associated with a relative increased risk of about 25% for a shortened lifespan. Furthermore, AAR, EEAA, and PEAA indicated significant age acceleration associated with living with uncontrolled HIV over a longer period of time, with an expected increase in biological age acceleration of 5.72 years, 5.40 years, and 10.78 years, respectively, in SC relative to SN, when comparing Visits C vs. A, after adjusting for chronologic age (Figure 5-1).

On the other hand, initial HAART utilization was also associated with some reversal of epigenetic outcomes. When comparing Visits D vs. C, after adjusting for chronologic age, EEAA and PEAA signs were significantly reversed with ART initiation, with an expected decrease in biological age acceleration of 2.68 years and 4.45 years, respectively, in SC relative to SN (Figure 5-1). This is somewhat remarkable of an effect size from only utilizing HAART for approximately three years or less, but it is not a complete reversal of age acceleration associated with long-term uncontrolled HIV infection. To examine the cumulative effect of both HIV infection and initial HAART utilization, we compared Visits D vs. A, and found that some age acceleration remained in almost every epigenetic clock. After adjusting for chronologic age, AAR, EEAA, and PEAA still indicated significant age acceleration, with an expected overall

increase in biological age acceleration of 4.68 years, 2.73 years, and 6.24 years, respectively, in SC relative to SN (Figure 5-1). Due to the limitations of this study design, it is unknown whether this “rejuvenating” effect of ART would continue to the point of complete age acceleration reversal, but the magnitude of ART-associated age deceleration was not significantly associated with the efficacy of ART in completely controlling HIV VL to undetectable levels (< 50 copies.ml) (Figure 5-3). Specifically, cumulative age acceleration was not significantly different in PLWH who were detectable vs. undetectable at Visit D, although the trend was that people who had undetectable levels of HIV at Visit D showed greater age rejuvenation versus those with uncontrolled viremia.

The age-adjusted DNAm-based estimate of telomere length (aaDNAmTL) (Lu, et al., 2019), characterizes accelerated biological aging in the opposite direction of the epigenetic clocks, as telomeres become shorter with repeated cellular divisions, yielding a lower value and a negative direction of change over time. Similar to acceleration in the PEAA epigenetic clocks, aaDNAmTL showed accelerated telomere shortening over the course of initial HIV infection in SC participants relative to SN participants, comparing Visits B vs. A (Figure 5-1). Moreover, the trend continued when comparing Visits C vs. A, consistent with the accelerations observed in AAR, EEAA, and PEAA. Furthermore, aaDNAmTL showed decelerated, rejuvenated telomere shortening over the course of initial HAART initiation, comparing Visits D vs. C, again consistent with the decelerations observed in EEAA and PEAA. All of these aforementioned observations were validated in continuous time model variants, in which time was modeled as the number of years since Visit A, rather than according to study design Visit letters (Figure 5-2).

The Grim Clock (GEAA) (Lu, et al., 2019), which focuses on predicting mortality and lifespan, showed no differences over time regardless of changes in HIV infection status, or ART initiation status. At visit A, there was a trend towards a difference between participants who would go on to acquire HIV versus those that did not. GEAA continued to demonstrate a marginal difference throughout time. This suggests that the Grim Clock includes factors that

predict mortality but are less affected by changes detected by other epigenetic measures when an individual first acquires HIV or starts ART.

It is important to highlight that in this retrospective study, which examines the effects of acquisition of HIV on epigenetic markers and following the initiation of ART. These initial findings at the point of HIV acquisition highlight the necessity for further research to explore whether these epigenetic markers can forecast the development of health complications, mortality and frailty. The findings that age acceleration begins shortly after HIV acquisition provides yet another compelling reason to aim for the rapid suppression of viremia in PLWH as soon as possible after infection. In this study, we confirm the previous findings that ART can partially reverse the age-acceleration effects caused by HIV acquisition.

The MACS/MWCCS is a comprehensive longitudinal cohort study that began in 1984 and continues to the present. A key strength of this cohort is its inclusion of both HIV-infected and uninfected participants from the same at-risk demographic, along with the collection of an extensive repository of biological samples every six months. Combined with a detailed demographic and clinical dataset, this creates a valuable resource for investigating a spectrum of questions relating to the basic science, clinical science, and epidemiology of HIV infection in the U.S., with a focus on comorbidities among men and women living with HIV.

This study assessed biological aging in over sixty individuals from the onset of HIV infection, alongside a matched cohort who remained HIV-negative. Results from this study highlights the need to understand the mechanisms behind the rapid age acceleration aging seen upon HIV acquisition as well as the mechanisms behind the less than total recovery in the rate of aging once the viral replication is controlled.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects

Participants for the study reported here represent a subsample of subjects from previously reported studies and selected from among participants of the Multicenter AIDS Cohort Study (MACS), now part of the MACS/WIHS Combined Cohort Study (MWCCS). The MACS is a continuing longitudinal study that investigates the progression and treatment of Human Immunodeficiency Virus (HIV) infection in men who have sex with men (Kaslow, et al., 1987). For this specific substudy, participants were selected from MACS participants who were part of a larger biomarker study, previously detailed elsewhere (Breen, et al., 2022; Sehl, et al., 2021; Wada, et al., 2015). The MWCCS adheres to all applicable ethical guidelines, including securing informed consent for research from all participants. The MACS/MWCCS substudy was granted exempt status. by the University of California, Los Angeles Medical Institutional Review Board IRB#15001179.

A total of 60 men who had experienced documented initial HIV infection and seroconversion (SC) after joining the MACS were selected, alongside 60 age and hepatitis C matched men who persistently tested HIV-negative (SN) who had data on all for four visits. The selection criteria are detailed in the section below. Clinical and demographic information is provided in Table 5-1. Figure 5-4 illustrates the overall study design.

For this re-examination of existing data, there were only 60 matched pairs that had complete data for all four visits (Visits A-D), thus reducing our effective sample size to 120.

METHOD DETAILS

Participant Selection and Samples

Viably-frozen peripheral blood mononuclear cells (PBMC) were obtained from the national repository of the MACS/MWCCS. MACS study visits typically occur at 6-month intervals, clinical and questionnaire data are collected, and peripheral blood samples are processed and frozen.

SC participants were selected who had PBMC samples available in the repository from two time periods: (1) up to 1.5 years prior to the first HIV-seropositive study visit (pre-HIV infection, Visit A), and (2) up to 2.5 years after the first seropositive visit (post-HIV infection, Visit B). The pre-HIV PBMC sample was required to be from a visit that was both HIV antibody seronegative and with undetectable plasma HIV RNA. All post-HIV infection PBMC samples were required to be before initiation of HAART; if multiple PBMC samples were available post-HIV infection, the visit closest to 3 years after the pre-HIV visit was selected.

Matched persistently HIV seronegative (SN) controls were then selected among MACS participants for each SC. SN were selected matched by age (± 2 years) and Hepatitis C Virus (HCV) status (HCV RNA positive/negative) at all visits, as well as by availability of PBMC at two visits (Visits A and B) with comparable time interval between visits (± 0.75 years). 60 matched SN controls were identified, but a PBMC sample matched on age and HCV status within a comparable time interval was not available from one control at Visit A (equivalent to pre-HIV infection in the matched SC).

HIV serostatus, plasma HIV viral load (VL), Hepatitis B Virus (HBV) status, and other demographic and clinical data were available from the MWCCS database, and are summarized in Table 5-1 and Figure 5-4 above. Where data are missing from any parameter, the exact n is shown in Table 5-1. An estimated date of HIV infection for each SC was calculated utilizing HIV serostatus (HIV antibody and Western Blot) and HIV VL data from all MACS study visits. Date of HIV infection was estimated as the midpoint between the last MACS study visit at which the participant was HIV seronegative and HIV VL undetectable (if VL data were available) and the first MACS study visit with either HIV-positive serostatus or detectable HIV VL, whichever came first. For 10 SC for whom VL data were missing at Visit B, the VL from the MACS visit immediately prior was used (0.3-0.5 years prior, approximately 3-6 months). For post-HIV infection visits with undetectable VL, a value equal to the lower limit of detection of the VL assay was assigned; 4 SC had VL <400 copies/mL (Roche Amplicor 2nd generation assay, Roche

Molecular Systems, Branchburg, NJ, USA), and 1 had <50 copies/mL (ultra-sensitive Roche Amplicor assay). HBV status at each visit was categorized as positive (HBV surface antigen [HBsAg] positive) or negative (HBsAg negative), and smoking history was evaluated by cumulative pack years reported.

