Genome Biology

RESEARCH

Open Access

Significant variation in the performance of DNA methylation predictors across data preprocessing and normalization strategies

Anil P. S. Ori[1*], Ake T. Lu[2], Steve Horvath[2,3] and Roel A. Ophoff[1,2,4*]

*Correspondence: anilori.contact@gmail.com; ophoff@ucla.edu

[1] Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095-176, USA
[2] Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
[3] Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA, USA
[4] Department of Psychiatry, Erasmus University Medical Center, Rotterdam, The Netherlands

Abstract

Background: DNA methylation (DNAm)-based predictors hold great promise to serve as clinical tools for health interventions and disease management. While these algorithms often have high prediction accuracy, the consistency of their performance remains to be determined. We therefore conduct a systematic evaluation across 101 different DNAm data preprocessing and normalization strategies and assess how each analytical strategy affects the consistency of 41 DNAm-based predictors.

Results: Our analyses are conducted in a large EPIC DNAm array dataset from the Jackson Heart Study ($N = 2053$) that included 146 pairs of technical replicate samples. By estimating the average absolute agreement between replicate pairs, we show that 32 out of 41 predictors (78%) demonstrate excellent consistency when appropriate data processing and normalization steps are implemented. Across all pairs of predictors, we find a moderate correlation in performance across analytical strategies (mean rho = 0.40, SD = 0.27), highlighting significant heterogeneity in performance across algorithms. Successful or unsuccessful removal of technical variation furthermore significantly impacts downstream phenotypic association analysis, such as all-cause mortality risk associations.

Conclusions: We show that DNAm-based algorithms are sensitive to technical variation. The right choice of data processing strategy is important to achieve reproducible estimates and improve prediction accuracy in downstream phenotypic association analyses. For each of the 41 DNAm predictors, we report its degree of consistency and provide the best performing analytical strategy as a guideline for the research community. As DNAm-based predictors become more and more widely used, our work helps improve their performance and standardize their implementation.

Keywords: DNA methylation, Infinium MethylationEPIC array, DNAm predictors, Consistency, Replicability, Biomarkers, Jackson Heart Study

Introduction

DNA methylation (DNAm) is a form of epigenetic regulation that is essential for human development and implicated in health and disease [1, 2]. Through advancements in biological technology, large-scale DNA methylation profiling has become more affordable and widely used. Microarray technologies now enable the simultaneous interrogation of DNAm states of more than 850,000 CpG dinucleotides across the genome, using the latest EPIC array [3]. An application of DNAm data has been in developing DNAm-based algorithms to predict health-related phenotypes, including blood cell type proportions [4, 5], aging [6–13], all-cause mortality risk [14–17], cancer risk [18, 19], body mass index (BMI), and smoking signatures [20], among others. These molecular predictors have great potential for clinical applications. A thorough and systematic investigation of their performance has however not been conducted so far.

Unlike the genome, the DNA methylome is of dynamic nature and largely explained by non-shared individual environments [21]. Like other high-throughput molecular data, DNAm can furthermore be impacted by variation in laboratory conditions, sample handling, reagents, and/or equipment used [22]. Technical variation is often widespread and tackling such effects is of critical importance to study biological variation in any -omic analysis, including DNAm. Over the years, a plethora of methods has been developed to identify and remove unwanted technical variations from DNAm data [23–29]. Previous studies have investigated the impact of specific methods on outcomes of DNAm analysis and demonstrated the importance of correcting for probe design type, batch effects, and hidden confounders while the effect of different normalization strategies gave mixed results [30–33]. A systematic and unbiased evaluation of commonly used data preprocessing and normalization strategies of DNAm data for the application of DNAm-based predictors has however not yet been conducted. DNAm is an important tool to study health and disease and understanding how analytical strategies impact algorithm performance is critical for method standardization and implementation for both research and clinical purposes.

Here, we performed a comprehensive investigation of 41 DNAm predictors and evaluated algorithm performance by measuring their consistency across 101 data preprocessing and normalization strategies in the Jackson Heart Study (JHS) [34]. The JHS has collected a large sample of 850 K EPIC DNAm arrays in blood that includes 146 pairs of technical replicates. These replicates represent identical DNA samples that were assayed twice at independent time points. The agreement in DNAm predictor estimate between technical replicates after data preprocessing and normalization allowed us to quantify the degree to which an analytical strategy can successfully remove unwanted technical variation. For each predictor, we report the analytic strategy that yields the most consistent estimates and demonstrate how reducing technical variation is critical for optimal algorithm performance in downstream phenotypic analyses. Our work emphasizes the importance of data processing and normalization of DNAm data and provides best practices to optimize the performance and consistency of DNAm predictors.

Results

To evaluate how unwanted technical variation in DNAm data impacts the performance of DNAm-based predictors, we implemented 101 data processing and normalization strategies in the JHS dataset. For each analytical strategy, which we will refer to as a

"pipeline", we then extracted beta values and calculated estimates of 41 DNAm-based predictors in (1) JHS data 1: a sample of 146 technical replicate pairs and (2) JHS data 2: a general sample of 1761 non-replicate samples that do not overlap with the individuals in the replicate dataset. Figure 1 shows an overview of our analysis plan. In the sample of technical replicates, we quantified the average absolute agreement between replicate pair values (i.e., consistency) by means of the ICC for each DNAm predictor and each pipeline separately (41 predictors × 101 pipelines = 4141 ICC analysis). We also generated DNAm estimates in the general sample. This allowed us to correlate the ICC of a pipeline that was estimated in the sample of replicates with predictor estimates in the independent general JHS sample.

We calculated the ICC estimates derived from a two-way random effect model to assess the consistency of each predictor for each data processing pipeline. The ICC is a zero to one estimate that quantifies the average absolute agreement across technical replicate pairs that were processed at a different occasion. We also calculated five other types of ICCs and found high concordance between the different ICC measures (mean rho = 0.99, SD = 0.01, see Additional file 1: Fig. S1). All ICC statistics for each DNAm predictor and pipeline are reported in Additional file 2. In the remainder of the paper, we will refer to ICC(2,1) as ICC, unless stated otherwise. The distribution of the ICC across pipelines for each predictor is shown in Additional file 1: Fig. S2.

Most DNAm-based predictors yield high consistency when the best analytical pipeline is implemented

Table 1 shows all 41 DNAm predictors alongside general information on each algorithm and corresponding ICC statistics, including the data processing and normalization pipeline that yielded the highest agreement between replicate pairs for each predictor. Across all predictors and pipelines ($N = 4141$), we observed a significant degree of similarity between replicates (all ICC P-values < 0.05/4,141). The median across all ICC estimates is 0.93 with a range of 0.22–0.99.

The GrimAge predictor reports the highest consistency (ICC = 0.994, $P = 6.6e-144$), followed by ZhangAge (ICC = 0.992, $P = 8.4e-132$), and TIMP_1 (ICC = 0.992, $P = 8.5e-133$). In fact, 32 out of 41 predictors (78%) reach an ICC > 0.9 with at least

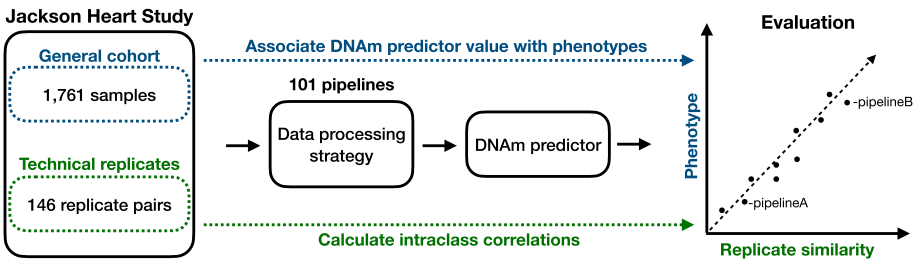

Fig. 1 Schematic overview of the analysis plan to evaluate DNAm algorithm performance. DNAm analyses are conducted using DNAm EPIC array samples in JHS. JHS includes a significant number of technical replicate pairs thereby allowing for a careful investigation of how the removal of unwanted technical variation impacts DNAm algorithm performance across 101 data processing pipelines. JHS has also collected information on disease-related phenotypes, including mortality status after follow-up. This allowed us to assess how the removal of technical variation in DNAm predictor estimates by a data processing pipeline impacts downstream phenotypic association analyses

