HRS | HEALTH AND RETIREMENT STUDY
A Longitudinal Study of Health, Retirement, and Aging
Sponsored by the National Institute on Aging

HRS DNA Methylation – VBS 2016

Jessica D. Faul, University of Michigan
Eileen Crimmins, University of Southern California
Sarah Munro, University of Minnesota
Bharat Thyagarajan, University of Minnesota
David R. Weir, University of Michigan

Survey Research Center
Institute for Social Research
University of Michigan
Ann Arbor, Michigan

October 2023

Table of Contents

I. Introduction

This document describes a data set consisting of values for DNA methylation data derived from the 2016 Health and Retirement Study Venous Blood Study. This is a HRS restricted health data release available from NIAGADS (https://www.niagads.org/). This release includes both a file with methylation beta values (CSV format) and BeadArray data output directly from the Illumina scanners (IDAT format). The file with methylation beta values is approximately 20 GB unzipped and 4 GB zipped in size. The full IDAT files require approximately 107 GB of storage (55 GB zipped).

A. Rationale

DNA Methylation (DNAm) is one mechanism by which exposure to adverse life circumstances and environments are linked to health outcomes related to aging. DNAm occurs with the addition of a methyl group to a CpG site in DNA. A number of researchers have identified portions of the genome where methylation changes are related to either age or, more recently, to health outcomes linked to age. DNA methylation is considered one of the Hallmarks of Aging (López-Otín, 2013).

B. The HRS Methylation Sample of Respondents

DNA methylation assays were done on a non-random subsample (n=4,104) people who participated in the 2016 Venous Blood Study. The sample includes all the participants of the 2016 Healthy Cognitive Aging Project (HCAP) who have provided blood samples, plus younger participants designated for future HCAP assessments, and a subsample of HCAP non-participants. This subsample, once weighted, fully represents the entire HRS sample. A total of 4,018 samples passed quality control (QC). The sample is 58% Female and has a median age of 68.7 years. It is racially diverse: Non Hispanic White (n=2,669, 66.4%), Non Hispanic Black (n=658, 16.4%), Hispanic (n=567, 14.11%), Non Hispanic Other (n=124, 3%). The sample is also socioeconomically diverse. The educational distribution is as follows: Less than High School (16.8%), High School / GED (52.12%), Some College (5.97%), College Degree or Higher (24.1%), Other (1%).

C. Collection

The 2016 VBS blood collection was managed by Hooper Holmes Health & Wellness. The phlebotomy service was provided with the names, addresses, and phone numbers of consenting respondents and contacted respondents to set appointments. Collection materials were mailed to the phlebotomists' homes in advance of the scheduled visit. Every attempt was made to schedule the blood draw within 4 weeks of the HRS core interview. Fasting was recommended and preferred but not required. Phlebotomists noted the fasting status of the samples. We collected 50.5 mL of blood in 6 tubes – 1 x 8 mL CPT tube, 3 x 10 mL double gel serum separator tubes (SST), 1 x 10 mL EDTA whole blood tube, and a 2.5 mL PAXgene RNA tube. The SST tubes are centrifuged in the field before being shipped overnight to the CLIA-certified Advanced Research and Diagnostic Laboratory at the University of Minnesota. Tube processing is done within 24 hours of arrival at the lab (within 48 hours of collection). DNA for methylation analysis was done using DNA extracted from the EDTA tube.

More information on the 2016 Venous Blood Study, including details on sampling, consent, and administration, is provided in the VBS 2016 Data Description.

D. Protocols for DNA Methylation Data

DNA methylation data are based on assays done using the Infinium Methylation EPIC BeadChip v1.0 at the University of Minnesota. Samples were randomized across plates by key demographic variables (i.e. age, cohort, sex, education, race/ethnicity) with 40 pairs of blinded duplicates. Analysis of duplicate samples showed a correlation >0.97 for all CpG sites.

The *minfi* package in R software was used for data preprocessing, and quality control. 3.4% of the methylation probes (n=29,431 out of 866,091) were removed from the final dataset due to suboptimal performance (using a detection P-value threshold of 0.01). Analysis for detection P-value failed samples was done after removal of detection P-value failed probes. Using a 5% cut-off (*minfi*) we removed 58 samples. We also removed sex mismatched samples and any controls (cell lines, blinded duplicates). High quality methylation data is available for 97.9% samples (n=4,018).