DNA Methylation Arrays

Methylation status at more than 850,000 potential methylation sites (CpGs) were measured using the Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA), by the UCLA Neuroscience Genomics Core (https://www.semel.ucla.edu/ungc). DNA methylation levels (beta values) were determined by calculating the ratio of intensities between methylated (signal A) and unmethylated (signal B) sites as previously described (Sehl, et al., 2020).

QUANTIFICATION AND STATISTICAL ANALYSES

Epigenetic Age Acceleration Measures

Five measures of epigenetic age acceleration were estimated for each of the 120 PBMC samples in total, representing samples from all participants at all visits, using the online epigenetic clock software (http://dnamage.genetics.ucla.edu). Each of these DNA methylation-based estimates was calculated using methylation beta values obtained from the Infinium MethylationEPIC BeadChip, on all samples at the same time without linkage to HIV serostatus group. There were no adjustments for multiple comparisons, as each epigenetic measure was developed taking this into account. Features of each clock examined are provided in Table S1. Briefly, Age Acceleration Residual (AAR) is based on the DNAm age estimated from 353 CpGs of Horvath's original epigenetic clock (Horvath, 2013), which is then regressed on chronologic age. AAR captures epigenetic age acceleration (i.e., older epigenetic or biologic age than chronologic age), is valid for a wide range of tissue types, and is known to be accelerated in disease states. Extrinsic epigenetic age acceleration (EEAA) is based on 71 CpGs of Hannum

(Hannum, et al., 2013), and was constructed to be positively correlated with senescent T lymphocytes and negatively correlated with naive T lymphocytes (Chen, et al., 2016; Hannum, et al., 2013). This measure captures both intrinsic methylation changes and extrinsic blood cell composition changes. Second generation clocks, including Phenotypic Age and Grim Age, were examined as they are much stronger predictors of mortality. Phenotypic Epigenetic Age Acceleration (PEAA), based on 513 CpGs, was developed by regressing a phenotypic measure of mortality risk on CpGs (Levine, et al., 2018). Grim Epigenetic Age Acceleration (GEAA), based on 1030 CpGs, was developed by regressing time-to-death on DNAm-based surrogate biomarkers of smoking pack-years and a selection of plasma proteins previously associated with mortality or morbidity (Lu, et al., 2019). Finally, a DNA methylation-based estimator of Telomere Length adjusted for chronologic age (aaDNAmTL) was examined in our analyses to evaluate whether HIV infection causes accelerated shortening of telomeres with increased age and/or rate of cellular replication (Lu, et al., 2019).

In this study, there are three time points occurring post-seroconversion to examine the difference between HIV-infected and uninfected, so it is not possible to determine exactly when and over what period of time the observed elevation in epigenetic age occurs. However, "epigenetic age acceleration" measures are compared, defined as the residual that results from regression of epigenetic age on chronologic age, at all visits for the SC and SN groups. Because these measures are age-adjusted, they are technically termed "accelerations" in the epigenetic clock literature, rather than elevations or advancements, even though the time course and rate of acceleration is unknown.

Absolute Counts of T Cell Subsets

Absolute total CD4 and total CD8 T cell counts (cells/mm^3) for each sample were available from the MACS/MWCCS database, which had been determined by standardized

protocols on the day each blood sample was originally obtained (Giorgi, et al., 1990; Hultin, et al., 2007). Percentages of naïve, activated, and senescent T cells among CD4 or CD8 T cells in thawed viable PBMC were utilized in combination with absolute total CD4 and CD8 T cell counts to calculate absolute cell counts for naïve, activated, and senescent CD4 and CD8 T cells in each sample.

Summary analyses were performed on the absolute counts of T cell subsets (Table 5-3) and percentages of T cell subsets within live lymphocytes (Table 5-4), comparing SC vs SN at Visits A through D.

Statistical analyses

No data were excluded from the analyses. Simple linear mixed models were used to compare HIV serostatus groups (SC vs. SN) on each age-adjusted epigenetic measure at visit A (all participants HIV-uninfected), visits B and C (post-HIV infection in SC, time interval-matched uninfected SN), and visit D (post-HAART initiation in SC, time interval-matched uninfected SN). Fitted values for each serostatus group, and 95% confidence bands, are shown for this model in Figure 5-1.

Because of the high variation in time difference between Visits B and C among subjects, we fitted a model which used continuous time since Visit A rather than discrete Visit numbers. LOESS (locally estimated scatterplot smoothing) models were used to accomplish this. For each of the five epigenetic outcome measures, we stratified the data by SN only and SC only, and two LOESS models were trained, for a total of ten. Fitted values for each serostatus group, and 95% confidence bands, are shown for this model in Figure 5-2.

Pairwise correlation analyses were performed for each of the absolute T cell subsets counts as well as all of the epigenetic measures, for all participants at Visit A, for SC at Visit B, and for SN at Visit B. The analyses were repeated for percentages of T cell subsets. Pearson correlation coefficients (rho) and p values are shown in Tables S3A-E, and Tables S4A-E.

Potential contributions of covariates (study visit, HIV serostatus group, interaction between study visit*HIV serostatus group, race [non-white vs. white], Body Mass Index [BMI], and smoking cumulative pack years) to the changes in each age-adjusted epigenetic measure between the four visits were analyzed in linear mixed effects models using random intercepts, with all participants in the same model. Due to missing data for some demographic covariates, n=100 samples for these mixed models. The F-values and p-values from the mixed effect models for all five epigenetic measures are shown in Table 5-2; individual parameter estimates and p-values from fixed effects analyses are reported in Table S2.

Potential contributions of absolute counts of T cell subsets (total CD4, total CD8, naïve CD4, naïve CD8, activated CD4, activated CD8, senescent CD4, senescent CD8) to the changes in each age-adjusted epigenetic measure between the four visits were analyzed in linear mixed effects models using random intercept, with all participants in the same model. All absolute T cell counts were natural log-transformed (log(cells/mm^3)) before inclusion into mixed models. Due to missing data for some flow cytometry variables, n=72 samples for these mixed models. We selected a consensus model to be used across all five epigenetic measures, which consisted of the following covariates: absolute counts of total CD8, percentage of naïve CD4, percentage of naïve CD8, percentage of senescent CD8, and percentage of activated CD8. The F-values and p-values from the mixed effect models for all five epigenetic measures are shown in Table 5-5; individual parameter estimates and p-values from fixed effects analyses are reported in Table S5.

The consensus model was selected in part using multivariate adaptive regression spline (MARS) models. MARS is a machine learning algorithm that can be used to aid model selection and fits flexible adaptive b-splines (Friedman, 1991). This allows the model to capture changes to epigenetic measures and T cell subset composition over time due to two highly exogenous events among SC subjects (i.e., HIV infection and HAART initiation). The consensus model was selected in three steps, as follows:

1. We fitted multivariate adaptive regression spline (MARS) models for each of the five epigenetic measures, using all 16 possible T cell measurements as covariates (8 absolute counts, 8 percentages, all natural log-transformed), along with serostatus group (SC vs. SN) and time since visit A (and an interaction term). Based on the MARS models' reports of estimated variable importance, removed exactly half of the T cell variables from our candidate model, keeping either an absolute count or a percentage for each of the 8 T cell subsets. The majority of the time, the percentage measurements were ranked higher in importance than the absolute count measurements, with the exception of total CD8 cells.
2. We examined the correlation of each covariate with each epigenetic measure, among only SC samples from Visit B, C, or D. From this, we found two T cell subsets, senescent CD4 and activated CD4, to be not significantly correlated (positively or negatively) with any of the epigenetic measures, and removed them from the candidate model.
3. We examined co-linearity between the remaining 6 covariates, and found total CD4 to be correlated with almost all of the other 5 covariates, and thus removed it from the candidate model.

These MARS models (above) were also used as the basis for identifying T cell cut-points that are shown in Figure S1. We curated said figure to show cut-points and epigenetic measures for which their final fitted splines (in the respective MARS models) had noticeably high-magnitude slopes, or had sharp changes in slope.