Table 1 Overview of predictor consistency and best performing data processing pipelines

Predictor information				Reliability (ICC statistics)			
Name	**Phenotype**	**Array**	**Probes**	**Median**	**Min**	**Max**	**Best analytical pipeline**
GrimAge [15]	Mortality	EPIC/450 K	1030	0.990	0.921	0.994	ENmix: bg = oob, dye = mean, norm = q3, probe = rcp
ZhangAge [8]	Chronological age	EPIC/450 K	514	0.991	0.987	0.992	ENmix: bg = neg, dye = mean, norm = q2, probe = rcp
TIMP_1 [15]	TIMP-1 serum protein	EPIC/450 K	42	0.988	0.973	0.992	ENmix: bg = oob, dye = relic, norm = q2, probe = rcp
Bcell [5]	B-lymphocyte cell fraction	EPIC	50	0.980	0.881	0.988	Minfi: no bg correction with control normalization
Neu [5]	Neutrophil cell fraction	EPIC	50	0.984	0.973	0.987	ENmix: bg = oob, dye = mean, norm = q2, probe = rcp
B2M [15]	B2M serum protein	EPIC/450 K	91	0.973	0.759	0.985	ENmix: bg = oob, dye = relic, norm = q1, probe = rcp
SkinBloodAge [9]	Chronological age	EPIC/450 K	391	0.979	0.908	0.982	ENmix: bg = neg, dye = relic, norm = q1, probe = rcp
Smoking_Lu [15]	Smoking pack years	EPIC/450 K	172	0.971	0.889	0.981	ENmix: bg = oob, dye = no, norm = no, probe = rcp
Smoking_McCartney [20]	Smoking pack years	EPIC	233	0.975	0.942	0.979	Minfi: noob with dye correction
HannumAge [7]	Chronological age	450 K	71	0.972	0.834	0.978	ENmix: bg = est, dye = relic, norm = no, probe = rcp
CD8T [5]	CD8 + T-cell fraction	EPIC	50	0.969	0.881	0.978	ENmix: bg = neg, dye = mean, norm = q1, probe = rcp
NK [5]	Natural killer cell fraction	EPIC	50	0.952	0.883	0.977	ENmix: bg = neg, dye = relic, norm = q3, probe = rcp
BioAge4HAStatic [17]	Chronological age	450 K	-	0.966	0.826	0.975	ENmix: bg = oob, dye = relic, norm = no, probe = rcp
Cystatin_C [15]	Cystatin-C serum protein	EPIC/450 K	87	0.954	0.829	0.973	ENmix: bg = oob, dye = no, norm = q2, probe = rcp
PhenoAge [14]	Mortality	EPIC/450 K/27 K	513	0.954	0.926	0.97	ENmix: bg = neg, dye = relic, norm = q1, probe = rcp

Table 1 (continued)

Predictor information				Reliability (ICC statistics)			
Name	**Phenotype**	**Array**	**Probes**	**Median**	**Min**	**Max**	**Best analytical pipeline**
Mono [5]	Monocyte cell fraction	EPIC	50	0.953	0.865	0.968	Minfi: illumine bg correction with control normalization
DNAmTL [12]	Telomere length	EPIC/450 K	140	0.952	0.912	0.965	ENmix: bg = oob, dye = relic, norm = q1, probe = rcp
HorvathAge [6]	Chronological age	450 K/27 K	353	0.950	0.867	0.964	WateRmelon: naten
CD4T [5]	CD4 + T-cell fraction	EPIC	50	0.959	0.951	0.964	ENmix: bg = neg, dye = no, norm = no, probe = rcp
epiTOC [18]	Mitotic divisions	450 K	385	0.911	0.498	0.962	ENmix: bg = oob, dye = mean, norm = q2, probe = rcp
Leptin [15]	Leptin serum protein	EPIC/450 K	187	0.896	0.447	0.953	ENmix: bg = oob, dye = relic, norm = q3, probe = rcp
VidalBraloAge [13]	Chronological age	27 K	8	0.945	0.922	0.952	ENmix: bg = neg, dye = mean, norm = no, probe = rcp
MiAge [19]	Mitotic divisions	450 K	268	0.884	0.348	0.947	WateRmelon: nanes
LinAge [10]	Chronological age	450 K	99	0.930	0.878	0.939	ENmix: bg = est, dye = relic, norm = no, probe = no_rcp
ADM [15]	ADM serum protein	EPIC/450 K	186	0.900	0.756	0.938	ENmix: bg = neg, dye = mean, norm = q3, probe = rcp
WHR [20]	Waist-to-hip ratio	EPIC	226	0.878	0.634	0.925	ENmix: bg = oob, dye = relic, norm = q2, probe = rcp
ZhangMortality [16]	Mortality	450 K	10	0.877	0.807	0.92	Minfi: no bg correction with control normalization
BodyFat [20]	Body fat	EPIC	968	0.893	0.843	0.918	ENmix: bg = est, dye = relic, norm = no, probe = rcp
Cholesterol [20]	Total cholesterol	EPIC	204	0.888	0.762	0.917	ENmix: bg = oob, dye = no, norm = q2, probe = rcp
BMI [20]	BMI	EPIC	1109	0.904	0.877	0.914	ENmix: bg = neg, dye = mean, norm = no, probe = rcp
GDF_15 [20]	GDF-15 serum protein	EPIC/450 K	137	0.819	0.502	0.903	ENmix: bg = est, dye = mean, norm = q1, probe = rcp

Table 1 (continued)

Predictor information				Reliability (ICC statistics)			
Name	**Phenotype**	**Array**	**Probes**	**Median**	**Min**	**Max**	**Best analytical pipeline**
LDL [20]	LDL	EPIC	233	0.846	0.732	0.901	ENmix: bg = oob, dye = relic, norm = no, probe = rcp
HDLratio [20]	Total to HDL cholesterol ratio	EPIC	412	0.848	0.643	0.890	ENmix: bg = oob, dye = relic, norm = q1, probe = rcp
Alcohol [20]	Alcohol	EPIC	450	0.807	0.551	0.878	ENmix: bg = neg, dye = relic, norm = no, probe = rcp
WeidnerAge [11]	Chronological age	27 K	3	0.826	0.583	0.865	ENmix: bg = neg, dye = relic, norm = no, probe = rcp
Education [20]	Educational attainment	EPIC	373	0.774	0.506	0.865	Cross: noob with dye correction + BMIQ
HDL [20]	HDL cholesterol	EPIC	737	0.835	0.694	0.853	ENmix: bg = est, dye = relic, norm = q1, probe = rcp
CD8pCD-28nCD45Ran [6]	Specific T-cell fraction	27 K	-	0.814	0.756	0.845	ENmix: bg = oob, dye = relic, norm = no, probe = rcp
PlasmaBlast [6]	Plasma B cell fraction	27 K	-	0.718	0.638	0.840	Cross: noob with dye correction + BMIQ
PAI_1 [15]	PAI-1 serum protein	EPIC/450 K	211	0.744	0.22	0.838	ENmix: bg = neg, dye = relic, norm = q3, probe = rcp
CD8naive [6]	CD8 T-cell fraction	27 K	-	0.777	0.659	0.830	WateRmelon: danen

Shown is general information on each DNAm-based predictor alongside their corresponding ICC statistics. The name of the predictor, the phenotype it is trained on, the array platform it can be applied on, and the number of predictor probes (if available) are listed on the left side of the table. ICC statistics are listed on the right side of the table. The ICC quantifies the degree of absolute agreement between estimator values of a pair of technical replicates. For each predictor, across 101 pipelines, the median, minimum, and maximum ICC are listed. Predictors are ranked by the maximum ICC. The final column reports methodological details of the best performing data processing pipelines (i.e., the pipeline with the highest consistency). *Bg* background correction, *dye* dye-bias correction, *norm* normalization method, *probe* probe-type bias correction. Full details on analytical pipelines and how they were implemented are available in Additional file 4.

one data processing pipeline. The predictors with higher ICCs have more narrow ICC distributions than predictors with lower ICCs (Additional file 1: Fig. S2), suggesting that predictors with higher consistency are more robust to the choice of data processing pipelines. The predictors with the lowest consistency are CD8pCD28nCD45RAn (ICC = 0.85, $P = 1.63e-41$), PlasmaBlast (ICC = 0.84, $P = 7.19e-52$), PAI-1 (ICC = 0.84, $P = 2.80e-40$), and CD8_naive (ICC = 0.83, $P = 1.17e-39$).