E. Subsample Weights

Respondents with at least one valid venous blood result (VBS16VALID in the HRS Tracker data file) were assigned a VBS weight. The weights were adjusted for the differential probabilities of participation by dividing the HRS 2016 sample weight by the predicted probability of having a valid venous blood result among community-dwelling 2016 HRS respondents born prior to 1960, excluding all members of the LBB cohort. The resulting interim weight was trimmed at the 1st and 99th percentiles and was then post stratified back to the entire 2016 HRS sample born prior to 1960 by age, sex, and race/ethnicity. Two separate respondent-level weights were created for the VBS 2016 Innovative Sub Sample and should be used for analyses of data from that sample. **VBSI16WGTRA should be used for analyses including DNA methylation**. Sample weights can be found in the HRS Tracker data file.

II. DNA methylation QC Report

A. Overview

The following section outlines the QC process we conducted, including summary plots and statistics, for the full set of HRS DNA Methylation Data. There are 5 sections:

1. Comparison of Reported and Predicted Sample Sex Values
2. Evaluation of Median Methylated and Unmethylated Intensities
3. Detection P-value Probe and Sample Flagging Statistics
4. Blinded Duplicate Analysis
5. Data Included with the NIAGADS Release

For detection p-value probe and sample flagging, we found a significant difference in the flagging rate for detection p-values calculated using the minfi package vs. the ewastools package. The decision was made to find the detection P-value flagged probes that overlap between the minfi and ewastools methods (the minfi flagged probes contained within the ewastools flagged probes). Dropped probes and dropped samples were selected based on the minfi detection P-value stats. An additional set of probes were flagged based on ewastools detection P-values.

```
The University of Minnesota processed 4224 total arrays for the HRS project.

 These included 81 cell line controls and 40 blinded duplicates.

 There were 4103 unique study samples remaining after removing any control samples at
the start of the QC process.
```

B. Comparison of Reported and Predicted Sample Sex Values

During QC analysis we found and 33 samples where the predicted and reported sex information were discordant. The sample identities could not be resolved and were dropped from the final sample.

1. Cell Line Control Reported vs Predicted Plots

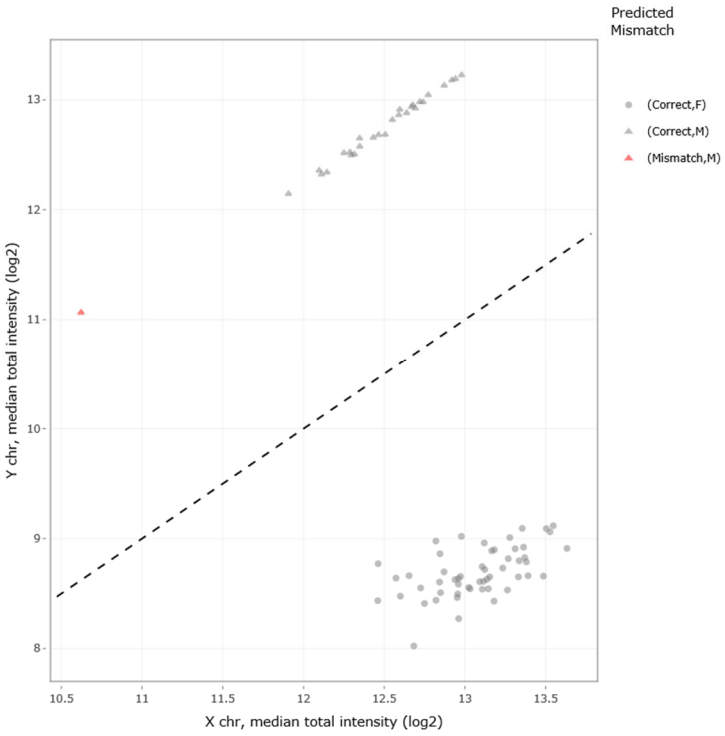
The University of Minnesota Genomics Center (UMGC) runs methylation arrays on a trio of cell line controls:

NA10858 -- male

NA10859 -- female

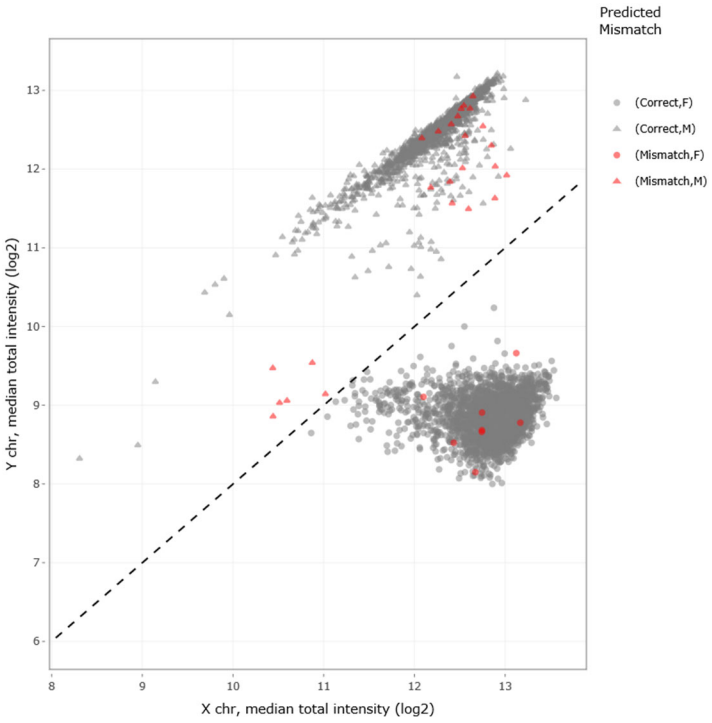
NA11875 -- female

The following plot shows the predicted and reported sex values for the cell line controls.



2. Unique Study Sample Reported vs. Predicted Sex Plot

This plot shows whether the reported or predicted sex values match for the unique study samples plus controls. Blinded duplicates are not included in this plot.



C. Evaluation of Median Methylated and Unmethylated Intensities

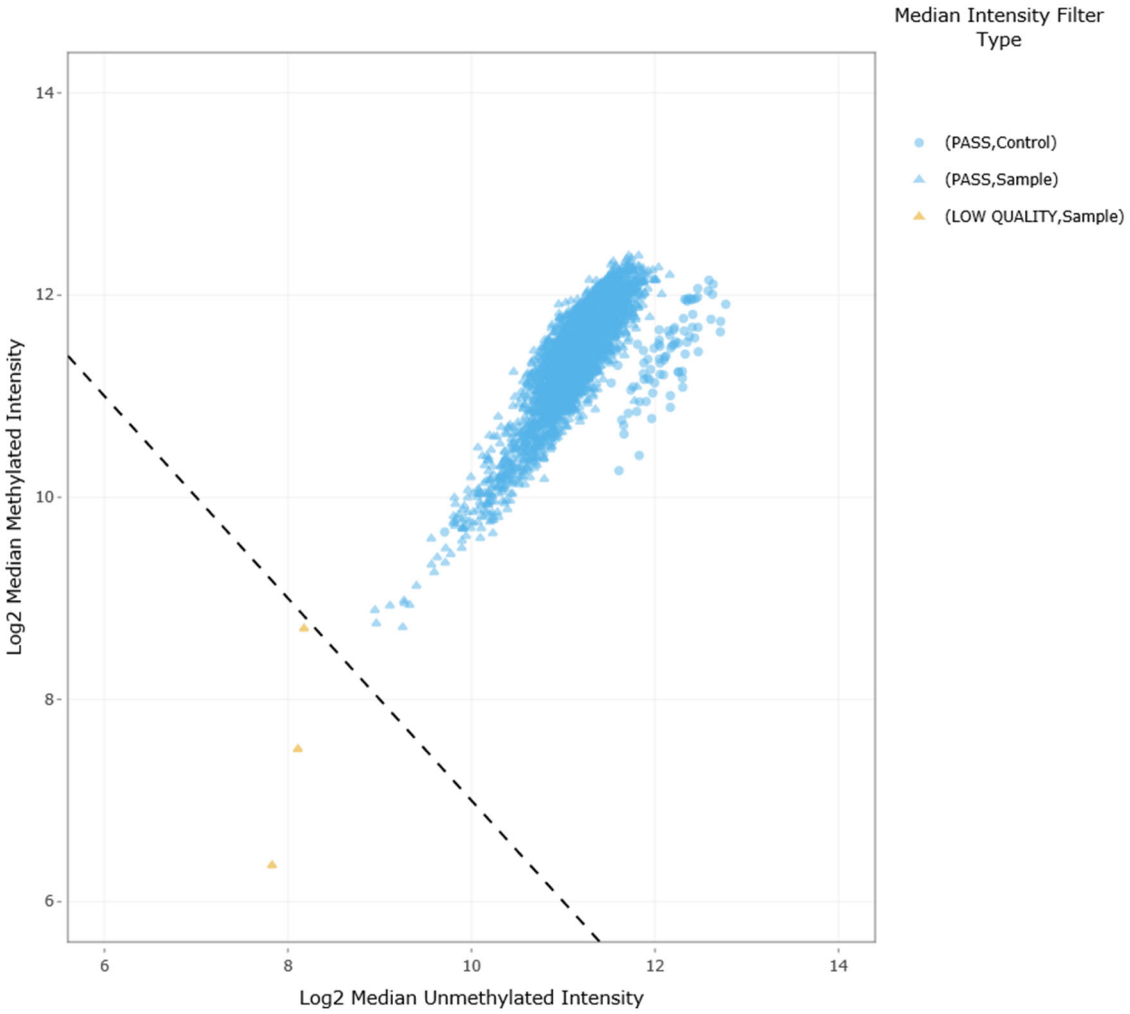
Based on an evaluation of the Log2 Median Methylated and Median Unmethylated intensities for each sample, the samples are labeled as either "PASS" or "LOW QUALITY". The average methylated and unmethylated intensity cut-off of 8.5 is a heuristic based on EPIC array data:

average(log2(Median Methylated) + log2(Median Unmethylated)) > 8.5

The few samples that were below this threshold were not removed from the data based solely on this threshold but may have been removed based on either sex mismatch or high probe failure rate.

Most Control samples are clustered together as a subgroup of those marked on the figure as "PASS".

1. M vs. U Plot for all Unique Study Samples (including cell line controls)



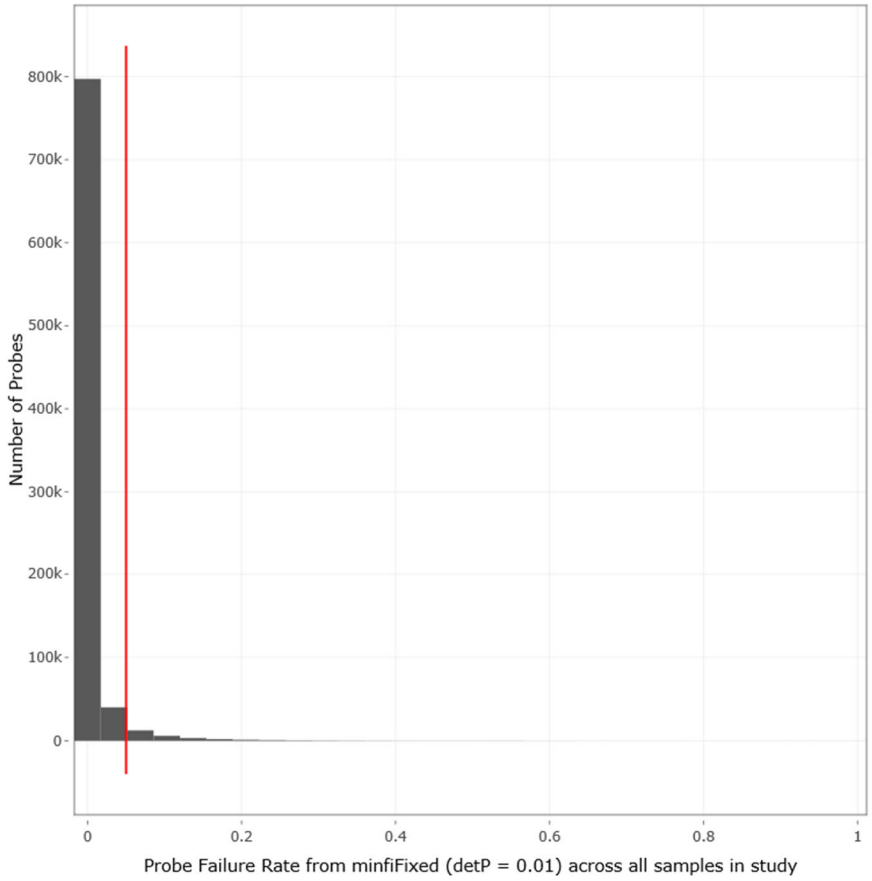
D. Detection P-value Probe and Sample Flagging Statistics