For the analyses with HIV detectability at Visit D, the mixed model is essentially the same as the one shown in Figure 5-1, except replacing the dichotomous variable for serostatus group (SC vs. SN) with a trichotomous variable for HIV detectability (ID vs. IU vs. UU). Fitted values for each HIV detectability group, and 95% confidence bands, are shown for this model in

Figure 5-3; individual parameter estimates and p-values from fixed effects analyses are reported in Table 5-6.

APPENDIX

Table S1: Features of Epigenetic Measures of Biologic Aging that apply to the Illumina 450 and EPIC array platforms, related to Figures 1, 2, and 3, Tables 2 and 4

Epigenetic measure	Pan-tissue age clock (Horvath)	Extrinsic age clock	Phenotypic age clock	Grim age clock	DNA methylation-based estimate of telomere length
Age-adjusted measure used in analyses	Age Acceleration Residual (AAR)	Extrinsic Epigenetic Age Acceleration (EEAA)	Phenotypic Epigenetic Age Acceleration (PEAA)	Grim Epigenetic Age Acceleration (PEAA)	Age-adjusted DNAmTL (aaDNAmTL)
Number of CpGs	353	71	513	1030	140
Tissue(s) in which the measure was developed	Many human tissues and cell types	Peripheral blood	Peripheral blood	Peripheral blood	Leukocytes, adipose, liver, monocytes
Methodology of development	Estimator of chronologic age developed on the basis of a wide spectrum of human tissues and cell types. It applies to all tissues and nucleated cell types.	Estimator of chronologic age for blood methylation data. It is defined as a weighted average of the Hannum clock and 3 cell types known to change with age: naïve and senescent T cells, and plasmablasts.	Methylation-based estimator of "phenotypic age" which was defined as a linear combination of several clinical parameters. While it is less predictive of chronologic age than the pan tissue clock, it is more predictive of mortality risk and many age-related conditions.	Methylation-based predictor of mortality risk, developed as an estimator of log transformed hazard ratio from Cox regression model analysis of time-to-death. Covariates include chronologic age, sex, and methylation-based surrogate biomarkers of smoking pack-years and plasma proteins.	Methylation-based estimator of leukocyte telomere length. It correlates negatively (i.e., shorter telomere length) with increasing age and obesity in blood and other tissues. Its correlation to actual telomere length is relatively weak (r=0.35).
Special features	Widely used in epidemiologic studies. Accelerated in HIV, obesity, neurodegenerative disease, and many disease states. Weakly predictive of mortality risk. Weak association with blood cell composition. Not related to smoking.	Captures both cell-intrinsic methylation changes and age-related changes in blood cell composition. Strongly correlated with blood cell counts: positively correlated with senescent T lymphocytes and negatively correlated with naïve T lymphocyte counts. Reflects aspects of immunosenescence, underlying age-related decline in the protective immune response.	Strongly predictive of healthspan and lifespan. Correlated with multi-morbidity, frailty. Correlated with smoking and markers of immunosenescence.	Strong predictor of mortality risk. Predictive of time to cancer. Associated with age at menopause, frailty, heart disease, metabolic syndrome, fatty liver. Correlated with markers of immunosenescence. Statistical analysis needs to adjust for chronologic age and sex as these variables are in the definition.	Reflects cell replicative history. Useful marker of age-related pathologies. Outperforms measured leukocyte telomere length in predicting time to death, time to coronary heart disease, time to congestive heart failure, and association with smoking history. Associated with physical functioning, dietary variables, education and income.

Table S2: Individual parameter estimates and p values from mixed effects models incorporating demographic factors for all five epigenetic measures, related to Table 5-2

Regression Fixed Effects	Coefficient Estimates (Standard Errors) [a]				
	AAR	EEAA	PEAA	GEAA	aaDNAmTL
(Intercept)	-3.3 (2.25)	-2.6 (2.26)	-2.72 (3.19)	1.4 (1.37)	0.22 (0.12)
Study Visit, Visit B vs A	0.18 (0.58)	-0.08 (0.55)	-1 (0.86)	-0.56 (0.33)	-0.01 (0.03)
Study Visit, Visit C vs A	-0.56 (0.6)	-0.92 (0.58)	-1.61 (0.9)	-0.83 (0.35)	0.06 (0.03)
Study Visit, Visit D vs A	-1.12 (0.61)	-0.96 (0.59)	-2.14 (0.91)	-1.07 (0.35)	0.08 (0.03)
HIV Serostatus Group, SC vs SN [b]	-0.07 (1.06)	-0.53 (1.09)	1.31 (1.48)	1.75 (0.66)	-0.03 (0.06)
Visit B vs A * SC vs SN	1.21 (0.85)	0.61 (0.81)	4.99 (1.26)	0.28 (0.49)	-0.22 (0.04)
Visit C vs A * SC vs SN	5.07 (0.85)	4.85 (0.81)	9.57 (1.27)	0.14 (0.49)	-0.45 (0.04)
Visit D vs A * SC vs SN	4.3 (0.85)	2.58 (0.82)	5.96 (1.28)	0.57 (0.49)	-0.29 (0.04)
Race, white vs non-white	-1 (1.23)	-1.16 (1.29)	-0.38 (1.67)	-2.01 (0.79)	-0.04 (0.07)
BMI, kg/m^2	0.11 (0.08)	0.11 (0.08)	0.03 (0.11)	-0.04 (0.05)	0 (0)
Smoking, cumulative pack years	0.02 (0.03)	0.05 (0.03)	0.04 (0.04)	0.1 (0.02)	0 (0)

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length, BMI = Body Mass Index

a: Coefficient values and standard errors for fixed effects (bold if Wald test p-value <0.01) from mixed models (n=400 out of 480 total observations due to missing data for some covariates, 54 SN and 46 SC) in a single model

b: HIV serostatus groups classified as SC (became HIV infected and seroconverted between visits A and B) vs SN (persistently HIV uninfected and seronegative at visits A through D)

Table S3A: Pairwise correlations of absolute T cell counts to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SN participants at all Visits (all HIV-uninfected), related to Table 5-3

Visits A-D, SN	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	0.46 <0.001 234	0.78 <0.001 231	0.47 <0.001 231	0.21 0.0012 231	0.24 <0.001 231	0.74 <0.001 226	0.11 0.095 226	0 0.97 234	-0.05 0.49 234	-0.16 0.012 234	0.07 0.28 234	0.09 0.17 234
CD8	—	1.00	0.23 <0.001 231	0.5 <0.001 231	0.5 <0.001 231	0.77 <0.001 231	0.34 <0.001 226	0.56 <0.001 226	0.25 <0.001 234	0.15 0.019 234	0.01 0.88 234	0.19 0.0031 234	-0.29 <0.001 234
Naïve CD4	—	—	1.00	0.62 <0.001 231	-0.15 0.019 231	0.05 0.42 231	0.71 <0.001 225	0.04 0.57 225	-0.15 0.026 231	-0.32 <0.001 231	-0.43 <0.001 231	-0.09 0.18 231	0.38 <0.001 231
Naïve CD8	—	—	—	1.00	-0.07 0.32 231	0.05 0.47 231	0.34 <0.001 225	0.12 0.064 225	-0.12 0.069 231	-0.41 <0.001 231	-0.44 <0.001 231	-0.05 0.42 231	0.42 <0.001 231
Senescent CD4	—	—	—	—	1.00	0.54 <0.001 231	0.1 0.15 225	0.15 0.027 225*	0.38 <0.001 231	0.41 <0.001 231	0.31 <0.001 231	0.21 0.0012 231	-0.36 <0.001 231
Senescent CD8	—	—	—	—	—	1.00	0.25 <0.001 225	0.5 <0.001 225	0.44 <0.001 231	0.41 <0.001 231	0.23 <0.001 231	0.16 0.016 231	-0.49 <0.001 231
Activated CD4	—	—	—	—	—	—	1.00	0.34 <0.001 226	0.05 0.47 226	-0.05 0.49 226	-0.14 0.037 226	-0.05 0.44 226	0.06 0.35 226
Activated CD8	—	—	—	—	—	—	—	1.00	0.24 <0.001 226	0.2 0.0021 226	0.09 0.19 226	0.11 0.11 226	-0.36 <0.001 226