Across pipelines and predictors ($N = 4141$), the ENmix package yielded higher consistency (median ICC = 0.93, range = 0.61–0.99) than the minfi (median ICC = 0.91, range = 0.22–0.99) and wateRmelon (median ICC = 0.91, range = 0.49–0.99) packages.

Among the best performing pipeline for each of the 41 DNAm predictors, i.e., achieving the highest consistency, 32 (78%), 4 (10%), and 3 (7%) predictors were from the ENmix, minfi, and wateRmelon package, respectively. Among ENmix pipelines, out-of-band (OOB) background estimation (15 out of 32), REgression on Logarithm of Internal Control probes (RELIC) dye-bias correction (19 out of 32), no quantile normalization (12 out of 32), and the Regression on Correlated Probes (RCP) probe-type bias correction (31 out of 32) yielded the highest consistency most often (Additional file 1: Fig. S3). Two ENmix pipelines achieved the highest consistency for three predictors. The analytical pipeline that included OOB background estimation, RELIC dye-bias correction, no normalization, and RCP probe-type bias correction (i.e., "ENmix:oob_relic_nonorm_rcp") performed best for the BioAge4HAStatic, LDL, and CD8pCD28nCD45RAn predictors. The pipeline that included OOB background estimation, RELIC dye-bias correction, quantile normalization, and RCP probe-type bias correction (i.e., "ENmix: oob_relic_q1_rcp") performed best for the B2M, DNAmTL, and HDLratio predictors.

Best performing analytical pipelines are less impacted by batch effects

Next, we assessed how corrections for batch effects impacted our measures of consistency across predictors across analytical pipelines. We included four covariates with potential batch effects in our analysis, i.e., array ID, array position, sample plate, and sample well. Such data was available for 1888 samples in the full JacksonHeart dataset. Across pipelines and predictors ($N=4141$), batch effects collectively explained a median of 4.45% (Q1–Q3 = 1.48–7.20%) of the variance in predictor estimates. Batch effects explained less variance in output estimates when the best performing pipelines were applied (median = 3.37%, Q1–Q3 = 1.35–7.23%) compared to the worst performing pipelines (median = 5.80%, Q1–Q3 = 3.08–9.43%). This indicates that analytical pipelines that yield a higher agreement between technical replicates are more successful in removing technical variation introduced by batch effects. This is further emphasized by a negative correlation between the variance explained by batch effects and the calculated ICC across predictors and pipelines ($N=4141$, rho = -0.05, $P=3.7e-04$). This relationship is stronger in the worst performing pipelines ($N=41$, rho = -0.37, $P=0.02$) and not present within the best performing pipelines ($N=41$, rho = 0.09, $P=0.58$). After regressing out batch effects from the predictor estimates, we found a strong correlation with the consistency measured without correcting for batch effects ($N=4141$, rho = 0.81, $P<2.2e-16$, see Additional file 1: Fig. S4). Correction for batch effects on average produced lower consistency (median ICC = 0.81, Q1–Q3 = 0.76–0.84) compared to unadjusted predictor estimates (median ICC = 0.96, Q1–Q3 = 0.91–0.98).

There is significant heterogeneity in pipeline performance across predictors

Among the 41 best performing pipelines (i.e., the pipeline with the largest ICC value for each of the 41 predictors), there are 27 different data processing and normalization strategies, which highlights significant heterogeneity in the choice of best pipeline between predictors. As ICC differences between pipelines of a predictor can be small and pipelines beyond the highest ICC may also be informative, we calculated the median rank across the 41 predictors for each of the 101 pipelines (Additional file 3). The pipeline with the best median rank (at 15) across predictors is the "ENmix: oob_relic_q1_rcp."

While this observation suggests this pipeline yields the best average performance across predictors, it still scored average to low for multiple predictors. For example, for the BMI predictor, the "ENmix: oob_relic_q1_rcp" pipeline had one of the lowest ranks (ICC = 0.89, rank = 91). It is also important to note that a data processing pipeline can also introduce more spurious variation instead of removing technical variation. That is, the raw data pipeline that does not apply any data processing and normalization yielded a median rank of 85 (range: 7 to 100). For the CD4T and CD8 naive predictors, the raw data pipeline ranked as the seventh best performing pipeline highlighting that most pipelines perform worse than no data processing at all for these two predictors. The "Minfi: raw_quantile_strat" and "Minfi: illumina_bg_quantile_strat" had the lowest median rank of 100 and yielded the lowest consistency for 17 and 9 predictors, respectively (Additional file 3).

To assess the concordance in pipeline performance across predictors more formally, we calculated the rank correlation in pipeline consistency between all pairs of predictors. In Fig. 2, we visualize the result of this analysis via a clustered correlation heatmap.

For some predictors, the ranking in pipeline performance is very similar. For example, the GrimAge, Smoking_Lu, Cystatin_C, and GDF_15 predictors show strong concordance (mean rho = 0.92). As noted, these four predictors were developed in the same dataset and the Cystatin_C, GDF_15, and Smoking_Lu estimates are included in the GrimAge algorithm. Across all pairs of predictors, we find a moderate correlation in pipeline performance (mean rho = 0.40, SD = 0.27). Some predictors however show little to no concordance with other predictors. The ranking of pipelines of the BMI and NK predictor, for example, have a mean rank correlation of 0.14 (SD = 0.20) and 0.21 (SD = 0.24), respectively, with that of other predictors. For a handful of predictor pairs, we even observe a negative correlation, suggesting that pipelines that yield high consistency for one predictor yield low consistency for another. Pipeline performance of the BioAge4HAStatic and Mono predictors for example has a correlation of − 0.45 ($P = 2.1e-06$). Our findings thus far show that specific pipelines are more effective in removing unwanted technical variation for a predictor and that significant heterogeneity exists in pipeline performance across predictors.