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S3B: Pairwise correlations of absolute T cell counts to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SC participants at Visit A (all HIV-uninfected), related to Table 5-3

	Pearson Correlation Coefficient (p-value) n [a]												
Visit A, SC	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	0.59 <0.001 53	0.65 <0.001 53	0.42 0.0016 53	0.3 0.03 53	0.25 0.067 53	0.52 <0.001 51	0.09 0.52 51	0.1 0.46 53	-0.03 0.83 53	-0.14 0.33 53	0.02 0.87 53	0.16 0.24 53
CD8	—	1.00	0.49 <0.001 53	0.53 <0.001 53	0.32 0.021 53	0.73 <0.001 53	0.03 0.85 51	0.3 0.034 51	0.37 0.0059 53	0.12 0.4 53	-0.02 0.87 53	0.04 0.78 53	-0.08 0.58 53
Naïve CD4	—	—	1.00	0.68 <0.001 53	0.02 0.88 53	0.22 0.11 53	0.35 0.012 51	0.03 0.83 51	-0.16 0.24 53	-0.48 <0.001 53	-0.45 <0.001 53	-0.26 0.057 53	0.52 <0.001 53
Naïve CD8	—	—	—	1.00	0.04 0.76 53	0.16 0.27 53	0.14 0.34 51	-0.01 0.92 51	-0.07 0.6 53	-0.37 0.0061 53	-0.42 0.0018 53	-0.2 0.15 53	0.54 <0.001 53
Senescent CD4	—	—	—	—	1.00	0.25 0.074 53	0.2 0.15 51	0.04 0.8 51	0.31 0.026 53	0.24 0.081 53	0.15 0.29 53	0.08 0.59 53	-0.21 0.14 53
Senescent CD8	—	—	—	—	—	1.00	-0.16 0.25 51	0.25 0.075 51	0.36 0.0085 53	0.24 0.089 53	-0.06 0.66 53	0.05 0.7 53	-0.24 0.079 53
Activated CD4	—	—	—	—	—	—	1.00	0.25 0.077 51	-0.24 0.095 51	-0.19 0.18 51	-0.24 0.093 51	-0.23 0.11 51	0.18 0.2 51
Activated CD8	—	—	—	—	—	—	—	1.00	0.03 0.82 51	-0.03 0.82 51	-0.03 0.85 51	-0.09 0.51 51	-0.16 0.25 51

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S3C: Pairwise correlations of absolute T cell counts to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SC participants at Visit B (recently HIV-infected), related to Table 5-3

Visit B, SC	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	0.08 0.56 59	0.74 <0.001 59	0.29 0.027 59	0.29 0.024 59	0.11 0.4 59	0.35 0.0081 57	-0.14 0.28 57	-0.12 0.38 59	-0.12 0.35 59	-0.28 0.034 59	-0.13 0.34 59	0.24 0.072 59
CD8	—	1.00	0.02 0.9 59	0.05 0.71 59	0.34 0.0075 59	0.72 <0.001 59	0.35 0.0082 57	0.75 <0.001 57	0.34 0.0081 59	0.43 <0.001 59	0.42 <0.001 59	0.14 0.3 59	-0.52 <0.001 59
Naïve CD4	—	—	1.00	0.55 <0.001 59	-0.03 0.85 59	0.13 0.33 59	0.09 0.53 57	-0.19 0.17 57	-0.17 0.2 59	-0.23 0.086 59	-0.31 0.017 59	-0.15 0.24 59	0.42 <0.001 59
Naïve CD8	—	—	—	1.00	-0.02 0.86 59	0.05 0.71 59	-0.18 0.18 57	-0.25 0.059 57	-0.14 0.28 59	-0.21 0.1 59	-0.27 0.038 59	-0.07 0.6 59	0.45 <0.001 59
Senescent CD4	—	—	—	—	1.00	0.28 0.029 59	0.5 <0.001 57	0.3 0.024 57	0.18 0.16 59	0.13 0.31 59	0.13 0.32 59	0.05 0.72 59	-0.29 0.024 59
Senescent CD8	—	—	—	—	—	1.00	0.13 0.35 57	0.35 0.0074 57	0.35 0.0074 59	0.39 0.0025 59	0.27 0.036 59	0.09 0.51 59	-0.35 0.0073 59
Activated CD4	—	—	—	—	—	—	1.00	0.58 <0.001 57	-0.03 0.8 57	0.05 0.72 57	0.06 0.66 57	-0.06 0.67 57	-0.22 0.11 57
Activated CD8	—	—	—	—	—	—	—	1.00	0.14 0.3 57	0.23 0.091 57	0.35 0.0075 57	0 0.99 57	-0.5 <0.001 57

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S3D: Pairwise correlations of absolute T cell counts to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SC participants at Visit C (living with uncontrolled HIV), related to Table 5-3

Visit C, SC	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	0.03 0.84 59	0.87 <0.001 57	0.3 0.023 57	0.05 0.72 57	-0.12 0.39 57	0.46 <0.001 57	-0.03 0.83 57	-0.34 0.0076 59	-0.29 0.027 59	-0.35 0.0065 59	-0.1 0.44 59	0.48 <0.001 59
CD8	—	1.00	-0.1 0.45 57	0.25 0.056 57	0.35 0.0076 57	0.72 <0.001 57	0.26 0.052 57	0.69 <0.001 57	0.29 0.027 59	0.41 0.0014 59	0.27 0.036 59	-0.11 0.39 59	-0.56 <0.001 59
Naïve CD4	—	—	1.00	0.51 <0.001 57	-0.19 0.17 57	-0.2 0.14 57	0.13 0.34 56	-0.2 0.13 56	-0.36 0.006 57	-0.38 0.0037 57	-0.44 <0.001 57	-0.15 0.25 57	0.57 <0.001 57
Naïve CD8	—	—	—	1.00	-0.14 0.3 57	-0.04 0.78 57	-0.16 0.24 56	-0.03 0.83 56	-0.06 0.67 57	-0.23 0.08 57	-0.23 0.087 57	-0.13 0.33 57	0.29 0.03 57
Senescent CD4	—	—	—	—	1.00	0.55 <0.001 57	0.19 0.16 56	0.17 0.21 56	0.03 0.84 57	0.12 0.37 57	0.04 0.75 57	0.01 0.94 57	-0.14 0.3 57
Senescent CD8	—	—	—	—	—	1.00	0.09 0.51 56	0.34 0.0095 56	0.25 0.066 57	0.4 0.0018 57	0.23 0.088 57	-0.1 0.46 57	-0.48 <0.001 57
Activated CD4	—	—	—	—	—	—	1.00	0.61 <0.001 57	-0.12 0.37 57	-0.08 0.56 57	0.08 0.53 57	0.12 0.37 57	0.01 0.95 57
Activated CD8	—	—	—	—	—	—	—	1.00	0.21 0.12 57	0.16 0.24 57	0.3 0.024 57	0.04 0.78 57	-0.41 0.0015 57

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S3E: Pairwise correlations of absolute T cell counts to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SC participants at Visit D (recently receiving HAART), related to Table 5-3

Visit D, SC	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	-0.03 0.82 60	0.85 <0.001 59	0.2 0.13 59	0.11 0.4 59	-0.16 0.24 59	0.48 <0.001 59	-0.32 0.014 59	-0.28 0.028 60	-0.27 0.037 60	-0.38 0.003 60	-0.32 0.013 60	0.3 0.019 60
CD8	—	1.00	-0.1 0.47 59	0.35 0.0064 59	0.01 0.97 59	0.76 <0.001 59	0.03 0.83 59	0.65 <0.001 59	0.33 0.01 60	0.19 0.14 60	0.17 0.2 60	0.01 0.96 60	-0.42 <0.001 60
Naïve CD4	—	—	1.00	0.41 0.0014 59	-0.11 0.4 59	-0.26 0.047 59	0.42 0.0011 59	-0.28 0.03 59	-0.33 0.011 59	-0.38 0.0032 59	-0.45 <0.001 59	-0.42 <0.001 59	0.46 <0.001 59
Naïve CD8	—	—	—	1.00	-0.14 0.28 59	0.01 0.96 59	0.11 0.41 59	0 0.98 59	0.01 0.95 59	-0.35 0.0073 59	-0.22 0.094 59	-0.19 0.16 59	0.28 0.03 59
Senescent CD4	—	—	—	—	1.00	0.15 0.25 59	0.1 0.43 59	-0.14 0.28 59	0.13 0.31 59	0.08 0.54 59	0.16 0.21 59	0.08 0.55 59	-0.23 0.078 59
Senescent CD8	—	—	—	—	—	1.00	-0.08 0.55 59	0.49 <0.001 59	0.26 0.045 59	0.21 0.1 59	0.07 0.58 59	-0.02 0.9 59	-0.37 0.0034 59
Activated CD4	—	—	—	—	—	—	1.00	0.15 0.26 59	-0.14 0.3 59	-0.15 0.26 59	-0.09 0.49 59	-0.15 0.26 59	0.15 0.27 59
Activated CD8	—	—	—	—	—	—	—	1.00	0.14 0.3 59	0.09 0.51 59	0.14 0.29 59	0.09 0.5 59	-0.24 0.062 59