The choice of data processing pipeline impacts the downstream analysis of predictors

Next, we evaluated if the performance of a pipeline can also affect downstream phenotypic analyses of a predictor. For these analyses, we used the general JHS data 2 sample. For each pipeline, we calculated the mean and standard deviation (SD) of the predictor estimate distribution in the general JHS sample. For each predictor, we then correlated these two statistics (i.e., the mean and SD) with the ICC estimates of the pipelines obtained in the technical replicate sample. We find that the choice of the pipeline has a significant impact on the distribution of the predictor estimate. Of the 41 predictors, 33 (80%) are significantly impacted on the distribution of their estimates after Bonferroni correction ($P < 0.0012$). For 22 predictors (54%), we find a significant correlation for both the mean and standard deviation. For DNAmTL, we, for example, observe a negative correlation between the performance of a pipeline and the mean of the estimate distribution (rho = − 0.71, $P < 2.2e-16$) and a positive correlation with the standard deviation of the estimate distribution (rho = 0.79, $P < 2.2e-16$). The best performing pipeline

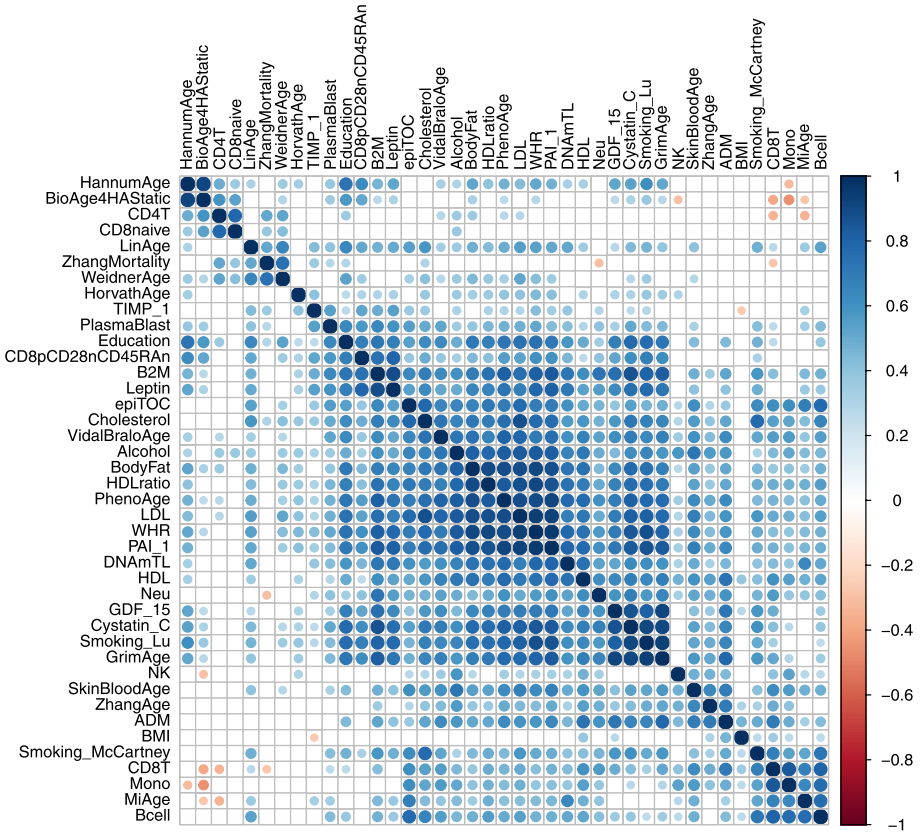

Fig. 2 DNAm predictors have a moderate degree of concordance in performance between pipelines. Shown is a clustered correlation heatmap of consistency across pipelines that visualizes the heterogeneity in pipeline performance between predictors. The color coding depicts Spearman's rho and clustering is performed using hierarchical clustering. Only correlations with a *P*-value < 0.01 are colored

yields a mean estimate of 6.83 kilobases (SD = 0.34). The least performing pipeline yields a mean estimate of 7.20 kilobases (SD = 0.29). This shows that the more effective a pipeline is in removing technical variation, the lower the DNAm-based predicted estimate of telomere length and the larger the variation between individuals. The direction of effect of the relationship between pipeline performance and the mean and standard deviation of the DNAm variables varies between predictors as well. HorvathAge, for example, is impacted on its standard deviation (rho = 0.39, $P = 5.6e-05$) but not on the mean (rho = −0.10, $P = 0.27$). HDLratio is impacted on its mean but unlike DNAmTL shows a positive correlation with pipeline performance (rho = 0.38, $P = 9.8e-05$). HDLratio is not impacted on the standard deviation of its distribution (rho = 0.00, $P = 0.96$). Correlation plots and correlation statistics of all predictors are shown in Additional file 5. A full overview of test statistics can be found in Additional file 6.

Several DNAm age predictors are known to predict all-cause mortality risk. We therefore examined if pipeline performance also impacts their association with mortality risk. We focus on four predictors: HorvathAge, PhenoAge, GrimAge, and ZhangAge. Each predictor has different training characteristics and captures a different aspect of biological age and/or mortality risk [35]. ZhangAge is a blood-based DNAm clock and was developed on the largest training dataset and shown not to be associated with mortality

risk despite its improved precision [8]. We find that pipeline performance significantly impacts downstream analysis for all four predictors (Fig. 3).

For HorvathAge, pipelines that achieve greater consistency also achieve a greater correlation between HorvathAge and chronological age (rho = 0.47, $P = 1.8e06$). Better performing pipelines furthermore achieve greater power to predict all-cause mortality (rho = 0.52, $P = 3.3e-08$). For PhenoAge, we did not find an effect on the correlation with chronological age but did find the survival analysis to be significantly impacted. Better performing pipelines achieve greater power for PhenoAge (rho = 0.68, $P < 2.2e-16$) but also a smaller hazard ratio (rho = -0.39, $P = 5.3e-05$), suggesting that unsuccessful removal of technical variation in DNAm data can inflate the magnitude of mortality

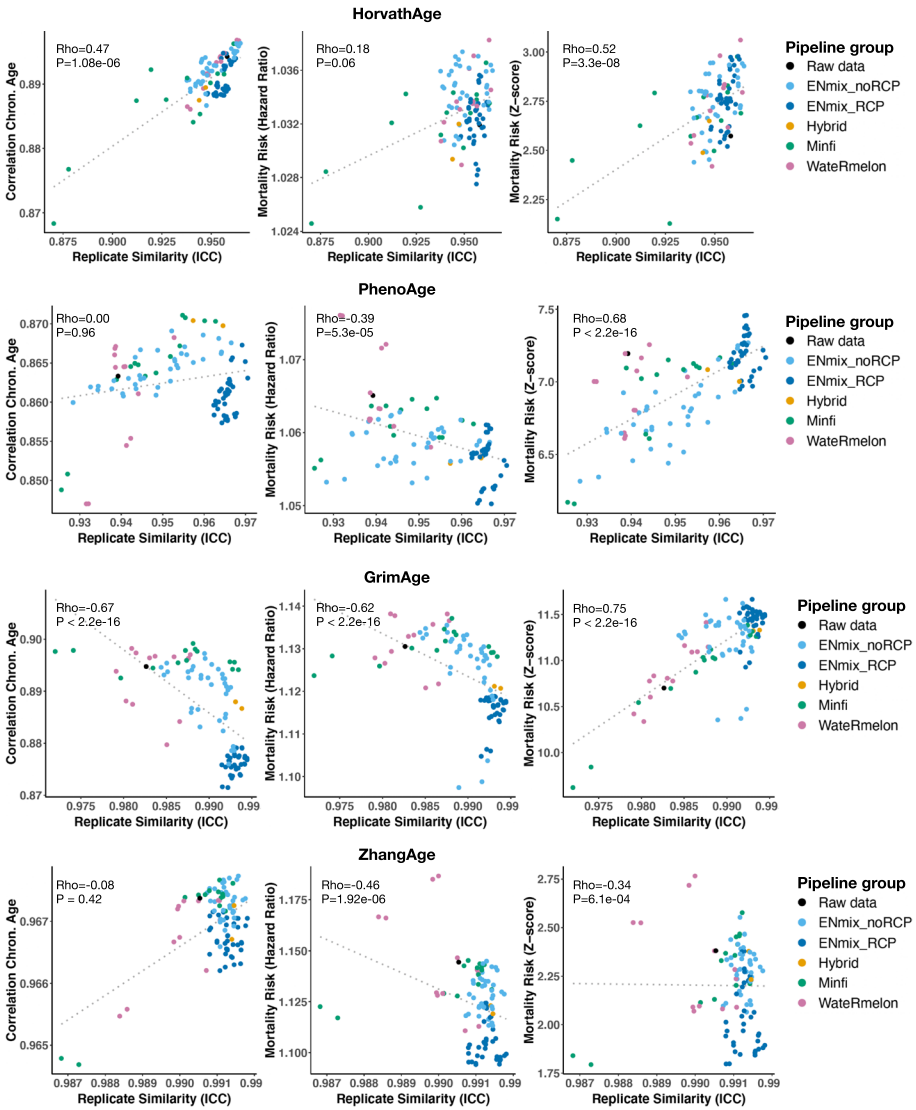

Fig. 3 Pipeline performance impacts downstream analyses of DNAm age predictors. Shown are the association between pipeline ICC and the correlation with chronological age (left panels), the hazard ratio of mortality risk prediction (middle panel), and the *z*-score of the mortality risk prediction (right panels) for HorvathAge (top row), PhenoAge (2nd row), GrimAge (3rd row), and ZhangAge (bottom row). Pipelines are color-coded by package/method. Spearman rank correlation statistics are shown in the top left corners