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S4A: Pairwise correlations of percentages of T cell subsets within live lymphocytes to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SN participants at all Visits (all HIV-uninfected), related to Table 5-4

Visits A-D, SN	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	-0.97 <0.001 204	0.46 <0.001 204	-0.19 0.0061 204	-0.14 0.044 204	-0.59 <0.001 204	0.28 <0.001 204	-0.53 <0.001 204	-0.26 <0.001 204	-0.16 0.018 204	-0.14 0.041 204	-0.08 0.24 204	0.43 <0.001 204
CD8	—	1.00	-0.41 <0.001 204	0.2 0.0035 204	0.18 0.011 204	0.66 <0.001 204	-0.24 <0.001 204	0.58 <0.001 204	0.27 <0.001 204	0.18 0.011 204	0.14 0.043 204	0.12 0.079 204	-0.44 <0.001 204
Naïve CD4	—	—	1.00	0.47 <0.001 207	-0.25 <0.001 207	-0.17 0.014 207	0.46 <0.001 204	-0.17 0.013 204	-0.18 0.0094 207	-0.4 <0.001 207	-0.48 <0.001 207	-0.13 0.062 207	0.48 <0.001 207
Naïve CD8	—	—	—	1.00	-0.21 0.0021 207	-0.13 0.07 207	0.05 0.51 204	0.02 0.74 204	-0.11 0.11 207	-0.54 <0.001 207	-0.5 <0.001 207	-0.14 0.039 207	0.47 <0.001 207
Senescent CD4	—	—	—	—	1.00	0.37 <0.001 207	-0.03 0.66 204	0.03 0.67 204	0.41 <0.001 207	0.42 <0.001 207	0.32 <0.001 207	0.02 0.8 207	-0.29 <0.001 207
Senescent CD8	—	—	—	—	—	1.00	0.04 0.6 204	0.5 <0.001 204	0.48 <0.001 207	0.44 <0.001 207	0.31 <0.001 207	0.12 0.099 207	-0.53 <0.001 207
Activated CD4	—	—	—	—	—	—	1.00	0.18 0.012 204	-0.04 0.62 204	-0.07 0.3 204	-0.16 0.021 204	-0.16 0.022 204	0.08 0.27 204
Activated CD8	—	—	—	—	—	—	—	1.00	0.2 0.0047 204	0.17 0.015 204	0.15 0.032 204	0.15 0.031 204	-0.39 <0.001 204

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S4B: Pairwise correlations of percentages of T cell subsets within live lymphocytes to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SC participants at Visit A (all HIV-uninfected), related to Table 5-4

Visit A, SC	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	-0.97 <0.001 53	0.36 0.0085 53	-0.13 0.36 53	-0.06 0.64 53	-0.58 <0.001 53	0.46 <0.001 53	-0.26 0.065 53	-0.26 0.062 53	-0.26 0.062 53	-0.17 0.24 53	-0.1 0.49 53	0.4 0.0026 53
CD8	—	1.00	-0.29 0.037 53	0.22 0.12 53	0.08 0.59 53	0.6 <0.001 53	-0.5 <0.001 53	0.19 0.16 53	0.27 0.052 53	0.24 0.084 53	0.11 0.43 53	0.08 0.55 53	-0.36 0.0085 53
Naïve CD4	—	—	1.00	0.47 <0.001 56	-0.27 0.041 56	-0.1 0.47 56	0.14 0.33 53	-0.05 0.71 53	-0.34 0.0098 56	-0.64 <0.001 56	-0.56 <0.001 56	-0.39 0.0033 56	0.63 <0.001 56
Naïve CD8	—	—	—	1.00	-0.07 0.6 56	-0.04 0.77 56	-0.08 0.57 53	-0.15 0.28 53	-0.2 0.15 56	-0.39 0.0032 56	-0.46 <0.001 56	-0.22 0.11 56	0.51 <0.001 56
Senescent CD4	—	—	—	—	1.00	0.16 0.23 56	0.05 0.74 53	-0.03 0.85 53	0.21 0.12 56	0.17 0.2 56	0.15 0.28 56	0.07 0.62 56	-0.18 0.19 56
Senescent CD8	—	—	—	—	—	1.00	-0.31 0.026 53	0.24 0.087 53	0.24 0.069 56	0.21 0.12 56	-0.07 0.6 56	0.01 0.97 56	-0.34 0.011 56
Activated CD4	—	—	—	—	—	—	1.00	0.27 0.055 53	-0.44 <0.001 53	-0.33 0.015 53	-0.26 0.064 53	-0.27 0.052 53	0.25 0.072 53
Activated CD8	—	—	—	—	—	—	—	1.00	-0.09 0.53 53	-0.07 0.63 53	0.02 0.91 53	-0.09 0.53 53	-0.21 0.14 53

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S4C: Pairwise correlations of percentages of T cell subsets within live lymphocytes to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SC participants at Visit B (recently HIV-infected), related to Table 5-4

Visit B, SC	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	-0.99 <0.001 52	0.77 <0.001 52	0.4 0.0032 52	0.03 0.82 52	-0.45 <0.001 52	0.17 0.23 52	-0.64 <0.001 52	-0.41 0.0027 52	-0.44 0.0012 52	-0.52 <0.001 52	-0.16 0.26 52	0.73 <0.001 52
CD8	—	1.00	-0.74 <0.001 52	-0.39 0.0046 52	-0.02 0.88 52	0.44 0.001 52	-0.18 0.2 52	0.65 <0.001 52	0.39 0.0046 52	0.4 0.0031 52	0.49 <0.001 52	0.17 0.22 52	-0.7 <0.001 52
Naïve CD4	—	—	1.00	0.62 <0.001 54	-0.16 0.25 54	-0.22 0.11 54	0.02 0.86 52	-0.5 <0.001 52	-0.36 0.0074 54	-0.42 0.0017 54	-0.48 <0.001 54	-0.16 0.24 54	0.65 <0.001 54
Naïve CD8	—	—	—	1.00	0 0.98 54	-0.2 0.16 54	-0.21 0.13 52	-0.49 <0.001 52	-0.17 0.23 54	-0.38 0.0045 54	-0.41 0.0023 54	-0.17 0.21 54	0.52 <0.001 54
Senescent CD4	—	—	—	—	1.00	0.16 0.25 54	0.25 0.069 52	-0.02 0.89 52	0.2 0.14 54	0.08 0.55 54	-0.01 0.97 54	-0.01 0.96 54	-0.1 0.45 54
Senescent CD8	—	—	—	—	—	1.00	-0.09 0.53 52	0.13 0.34 52	0.37 0.0063 54	0.35 0.0094 54	0.21 0.13 54	0.01 0.96 54	-0.32 0.018 54
Activated CD4	—	—	—	—	—	—	1.00	0.4 0.0035 52	-0.24 0.091 52	-0.19 0.19 52	-0.18 0.19 52	-0.08 0.59 52	0.14 0.33 52
Activated CD8	—	—	—	—	—	—	—	1.00	0.06 0.67 52	0.12 0.4 52	0.27 0.057 52	0.02 0.91 52	-0.5 <0.001 52

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S4D: Pairwise correlations of percentages of T cell subsets within live lymphocytes to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SC participants at Visit C (living with uncontrolled HIV), related to Table 5-4