risk. In contrast to our findings for HorvathAge, we found that better performing pipelines produced a lower correlation with chronological age for GrimAge (rho = −0.67, $P < 2.2e-16$). Similar to PhenoAge, we found that pipelines that achieve greater consistency yield more significant associations with mortality for GrimAge (rho = 0.75, $P < 2.2-16$) but also a smaller hazard ratio (rho = −0.62, $P < 2.2e-16$). The most reliable pipeline reports a significant hazard ratio of 1.12 (SE = 0.01, $P = 1.60e-30$), which verifies GrimAge as a strong predictor of all-cause mortality, especially when spurious technical variation is appropriately accounted for. For ZhangAge, we found no impact on the correlation with chronological age. Better performing pipelines produced smaller and less significant effects in associations with all-cause mortality. The most reliable pipeline produced a non-significant hazard ratio of 1.10 (SE = 0.05, $P = 0.06$), confirming that ZhangAge does not predict mortality risk. Taken together, using the general JHS sample, we demonstrate how pipeline performance has a significant impact on the downstream phenotypic analysis of DNAm predictors.

Predictor consistency is inversely associated with the sample size of the training dataset

To assess if specific features of the predictors are associated with higher consistency, we investigated the number of CpG probes and the sample size of the training dataset in relation to the ICC of the best performing pipeline (Additional file 1: Fig. S5). Using predictors for which such information was available, we find that the sample size of the dataset in which a predictor was developed is inversely associated with the observed degree of predictor consistency ($N = 37$, rho = −0.39, $P = 0.02$). We did not find a significant association between the number of predictor CpG probes and the consistency of a predictor ($N = 37$, rho = −0.21, $P = 0.20$).

A smaller number of replicate pairs can be used to measure consistency

In our analyses, we made use of a large number of replicate pairs. We therefore assessed how sample size affected our measure of consistency and if a smaller number of replicate pairs yield similar findings. Across reliabilities from all pipelines and predictors, we observe good concordance (rho > 0.94) with as low as ten replicate pairs compared with measures obtained from larger sample sizes (Additional file 1: Fig. S6). Differences however exist between predictors with some predictors still requiring a larger number of replicate pairs (Additional file 7).

Discussion

DNAm-based predictors are emerging as powerful new methods to study health and disease, but little is known about the consistency and replicability of the estimates they produce. To investigate their performance, we carried out a systematic evaluation of 41 predictors across 101 data processing and normalization strategies and assessed to what degree algorithm performance is impacted by (un)successful removal of technical variation. Leveraging a large technical replicate sample in the JHS, we demonstrate that the choice of the analytical pipeline has a significant impact on the consistency of predictors as well as on the outcomes of downstream phenotypic analyses. We highlight that specific pipelines are more effective in removing unwanted technical variation for a predictor but that significant heterogeneity exists in pipeline performance across predictors.

Pipelines of the ENmix package achieved the highest consistency and were most frequently represented among the best performing pipelines. As research on DNAm-based predictors will continue to grow, our work provides best practices for the research community to help standardize their implementation and improve their performance.

To quantify method performance, we used a type of intraclass correlation that measures consistency by assessing the degree of absolute similarity between technical replicate pairs. Guidelines from reliability research suggest that ICC values less than 0.5 are indicative of poor consistency, values between 0.5 and 0.75 indicate moderate consistency, values between 0.75 and 0.9 indicate good consistency, and values greater than 0.90 indicate excellent consistency [36]. The ICC range of best performing pipelines across predictors was 0.83–0.99, indicating good to excellent consistency for these predictors. For 32 out of 41 predictors (78%), we found excellent consistency (ICC > 0.9) for at least one data processing pipeline. Several predictors show a degree of consistency close to 1, which demonstrates that repeated collections of DNAm data yield almost the same predictor estimate and highlights their potential as a biomarker for health-related outcomes. Among predictors with high consistency are predictors of mortality risk, smoking behavior, blood cell types, and cancer risk. Demonstrating internal validity for these DNAm tools is important for research purposes but even more so for their potential utilization for health management and disease prediction in the clinic. GrimAge, a strong predictor of all-cause mortality, for example, showed the greatest agreement between replicates with an ICC of 0.994. This finding demonstrates excellent consistency based on technical replicates from the same biological sample. It remains an open question if the measured consistency translates to repeated measures of DNA samples extracted from different blood draws at the same time point or across time points. The analytical framework we applied can however be easily extended to study design of other types of (biological) replicates. Establishing method consistency and replicability in other contexts of technical and biological variation is an important next step for future research.

We found that the choice of the analytical pipeline is essential as multiple data processing strategies produced poor consistency (ICC < 0.5) for several predictors. For some predictors, like for CD4T and CD8 naive T cells, using the raw data achieves higher consistency than most data processing pipelines. This highlights that analytical decisions on how to best prepare DNAm data require careful consideration as certain data processing and normalization steps can even reduce algorithm performance. Among the best performing pipelines of each predictor, we found significant heterogeneity across predictors. That is, there are 27 unique pipelines across the 41 predictors. On average, pipelines of the Enmix package achieved the highest consistency most frequently. While there is no one optimal pipeline to use for all predictors, several data processing steps stand out as producing high consistency for multiple predictors. For example, almost half of the best performing pipelines make use of the RELIC dye-bias correction method. RELIC uses the information between pairs of internal normalization control probes to correct for differences between color channels that measure intensity levels of the array [28]. The EPIC array contains 85 pairs of controls that target the same DNA region in housekeeping genes and contain no underlying CpG sites. RELIC uses the relationship between the pairs of controls to correct for dye bias on intensity values for the whole array. Another data processing step that produced high consistency is the RCP

probe-type bias correction method. Thirty-one out of 41 of the best performing pipelines make use of this data processing step. RCP uses the existing correlation between pairs of nearby type I and II probes to adjust the beta values of all type II probes [27]. Both RELIC and RCP have been shown to reduce technical variation in DNAm data and are implemented in the ENmix package.

When only one processing strategy is desired, we recommend using the following stepwise sequence of ENmix methods: OOB background correction, RELIC dye-bias correction, quantile normalization applied separately for methylated and unmethylated intensities of Infinium I and II probes, and RCP to correct for probe design type bias. This analytical pipeline achieved the best median rank across predictors. We further recommend careful quality control whether predictor estimates in downstream analyses are not dependent on (un)successful removal of technical variation, ideally with the use of replicates. In case no replicates are available, users can use the ICC values of each analytical pipeline reported in this study and assess if these ICC values correlate at all with their outcome of interest across the different processing strategies. While this strategy is suboptimal compared to including study-specific replicates, it can help assess if technical variation impacts the analyses if one finds a significant correlation between the ICC value calculated in our study and the estimated outcome measure in their study across analytical pipelines (like we show in Fig. 3). Finally, as there is significant heterogeneity in the best performing analytical pipeline between predictors, we recommend evaluating multiple data preprocessing and normalization strategies when comparing the performance of DNAm predictors, ideally using the pipelines that yield their best performance, as reported in Table 1.

We found that the best performing analytical pipelines are less impacted by batch effects, confirming that technical variation is more successfully removed, compared to pipelines that yielded lower consistency. We furthermore recommend careful consideration when adjusting for batch effects as it can result in biologically meaningful information being lost. In our analyses, we found that measures of consistency were on average lower when batch effects were corrected for by regressing out the effect of array chip, array position, and laboratory plate compared to the unadjusted predictor estimates. Batch correction methods that can consider a trade-off between batch noise and signal preservation may help users if such crude statistical adjustments are needed [37]. Development of new data preprocessing and normalization methods that improve our ability to remove technical variation from DNAm estimates is an important avenue to pursue as well.

The choice of the analytical pipeline does not only impact the consistency of a predictor but also significantly affects downstream phenotypic analyses. We show that 80% of predictors are impacted on the mean and/or standard deviation of their distribution in the general JHS cohort. We furthermore analyzed DNAm clocks and showed that the strength of the correlation between DNAm age and chronological age is affected in opposite directions for HorvathAge and GrimAge. While the correlation with chronological age becomes stronger with better performing pipelines for HorvathAge, the correlation becomes weaker for GrimAge. For DNAm clocks that are shown to be associated with mortality risk, successful removal of technical variation produced smaller hazard ratios but more significant associations. This highlights that not appropriately

accounting for technical variation can decrease statistical power and inflate risk estimates for these predictors. It also shows that despite the narrow distribution of the ICC for these predictors, for example GrimAge has an ICC range of 0.921–0.994 indicating excellent consistency across all pipelines, the choice of the pipeline still impacts downstream association analyses. We note that in our association analysis with mortality risk, we adjusted for chronological age, and still found that the choice of pipeline influences the outcome of the analysis. This is different from the findings of a previous study that reported that the choice of analytical pipeline influences the mean of DNAm age but not the DNAm age acceleration residual [38]. This study however only compared three data processing and normalization strategies and could have missed this effect as it did not perform a systematic evaluation across many pipelines. Finally, we confirm that ZhangAge, a DNAm clock developed in the largest blood-based DNAm dataset, does not associate with mortality risk.

We also investigated if specific characteristics of a predictor impacted their consistency. We found that the sample size of the training dataset has a moderate inverse relationship with the consistency of a predictor. This suggests that predictors developed in larger training datasets are more sensitive to technical variation than predictors developed in a smaller dataset. This relationship could for example arise if larger training datasets on average have more technical factors that are not properly accounted for. The ZhangAge predictor, however, was developed in the largest training dataset and shows the second to highest consistency of all predictors we investigated. This indicates that other factors in addition to the sample size of the training dataset are likely to play a role as well. ZhangAge was developed using 65 training sets across 14 cohorts, where each training set had a certain number (ranging between 1 and 13) of cohorts randomly sampled from the 14 cohorts [8]. This strategy is, as far as we know, unique to this predictor and may have helped select for CpG probes that are less impacted by technical variation due to its many training sets of different randomly assigned cohort compositions. As training datasets with large sample sizes are essential to developing more accurate DNA-based predictors, a strategy to randomize the potential effect of technical factors, like was implemented for the development of ZhangAge, could be worthwhile to consider for new predictors as well. We did not find a significant relationship between the number of CpG probes and the observed consistency of a predictor.

Our study comes with limitations. First, we measured consistency using technical replicates in one study. A different cohort or different types of repeated measures, for example biological replicates, may yield different outcomes. Ideally, one would use study-specific replicate samples and assess if similar best practices are achieved or if alternative strategies are more appropriate to remove technical variation most optimally for that specific study. If future studies have the means to include replicate samples, they should aim to include at least ten replicate pairs. We determined that for most predictors a sample size of ten replicate pairs can already provide meaningful insights into their consistency and replicability. Second, several predictors were not fully compatible with the EPIC array platform. Predictors that were developed on older DNAm array platforms showed lower consistency. Missing probes could have affected the outcome of our analysis. Having said that, as the older 27 K and 450 K DNAm array platforms are discontinued, any future application of predictors that are not fully compatible with the

EPIC array will face a similar challenge. Third, we included all DNAm probes of predictors in our analyses and did not assess how accounting for data quality of probes impacts measures of consistency. Overall predictor probes were of good quality in the Jackson Heart Study sample. Future work should however investigate this further, for example by using imputation-based methods to impute probes that need to be excluded because of lower data quality or bad mapping, as not all datasets will have good-quality data. Fourth, the ICC is limited in that it focuses on the reduction of inter-pair variance and does not track the loss of possibly informative variance. Future studies may evaluate the normalization pipelines with respect to predictive accuracy for morbidity risk as well. Finally, we assessed the impact of (un)successful removal of unwanted technical variation on downstream phenotypic analyses of DNAm clocks and mortality risk, which may yield different results for other phenotypes that were not measured in JHS.

Conclusions

In summary, this study demonstrates that considerable variation exists in the performance of DNAm-based predictors depending on the data processing and normalization strategy implemented. Analytical pipelines that best remove unwanted technical variation in DNAm data achieve excellent consistency for most predictors thereby demonstrating their potential as biomarkers for health-related outcomes. DNAm is an important tool to study health and disease. As the number of DNAm predictors continues to rise, understanding how best to improve and implement these algorithms will be essential for downstream clinical applications.

Methods

Cohort descriptions

The Jackson Heart Study is a large observational study of African American individuals from the Jackson, Mississippi (USA), metropolitan area [34]. JHS seeks to study the causes and disparities in cardiovascular health and related phenotypes in African Americans. Data and biological materials have been collected from 5306 participants. For a subset of the cohort, peripheral blood samples were collected at baseline and subsequently used to quantify DNA methylation using the Illumina Infinium MethylationEPIC BeadChip that covers over 850,000 CpG sites. These samples have been included in previous DNAm studies [15, 39]. See Additional file 8 for cohort characteristics. In our analysis, we included individuals for which DNAm data, phenotypic variables, and mortality data were available ($N = 1909$, 62.2% women, mean (SD) of age = 56.1 (12.4) years). For 146 individuals, technical replicates were collected. We therefore divided this dataset into two samples: (1) a general cohort sample that does not include technical replicate pairs ($N = 1761$, 62.6% women, mean (SD) of age = 56.0 (12.3)) and (2) a technical replicate sample ($N = 146$, 57.5% women, mean (SD) of age = 57.4 (14.0)). Replicate pairs represent DNAm samples that were assayed twice using the EPIC array at separate occasions but originate from the same DNA extraction sample.

Data preprocessing and normalization strategies

To perform a systematic evaluation of available data preprocessing and normalization strategies, we incorporated all methods that are available through the commonly used

R packages minfi [40], wateRmelon [23], and ENmix [25]. These methods facilitate analytical strategies that help remove unwanted technical variation while allowing for probe retention, which is important as the removal or masking of probes (i.e., missing data) will impact predictor estimates. Within the same package, we implemented all possible combinations of background correction, dye-bias correction, probe correction, and data normalizations as was feasible within the structure of the package. To be inclusive and unbiased in our approach, we did not make a selection on methods a priori. In total, this yielded 101 strategies to prepare DNAm data (Additional file 4). For each sample, raw intensity values were read from IDAT files into an RGChannelSetExtended object in the R programming environment using the read.metharray() function in minfi. Sample quality control was performed by excluding samples with more than 5% of CpG sites with a detection *P*-value greater than 0.05 (using the pfilter() function in the wateRmelon package) and by removing outlying samples based on a low median of chipwide (un) methylation across CpG sites (using the getQC() function in minfi). In total, 44 samples were removed. No probes were filtered out to minimize missing probes in downstream DNAm prediction analysis. We did check the quality of probes used by predictors to calculate estimates and found these overall to be of good quality (Additional file 9). Data processing and normalization were then executed in batches of 96 samples for computational efficiency. The output of each analytical pipeline was a matrix with beta values for each sample. Additional file 10 shows an overview of our sample quality control analysis.