Visit C, SC	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	-0.98 <0.001 52	0.83 <0.001 52	0.1 0.47 52	-0.09 0.51 52	-0.49 <0.001 52	0.32 0.021 52	-0.39 0.0039 52	-0.44 0.001 52	-0.45 <0.001 52	-0.39 0.0041 52	-0.06 0.69 52	0.71 <0.001 52
CD8	—	1.00	-0.8 <0.001 52	-0.14 0.33 52	0.1 0.49 52	0.49 <0.001 52	-0.29 0.04 52	0.44 0.0011 52	0.42 0.0017 52	0.45 <0.001 52	0.38 0.006 52	0.03 0.81 52	-0.71 <0.001 52
Naïve CD4	—	—	1.00	0.41 0.0027 52	-0.2 0.15 52	-0.38 0.0053 52	0.13 0.35 52	-0.34 0.013 52	-0.39 0.0047 52	-0.5 <0.001 52	-0.47 <0.001 52	-0.27 0.051 52	0.63 <0.001 52
Naïve CD8	—	—	—	1.00	-0.21 0.14 52	-0.28 0.046 52	-0.27 0.053 52	-0.27 0.055 52	-0.08 0.59 52	-0.38 0.0056 52	-0.3 0.033 52	-0.11 0.44 52	0.34 0.013 52
Senescent CD4	—	—	—	—	1.00	0.41 0.0029 52	0.07 0.62 52	0.03 0.83 52	-0.1 0.47 52	0.05 0.71 52	-0.03 0.85 52	-0.07 0.64 52	-0.04 0.8 52
Senescent CD8	—	—	—	—	—	1.00	-0.11 0.45 52	0.19 0.19 52	0.29 0.039 52	0.39 0.004 52	0.23 0.097 52	-0.14 0.31 52	-0.53 <0.001 52
Activated CD4	—	—	—	—	—	—	1.00	0.54 <0.001 52	-0.15 0.29 52	-0.22 0.12 52	0 0.99 52	0.02 0.89 52	0.13 0.36 52
Activated CD8	—	—	—	—	—	—	—	1.00	0.21 0.13 52	0.07 0.61 52	0.33 0.019 52	0.03 0.82 52	-0.34 0.014 52

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S4E: Pairwise correlations of percentages of T cell subsets within live lymphocytes to each other, and to each of the epigenetic clocks (AAR, EEAA, PEAA, GEAA) and estimated telomere length (aaDNAmTL), among SC participants at Visit D (recently receiving HAART), related to Table 5-4

Visit D, SC	Pearson Correlation Coefficient (p-value) n [a]												
	CD4	CD8	Naïve CD4	Naïve CD8	Senescent CD4	Senescent CD8	Activated CD4	Activated CD8	AAR	EEAA	PEAA	GEAA	aaDNAmTL
CD4	1.00	-0.98 <0.001 52	0.81 <0.001 52	0.02 0.9 52	0.03 0.83 52	-0.6 <0.001 52	0.56 <0.001 52	-0.51 <0.001 52	-0.35 0.011 52	-0.31 0.024 52	-0.32 0.019 52	-0.23 0.1 52	0.45 <0.001 52
CD8	—	1.00	-0.78 <0.001 52	0 0.98 52	-0.03 0.84 52	0.6 <0.001 52	-0.58 <0.001 52	0.5 <0.001 52	0.32 0.02 52	0.27 0.051 52	0.27 0.05 52	0.21 0.13 52	-0.41 0.0028 52
Naïve CD4	—	—	1.00	0.28 0.045 52	-0.08 0.56 52	-0.41 0.0024 52	0.47 <0.001 52	-0.38 0.0049 52	-0.32 0.02 52	-0.37 0.0065 52	-0.38 0.0059 52	-0.35 0.012 52	0.46 <0.001 52
Naïve CD8	—	—	—	1.00	-0.01 0.96 52	-0.13 0.34 52	-0.09 0.52 52	-0.26 0.06 52	0.04 0.8 52	-0.36 0.0085 52	-0.21 0.13 52	-0.15 0.27 52	0.34 0.012 52
Senescent CD4	—	—	—	—	1.00	0.2 0.15 52	0.09 0.51 52	-0.08 0.55 52	0.12 0.4 52	0.07 0.64 52	0.15 0.29 52	0.06 0.67 52	-0.22 0.12 52
Senescent CD8	—	—	—	—	—	1.00	-0.24 0.087 52	0.45 <0.001 52	0.27 0.049 52	0.21 0.13 52	0.14 0.33 52	0 0.97 52	-0.37 0.0077 52
Activated CD4	—	—	—	—	—	—	1.00	0.05 0.7 52	-0.31 0.027 52	-0.36 0.0087 52	-0.15 0.28 52	-0.17 0.22 52	0.32 0.022 52
Activated CD8	—	—	—	—	—	—	—	1.00	-0.02 0.88 52	-0.02 0.9 52	0.08 0.55 52	0.08 0.59 52	-0.18 0.21 52

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Cell in bold if p-value < 0.01

Table S5: Individual parameter estimates and p values from mixed effects models incorporating natural log-transformed absolute T cell subset counts and percentages of T cell subsets within live lymphocytes for all five epigenetic measures, related to Table 5-5

	Coefficient Estimates (Standard Errors) [a]				
Regression Fixed Effects	AAR	EEAA	PEAA	GEAA	aaDNAmTL
(Intercept)	-1.34 (4.95)	0.55 (4.6)	12.12 (6.17)	0.11 (2.67)	0.23 (0.21)
Study Visit, Visit B vs A	-0.41 (0.81)	-0.23 (0.76)	-1.6 (1.01)	-0.63 (0.41)	0.02 (0.04)
Study Visit, Visit C vs A	-1.1 (0.82)	-1.34 (0.76)	-2.52 (1.01)	-1.26 (0.41)	0.09 (0.04)
Study Visit, Visit D vs A	-1.38 (0.83)	-1.05 (0.77)	-3.41 (1.03)	-1.19 (0.42)	0.12 (0.04)
HIV Serostatus Group, SC vs SN [b]	-2.58 (1.17)	-1.82 (1.08)	-1.73 (1.47)	1.17 (0.82)	0.02 (0.05)
Visit B vs A * SC vs SN	0.49 (1.22)	-0.11 (1.14)	3.25 (1.52)	0.27 (0.62)	-0.08 (0.05)
Visit C vs A * SC vs SN	2.89 (1.38)	2.83 (1.28)	5.07 (1.71)	0.36 (0.7)	-0.16 (0.06)
Visit D vs A * SC vs SN	2.02 (1.2)	0.43 (1.11)	4.47 (1.49)	0.82 (0.61)	-0.13 (0.05)
Absolute Count CD8 T cells, log(cells/mm^3) [c]	0.62 (0.79)	0.84 (0.73)	-0.79 (0.98)	0.25 (0.42)	-0.07 (0.03)
% Naïve (CD45RA+CCR7+) CD4 T cells, log(%)	-24.9 (6.13)	-17.08 (5.69)	-35.86 (7.67)	-6.66 (3.48)	1.45 (0.26)
% Naïve (CD45RA+CCR7+) CD8 T cells, log(%)	3.53 (11.64)	-62.49 (10.8)	-50.5 (14.59)	-6.56 (6.78)	2.25 (0.5)
% Senescent (CD28-CD57+) CD8 T cells, log(%)	29.4 (7.94)	30.51 (7.37)	17.7 (9.91)	-0.75 (4.39)	-1.98 (0.34)
% Activated (HLA DR+CD38+) CD8 T cells, log(%)	-0.52 (10.46)	-18.31 (9.7)	24.82 (12.99)	-4.15 (5.44)	-0.87 (0.46)

AAR = Age-Acceleration Residual, EEAA = Extrinsic Epigenetic Age Acceleration, PEAA = Phenotypic Epigenetic Age Acceleration, aaDNAmTL = age-adjusted DNA methylation-based estimate of Telomere Length

a: Coefficient values and standard errors for fixed effects (bold if Wald test p-value <0.01) from mixed models (n=288 out of 480 total observations due to missing data for some covariates, 34 SN and 38 SC) in a single model
b: HIV serostatus groups classified as SC (became HIV infected and seroconverted between visits A and B) vs SN (persistently HIV uninfected and seronegative at visits A through D)
c: Absolute counts of T cell subsets as described in STAR Methods and Table S6; all cell counts and percentages were natural log-transformed (log) for analyses

Table S6: Multicolor flow cytometry panels for determination of total CD3 T cells, and CD4, CD8, naïve CD4 or CD8, senescent CD4 or CD8, and activated CD4 or CD8 T cell subsets, related to STAR Methods

	Attune NxT Flow Cytometer Channel Antibody Label							
	BL-2	RL-1	RL-3	VL-1	VL-2	VL-3	YL-1	YL-3
Flow Cytometry Panel	PerCP	AF647	APC-Cy7	V450	Zombie Aqua	BV605	PE	PE-Cy7
Naïve and Senescent tube (antibody volume/tube)	CD3 (20 µL)	CCR7 (20 µL)	CD8 (5 µL)	CD4 (5 µL)	Zombie Aqua (100 µL)	CD57 (5 µL)	CD28 (20 µL)	CD45RA (5 µL)
Activated tube (antibody volume/tube)	CD3 (20 µL)	----	CD8 (5 µL)	CD4 (5 µL)	Zombie Aqua (100 µL)	HLA-DR (5 µL)	CD38 (20 µL)	----
Isotype tube (antibody volume/tube)	CD3 (20 µL)	IgG2a (5 µL)	CD8 (5 µL)	CD4 (5 µL)	Zombie Aqua (100 µL)	IgG2a (2.5 µL)	IgG1 (20 µL)	IgG1 (5 µL)

PerCP = Peridinin-Chlorophyll-Protein, AF647 = Alexa Fluor 647, APC-Cy7 = Allophycocyanin-Cyanine7 tandem, V450 = horizon V450, BV605 = Brilliant Violet 605, PE = Phycoerythrin, PE-Cy7 = Phycoerythrin-Cyanine7 tandem

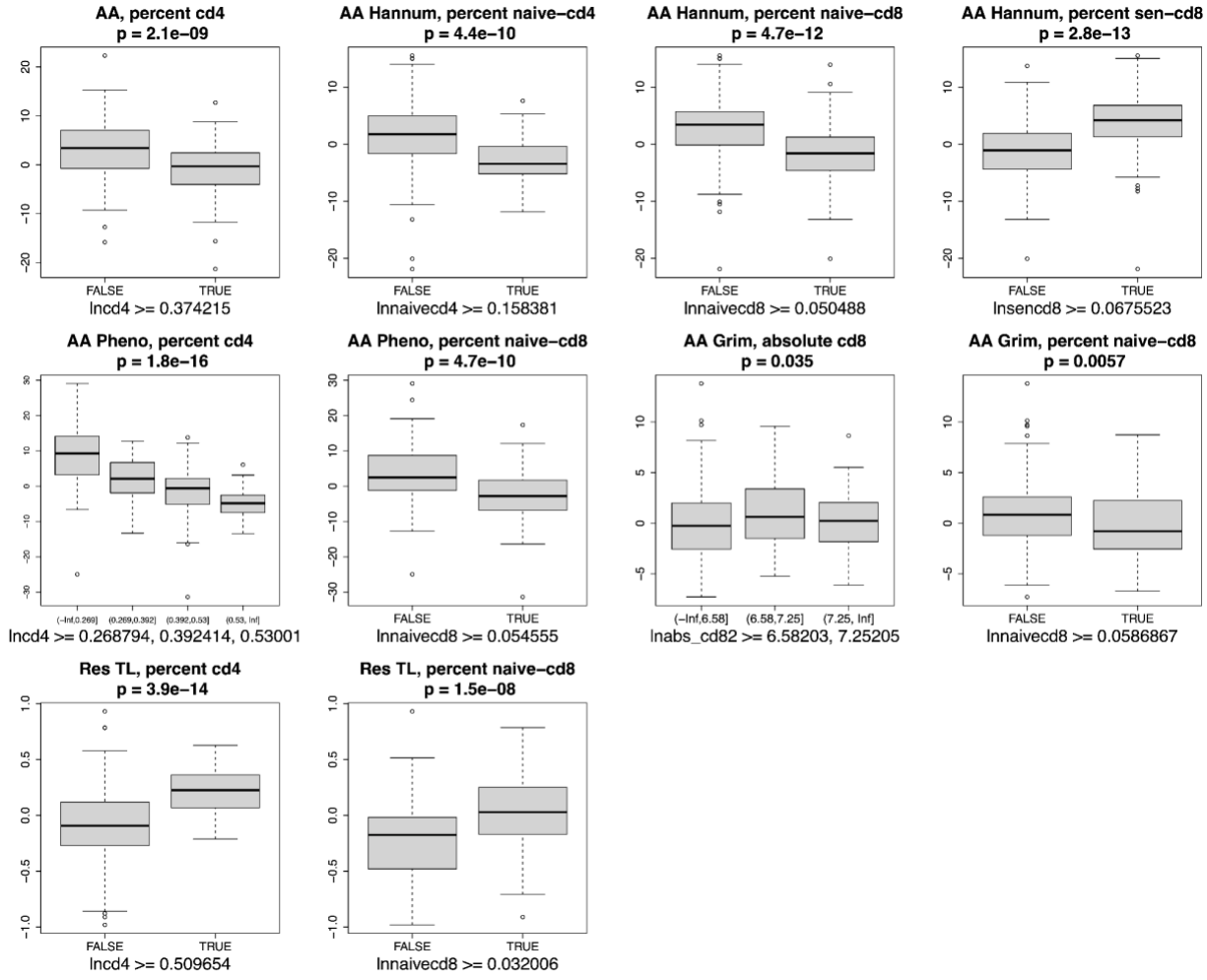

Figure S1: Multiple epigenetic measures demonstrate significant differences in aging across specific cut-points in specific T cell subset compositions, among all subjects, related to Table 5-5. The differences in epigenetic measures across specific cut-points are shown for: a) AAR, b-d) EEAA, e-f) PEAA, g-h) GEAA, and i-j) aaDNAmTL. All subjects with complete T cell subset data are used (namely, the same set of subjects used in Table 5-5). Each panel shows two or more boxplots (heavy line = median, box = 25th-75th percentile, whiskers = 5th-95th percentile) which in total show all n=72 subjects at all visits; data in boxplots are defined based on T cell subset cut-point criteria displayed on x-axis titles. 38 SC and 34 SN subjects were evaluated.

REFERENCES

Ahuja, N.*, et al.* Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Res* 1998;58(23):5489-5494.

Ambatipudi, S.*, et al.* DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *Eur J Cancer* 2017;75:299-307.

Arneson, A.*, et al.* A mammalian methylation array for profiling methylation levels at conserved sequences. *Nature Communications* 2022;13(1):783.

Barratclough, A.*, et al.* Accurate Epigenetic Aging in Bottlenose Dolphins (Tursiops truncatus), an Essential Step in the Conservation of at-Risk Dolphins. In, *Journal of Zoological and Botanical Gardens*. 2021. p. 416-420.

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

Berdyshev, G.D.*, et al.* [Nucleotide composition of DNA and RNA from somatic tissues of humpback and its changes during spawning]. *Biokhimiia* 1967;32(5):988-993.

Breen, E.C.*, et al.* Accelerated aging with HIV occurs at the time of initial HIV infection. *iScience* 2022;25(7).

Breeze, C.E.*, et al.* eFORGE v2.0: updated analysis of cell type-specific signal in epigenomic data. *Bioinformatics* 2019;35(22):4767-4769.

Castillo-Mancilla, J.R.*, et al.* Suboptimal Adherence to Combination Antiretroviral Therapy Is Associated With Higher Levels of Inflammation Despite HIV Suppression. *Clinical Infectious Diseases* 2016;63(12):1661-1667.

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

Chiavellini, P.*, et al.* Hippocampal DNA Methylation, Epigenetic Age, and Spatial Memory Performance in Young and Old Rats. *J Gerontol A Biol Sci Med Sci* 2022;77(12):2387-2394.

Christensen, B.C.*, et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet* 2009;5(8):e1000602.

Cossette, M.L.*, et al.* Epigenetics and island-mainland divergence in an insectivorous small mammal. *Mol Ecol* 2023;32(1):152-166.

de Magalhães, J.P., Costa, J. and Church, G.M. An Analysis of the Relationship Between Metabolism, Developmental Schedules, and Longevity Using Phylogenetic Independent Contrasts. *The Journals of Gerontology: Series A* 2007;62(2):149--160.