DNAm-based predictors

DNAm predictor estimates were calculated using regression coefficients as reported by the corresponding study unless stated otherwise. Custom R scripts were implemented that take as input a matrix of EPIC array beta values and output predicted estimates as a linear combination of weighted CpG methylation levels. For DNAm clocks, inverse transformation was applied to calibrate the DNAm age estimates in units of years, as required by the algorithm. For instance, Horvath's epigenetic clock regressed log-linear age (that leveraged age at 20) on DNA methylation levels and required this calibration step.

Next, we briefly describe the different predictors included in our study. Additional file 11 presents an overview of predictor characteristics. For full details on each predictor, we refer to their corresponding studies.

DNAm clocks

The following predictors all output a form of DNAm age and capture a different aspect of biological age depending on the characteristics of their training dataset. The Hannum clock uses 71 CpG probes and was developed in a whole blood 450 K DNAm dataset of 656 individuals [7]. The Horvath clock was developed using 3931 multi-tissue and multi-cell type samples using both 27 K and 450 K array samples [6]. The Horvath clock uses 353 CpG probes that are present on both arrays. The BioAge4HAStatic clock is an extended measure of the Hannum clock and defined by forming a weighted average of Hannum's estimate with 3 cell types that are known to change with age: naïve (CD45RA + CCR7 +) cytotoxic T cells, exhausted (CD28 − CD45RA −) cytotoxic T cells, and plasmablasts [17]. The Weidner clock uses 3 CpG and was developed in a 27 K

DNAm dataset of whole blood samples from 575 individuals [11]. The Lin clock uses 99 CpG and was developed in a dataset of 450 K array whole blood samples of 656 individuals [10]. The VidalBralo clock uses 8 CpG probes and was developed in a dataset of 450 K array whole blood tissue of 390 individuals [13]. The Skin & Blood clock uses 391 CpG probes and was developed in a dataset of 450 K and EPIC arrays of a mixture of human fibroblasts, skin tissue, buccal cells, endothelial cells, whole blood, and cord blood samples ($N=896$) [9]. The Zhang clock uses 514 CpG probes and was developed in a dataset of EPIC and 450 K arrays of 13,566 samples. The majority of the samples were derived from whole blood with a small subsample from saliva tissue [8].

Mitotic clocks

The MiAge calculator uses 268 CpG probes and was developed on 4020 samples of 8 cancer types using 450 K DNAm arrays [19]. MiAge outputs an estimate of mitotic age (total number of lifetime cell divisions) for a given human tissue. The epiTOC calculator was developed in a 450 K DNAm dataset of 650 whole blood samples. EpiTOC uses a subset of 385 Polycomb group targets promoter CpGs to predict an estimate of age acceleration in cancer. EpiTOC yields a score, denoted "pcgtAge," as the average DNAm over CpG sites, representing the age-cumulative increase in DNAm at these sites due to putative cell-replication errors [18].

Mortality risk estimators

The Zhang mortality score is defined by a weighted average of 10 CpGs that are associated with mortality status [16]. The Zhang mortality score predictor was trained on a discovery cohort of whole blood 450 K DNAm samples from 954 individuals ($N=402$ deceased at follow-up) and validated in a cohort of 1000 individuals ($N=231$ deceased at follow-up). The second mortality estimator, Levine clock, is a predictor of "phenotypic age," which is a DNAm surrogate of the composite score based on ten mortality markers (9 clinical markers + chronological age) [14]. A training cohort of 456 whole blood samples was then used to identify 513 CpGs predictive of phenotypic age. Only probes available on the 27 K, 450 K, and the EPIC array platform were used in their analysis. The linear combination of the weighted 513 CpGs is called "DNAm PhenoAge." The third mortality risk estimator is GrimAge from Lu et al., which is defined by a composite score based on seven DNAm-based plasma protein markers, DNAm-based pack years of smoking, chronological age, and gender [15]. GrimAge used a training dataset of whole blood samples of 1731 individuals. The DNA methylation profiling was based on the 450 K beadchip but the biomarker was trained on the CpGs present on both the 450 K and the EPIC array in order to ensure compatibility for both platforms. GrimAge was calculated using a python executable that was developed by the authors of the original study, which also outputs several DNAm-based plasma protein markers, three blood cell types, and pack years of smoking (see below).

Plasma protein markers

DNAm-based estimators were developed for the following seven plasma proteins: adrenomedullin (ADM), beta-2-microglobulin (B2M), cystatin-C, growth differentiation factor 15 (GDF-15), leptin, plasmin activator inhibitor 1 (PAI-1), and tissue inhibitor

metalloproteinases 1 (TIMP-1). These plasma proteins were measured using an immunoassay and the predictor trained using a whole blood 450 k DNAm dataset of 1731 individuals in the Framingham Heart Study (FHS) cohort [15]. ADM, B2M, cystatin-C, GDF-15, leptin, PAI-1, and TIMP-1 are defined by 186, 91, 87, 137, 187, 211, and 42 CpGs, respectively. Each of these individual estimates was calculated using the GrimAge python executable.

Smoking predictors

Two DNAm-based smoking predictors were included in our analysis. The Lu estimator was trained using a whole blood 450 K DNAm dataset of 1731 individuals in FHS and uses 172 CpGs for prediction, which is a component of GrimAge [15]. We estimated Lu pack years of smoking using the GrimAge python executable. The McCartney estimator was developed using EPIC DNAm data (only probes that are also present on the 450 K platform) of 3444 individuals [20]. The McCartney estimator uses 233 CpGs and outputs, similar to the Lu predictor, the number of pack years of smoking.

Blood cell type estimator

We included DNAm-based blood cell type estimators for nine cell types in our analysis. For neutrophils (Neu), B cells, monocytes (Mono), natural killer cells (NK), CD4 + T cells (CD4T), and CD8 + T cells (CD8T), estimators were developed using 850 K EPIC DNAm data from magnetic sorted cells [5]. These six cell types were estimated jointly using the estimateCellProp(refdata = "FlowSorted.Blood.EPIC", nprobes = 50) function of the ENmix R package. Plasma B cells (PlasmaBlasts), naive CD8 + T cells, and CD8 +, CD28 −, CD45RA − T cells (CD8pCD28nCD45RAn), were estimated based on the Horvath method [41] and computed using the same python executable as was used for the GrimAge estimator. These estimates are the same estimates that can be obtained through the online DNAm Age Calculator: https://dnamage.genetics.ucla.edu/.

Other estimators

We also included DNAm-based estimators that are developed for body mass index (BMI, in kg/m^2), alcohol (units: per week), educational attainment (Edu, in years), total cholesterol (in mmol/L), HDL cholesterol (in mmol/L), LDL with remnant cholesterol (in mmol/L), total to HDL cholesterol ratio (HDL_ratio), waist-to-hip ratio (WHR), and body fat (in %). These estimators were developed in a whole blood EPIC DNAm dataset (only probes that are also present on the 450 K platform) of between 2819 and 5036 individuals and used between 205 and 1109 CpG sites to predict DNAm-based estimates [20]. Finally, we also included an estimator of leukocyte telomere length (TL). This DNAm-based TL predictor was developed in a whole blood 450 K/EPIC DNAm dataset of 2256 individuals and uses 140 CpGs [12].

Statistical analyses

In the sample of technical replicates, the intraclass correlation (ICC) was calculated using the ICC() function of the R *psych* package (v2.1.3). More specifically, we use ICC(2,1), which is a type of ICC that calculates the degree of consistency from a single measurement using a two-way random effects model [36, 42]. ICC(2,1) assumes

absolute agreement, which means the estimates of the replicates are expected to have exactly the same value. We also calculated ICC(1,1), ICC(3,1), ICC(1,k), ICC(2,k), ICC(3,k) for comparison with other ICC types.

In the general JHS sample (i.e., without technical replicates), we calculated multiple statistical measures on the distribution of the output estimates of each predictor. The coefficient of variation was calculated by dividing the standard deviation by the mean of the distribution of the estimates. DNAm age acceleration residual (ΔAge) was calculated by regressing DNAm age on chronological age using the lm() function in R. To relate DNAm predictor estimates with mortality risk (15% of individuals in JHS are deceased), a Cox proportional hazards regression model was fitted using the coxph() function of the *survival* package (v3.2). Finally, to assess if the above statistical properties change depending on the type of data processing pipeline used, we calculated Spearman correlations between the ICC calculated in the replicate JHS sample and the various statistics generated in the general JHS sample across the 101 pipelines For this, we use the cor.test(method = "spearman") function of the *stats* package. The statistical analyses were performed in R (v4.0.3).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02793-w.

Additional file 1: Supplementary figures 1-6.

Additional file 2. Table that describes full ICC statistics.

Additional file 3. Table that describes ICC statistics (median, min, max) for each of the pipelines.

Additional file 4. Overview of all pipelines implemented with short descriptions.

Additional file 5. Figures that visualize the relationship between predictor estimates and replicate similarity (ICC) for each of the 41 DNAm predictors.

Additional file 6. Table that describes statistics of associations of ICC with phenotypes.

Additional file 7. Figures that visualize the relationship between ICC and replicate sample size for each DNAm predictor.

Additional file 8. Table that describes cohort characteristics.

Additional file 9. Supplementary methods.

Additional file 10. Table that describes sample quality control statistics.

Additional file 11. Table that describes general information on DNAm predictors implemented.

Additional file 12. Peer review history.

Acknowledgements

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I), and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I, and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staff and participants of the JHS. We furthermore thank James G. Wilson for his contributions to the study.

Review history

The review history is available as Additional file 12.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Disclaimers

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the US Department of Health and Human Services.

Authors' contributions
APSO, SH, and RAO conceived of the study. APSO performed data analyses and primary interpretation of results and writing of the manuscript. ATL developed the GrimAge executable and provided code to estimate the GrimAge predictors and run the survival analyses. SH and RAO oversaw the work. JGW and SH provided access to data of the Jackson Heart Study. All authors read, gave input on, and approved the final manuscript.

Funding
ATL and SH were supported by NIH Grant 1U01AG060908 – 01.

Availability of data and materials
All DNAm and phenotype data were obtained from the Jackson Heart Study. The data are not publicly available because they contain information that could compromise the study participant privacy and consent. Raw data can be obtained by submitting a request to the Jackson Heart Study research office: https://www.jacksonheartstudy.org/Research/Study-Data/Data-Access.
Information on data processing and normalization pipelines are available in Additional files 3 and 4. R code for these pipelines can be downloaded from GitHub [43] or Zenodo [44]. These repositories also contain R scripts to calculate estimates for DNAm predictors that are described in the manuscript.

Declarations

Ethics approval and consent to participate
All participants included in this study provided consent per procedures of the Jackson Heart Study.

Competing interests
UC Regents (the employer of SH and ATL) has filed patents surrounding several epigenetic biomarkers of aging (including GrimAge) which list SH and ATL as inventors. The other authors declare that they have no competing interests.

Received: 8 October 2021 Accepted: 11 October 2022
Published online: 24 October 2022

References

1. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. 2019;20:590–607.
2. Schübeler D. Function and information content of DNA methylation. Nature. 2015;517:321–6. https://doi.org/10.1038/nature14192.
3. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17:208.
4. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014;30:1431–9.
5. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19:64.
6. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.
7. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–67.
8. Zhang Q, Vallerga CL, Walker RM, Lin T, Henders AK, Montgomery GW, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. Genome Med. 2019;11:54.
9. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and studies. Aging. 2018;10:1758–75.
10. Lin Q, Weidner CI, Costa IG, Marioni RE, Ferreira MRP, Deary IJ, et al. DNA methylation levels at individual age-associated CpG sites can be indicative for life expectancy. Aging. 2016;8:394–401.
11. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 2014;15:R24.
12. Lu AT, Seeboth A, Tsai P-C, Sun D, Quach A, Reiner AP, et al. DNA methylation-based estimator of telomere length. Aging. 2019;11:5895–923.
13. Vidal-Bralo L, Lopez-Golan Y, Gonzalez A. Simplified assay for epigenetic age estimation in whole blood of adults. Front Genet. 2016;7:126.
14. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging. 2018;10:573–91.
15. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging. 2019;11:303–27.
16. Zhang Y, Wilson R, Heiss J, Breitling LP, Saum K-U, Schöttker B, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. Nat Commun. 2017;8:14617.
17. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai P-C, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging. 2016;8:1844–65.
18. Yang Z, Wong A, Kuh D, Paul DS, Rakyan VK, Leslie RD, et al. Correlation of an epigenetic mitotic clock with cancer risk. Genome Biol. 2016;17:205.

19. Youn A, Wang S. The MiAge Calculator: a DNA methylation-based mitotic age calculator of human tissue types. Epigenetics. 2018;13:192–206.
20. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19:136.
21. Hannon E, Knox O, Sugden K, Burrage J, Wong CCY, Belsky DW, et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. PLoS Genet. 2018;14:e1007544.
22. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9 Nature Publishing Group.
23. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14:293.
24. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014;15:503.
25. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Res. 2016;44:e20–e20. https://doi.org/10.1093/nar/gkv907.
26. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29:189–96.
27. Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics. 2016;32:2659–63.
28. Xu Z, Langie SAS, De Boever P, Taylor JA, Niu L. RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. BMC Genomics. 2017;18:4.
29. Maksimovic J, Gordon L, Oshlack A. SWAN: subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biol. 2012;13:R44.
30. van Rooij J, Mandaviya PR, Claringbould A, Felix JF, van Dongen J, Jansen R, et al. Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. Genome Biol. 2019;20:235.
31. Wu MC, Joubert BR, Kuan P-F, Håberg SE, Nystad W, Peddada SD, et al. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. Epigenetics. 2014;9:318–29.
32. Wang T, Guan W, Lin J, Boutaoui N, Canino G, Luo J, et al. A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. Epigenetics. 2015;10:662–9. https://doi.org/10.1080/15592294.2015.1057384.
33. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013;8:333–46.
34. Taylor HA Jr, Wilson JG, Jones DW, Sarpong DF, Srinivasan A, Garrison RJ, et al. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethn Dis. 2005;15:S6-4–17.
35. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet. 2018;19:371–84.
36. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155–63.
37. Oytam Y, Sobhanmanesh F, Duesing K, Bowden JC, Osmond-McLeod M, Ross J. Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets. BMC Bioinformatics. 2016;17:332.
38. McEwen LM, Jones MJ, Lin DTS, Edgar RD, Husquin LT, MacIsaac JL, et al. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. Clin Epigenetics. 2018;10:123.
39. Lee Y, Sun D, Ori APS, Lu AT, Seeboth A, Harris SE, et al. Epigenome-wide association study of leukocyte telomere length. Aging. 2019;11:5876–94.
40. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9. https://doi.org/10.1093/bioinformatics/btu049.
41. Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock. J Infect Dis. 2015;212:1563–73.
42. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420–8.
43. Ori APS. Analysis code of DNAm data preprocessing and normalization strategies and implementation of DNAm predictors. GitHub. 2021. https://github.com/anilpsori/DNAm_pipelines_and_biomarkers.
44. Ori APS. Analysis code of DNAm data preprocessing and normalization strategies and implementation of DNAm predictors. Zenodo. 2022. https://doi.org/10.5281/zenodo.7150375.

Publisher's Note