Dugué, P.A.*, et al.* DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. *Int J Cancer* 2018;142(8):1611-1619.

Friedman, J.H. Multivariate Adaptive Regression Splines. *The Annals of Statistics* 1991;19(1):1-67, 67.

Friedman, J.H., Hastie, T. and Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 2010;33(1):1-22.

Giorgi, J.V.*, et al.* Quality control in the flow cytometric measurement of T-lymphocyte subsets: The Multicenter AIDS Cohort Study experience. *Clinical Immunology and Immunopathology* 1990;55(2):173-186.

Gross, Andrew M.*, et al.* Methylome-wide Analysis of Chronic HIV Infection Reveals Five-Year Increase in Biological Age and Epigenetic Targeting of HLA. *Molecular Cell* 2016;62(2):157-168.

Haghani, A.*, et al.* DNA methylation networks underlying mammalian traits. *Science* 2023;381(6658):eabq5693.

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

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

Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* 2013;14(10):3156.

Horvath, S.*, et al.* DNA methylation clocks tick in naked mole rats but queens age more slowly than nonbreeders. *Nature Aging* 2021;2(1):46-59.

Horvath, S.*, et al.* DNA methylation aging and transcriptomic studies in horses. *Nature Communications* 2022;13(1).

Horvath, S.*, et al.* Pan-primate studies of age and sex. *GeroScience* 2023;45(6):3187-3209.

Horvath, S.*, et al.* Methylation studies in Peromyscus: aging, altitude adaptation, and monogamy. *Geroscience* 2022;44(1):447-461.

Horvath, S.*, et al.* Epigenetic clock and methylation studies in marsupials: opossums, Tasmanian devils, kangaroos, and wallabies. *GeroScience* 2022;44(3):1825-1845.

Horvath, S. and Levine, A.J. HIV-1 Infection Accelerates Age According to the Epigenetic Clock. *The Journal of Infectious Diseases* 2015;212(10):1563--1573.

Horvath, S.*, et al.* DNA methylation clocks for dogs and humans. *Proceedings of the National Academy of Sciences* 2022;119(21).

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

Horvath, S. and Ritz, B.R. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. *Aging (Albany NY)* 2015;7(12):1130-1142.

Horvath, S.*, et al.* Epigenetic clock and methylation studies in the rhesus macaque. *Geroscience* 2021;43(5):2441-2453.

Horvath, S.*, et al.* DNA methylation age analysis of rapamycin in common marmosets. *Geroscience* 2021;43(5):2413-2425.

Hultin, L.E.*, et al.* Assessing immunophenotyping performance: Proficiency-validation for adopting improved flow cytometry methods. *Cytometry Part B: Clinical Cytometry* 2007;72B(4):249-255.

Jasinska, A.J.*, et al.* Epigenetic clock and methylation studies in vervet monkeys. *Geroscience* 2022;44(2):699-717.

Kaslow, R.A.*, et al.* The Multicenter Aids Cohort Study: Rationale, Organization, And Selected Characteristics Of The Participants. *American Journal of Epidemiology* 1987;126(2):310-318.

Kordowitzki, P.*, et al.* Epigenetic clock and methylation study of oocytes from a bovine model of reproductive aging. *Aging Cell* 2021;20(5):e13349.

Langfelder, P. and Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9(1):559.

Larison, B.*, et al.* Epigenetic models developed for plains zebras predict age in domestic horses and endangered equids. *Commun Biol* 2021;4(1):1412.

Lemaître, J.F.*, et al.* DNA methylation as a tool to explore ageing in wild roe deer populations. *Mol Ecol Resour* 2022;22(3):1002-1015.

Leung, J.M.*, et al.* Longitudinal study of surrogate aging measures during human immunodeficiency virus seroconversion. *Aging* 2017;9(3):687-705.

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):690-700.

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):1198-1211.

Levine, M.E.*, et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* 2018;10(4):573--591.

Lu, A.T.*, et al.* Universal DNA methylation age across mammalian tissues. *Nature Aging* 2023.

Lu, A.T.*, et al.* Genetic architecture of epigenetic and neuronal ageing rates in human brain regions. *Nat Commun* 2017;8:15353.

Lu, A.T.*, et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* 2019;11(2):303--327.

Lu, A.T.*, et al.* DNA methylation-based estimator of telomere length. *Aging* 2019;11(16):5895-5923.

Maegawa, S.*, et al.* Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res* 2010;20(3):332-340.

McLean, C.Y.*, et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* 2010;28(5):495--501.

Murtagh, F. Multidimensional Clustering Algorithms. Wuerzburg: Physica-Verlag; 1985.

Parsons, K.M.*, et al.* DNA methylation-based biomarkers for ageing long-lived cetaceans. *Molecular Ecology Resources* 2023;23(6):1241-1256.

Peters, K.J.*, et al.* An epigenetic DNA methylation clock for age estimates in Indo-Pacific bottlenose dolphins (Tursiops aduncus). *Evolutionary Applications* 2023;16(1):126-133.

Pinho, G.M.*, et al.* Hibernation slows epigenetic ageing in yellow-bellied marmots. *Nat Ecol Evol* 2022;6(4):418-426.

Prado, N.A.*, et al.* Epigenetic clock and methylation studies in elephants. *Aging Cell* 2021;20(7).

Quach, A.*, et al.* Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging* 2017;9(2):419--446.

Raj, K.*, et al.* Epigenetic clock and methylation studies in cats. *Geroscience* 2021;43(5):2363-2378.

Rakyan, V.K.*, et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Research* 2010;20(4):434-439.

Rickabaugh, T.M.*, et al.* Acceleration of Age-Associated Methylation Patterns in HIV-1-Infected Adults. *PLOS ONE* 2015;10(3):e0119201.

Robeck, T.R.*, et al.* Multi-Tissue Methylation Clocks for Age and Sex Estimation in the Common Bottlenose Dolphin. *Frontiers in Marine Science* 2021;8.

Robeck, T.R.*, et al.* Multi-species and multi-tissue methylation clocks for age estimation in toothed whales and dolphins. *Commun Biol* 2021;4(1):642.

Robeck, T.R.*, et al.* Multi-tissue DNA methylation aging clocks for sea lions, walruses and seals. *Communications Biology* 2023;6(1):359.

Schachtschneider, K.M.*, et al.* Epigenetic clock and DNA methylation analysis of porcine models of aging and obesity. *GeroScience* 2021;43(5):2467-2483.

Sehl, M.E.*, et al.* Increased Rate of Epigenetic Aging in Men Living With HIV Prior to Treatment. *Frontiers in Genetics* 2021;12.

Sehl, M.E.*, et al.* Increased Rate of Epigenetic Aging in Men Living With HIV Prior to Treatment. *Frontiers in Genetics* 2022;12.

Sehl, M.E.*, et al.* The effects of anti-retroviral therapy on epigenetic age acceleration observed in hiv-1-infected adults. *Pathogens and Immunity* 2020;5(1):291.

Shiau, S.*, et al.* Longitudinal changes in epigenetic age in youth with perinatally acquired HIV and youth who are perinatally HIV-exposed uninfected. *AIDS* 2021;35(5):811-819.

Sugrue, V.J.*, et al.* Castration delays epigenetic aging and feminizes DNA methylation at androgen-regulated loci. *Elife* 2021;10.

Teschendorff, A.E.*, et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research* 2010;20(4):440-446.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996;58(1):267-288.

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

Vu, H. and Ernst, J. Universal chromatin state annotation of the mouse genome. *Genome Biology* 2023;24(1):153.

Wada, N.I.*, et al.* The effect of HAART-induced HIV suppression on circulating markers of inflammation and immune activation. *AIDS* 2015;29(4).

Wilkinson, G.S.*, et al.* Author Correction: DNA methylation predicts age and provides insight into exceptional longevity of bats. *Nature Communications* 2021;12(1).

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

Zhou, W.*, et al.* SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Research* 2018;46(20):e123--e123.

Zoller, J.A.*, et al.* DNA methylation clocks for clawed frogs reveal evolutionary conservation of epigenetic aging. *GeroScience* 2023.

Zou, H. and Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 2005;67(2):301-320.