If a probe had a detection P-value > 0.01 in > 5% of all samples (across all groups, the entire cohort), then the probe was removed from analysis for detection P-value failure. In total, 29,431 probes were cut from the released beta matrix file and an additional 168,615 probes identified by ewastools were flagged (see probesToFlag.csv).

```
40064 positions failed in > 5 % of all samples in all groups for minfi
 with Det P threshold = 0.01.
29431 positions failed in > 5 % of all samples in all groups for minfiFixed
 with Det P threshold = 0.01.
48108 positions failed in > 5 % of all samples in all groups for minfiFixed
 with Det P threshold = 1e-4.
82570 positions failed in > 5 % of all samples in all groups for minfiFixed
 with Det P threshold = 1e-10.
198046 positions failed in > 5 % of all samples in all groups for ewastools
 with Det P threshold = 0.01.
172165 positions failed in > 5 % of all samples in all groups for ewastools
 with Det P threshold = 0.05.
157067 positions failed in > 5 % of all samples in all groups for ewastools
 with Det P threshold = 0.1.
```

1. Histogram of Detection P-value per Probe

Histogram showing the per probe detection P-values. Red line indicates 5% cut off.



2. Dropping Detection P-value Failed Samples

Analysis for detection P-value failed samples was done after removal of detection P-value failed probes from the full sample cohort. Using a 5% cut-off we find that we have the following numbers of samples removed with different methods and thresholds:

```
Number of samples flagged by the minfi Fixed approach is 58
this is ~ 1.373106 % of all study samples.
Number of samples flagged by the ewastools (det P thresh = 0.01) approach is 149
this is ~ 3.527462 % of all study samples.
Number of samples flagged by the ewastools (det P thresh = 0.05) approach is 145
this is ~ 3.432765 % of all study samples.
```

We decided to remove the 58 samples flagged by the Minfi Fixed approach from the data set.

E. Blinded Duplicate Analysis

All of the original and duplicate samples had correctly matched predicted and reported sex values.

We then checked to make sure that the beta values were consistent for the original and duplicate samples for each pair. The majority of the samples were found to be properly paired. One pair was found not to be properly matched - the correlation for this original/duplicate pair was less than 0.52. It was determined that one of these samples was likely mislabeled at point. We dropped all duplicates plus the original sample that didn't have a proper snp match among the blinded duplicate pairs from the released file.

After sample removal for sex mismatch, p-value failure, and the failed blinded duplicate, the final methylation data includes **4018 study samples** out of 4224 total arrays.

F. Data Included with the NIAGADS Release

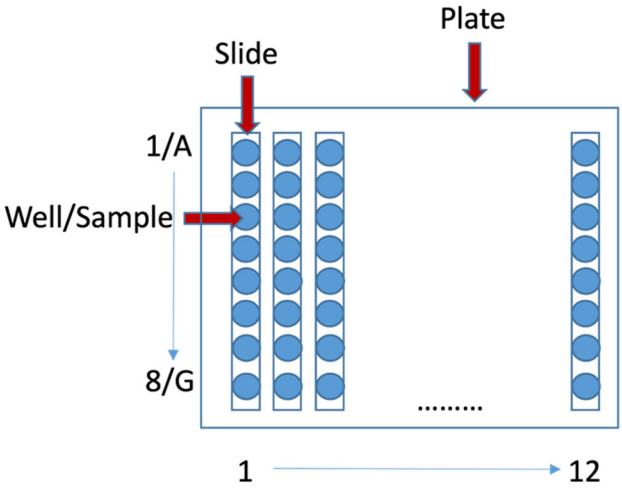
1. Subject IDs

The Subject IDs included in the data release are NOT the HRS survey IDs (HHID/PN). In order to link to the HRS survey data you will need to request the HRS-NIAGADS DNAm Cross-Reference File from HRS. See the HRS website for more information. IDAT files are labled using the slide, row and column information for the sample. The file idat_crosswalk.csv links the IDs used on the IDAT files to the Subject IDs used elsewhere.

2. Additional Variables

Additional variables have been added to the NIAGADS data release. Plate (N=48), array (same as row, N=8) and slide (combines column and plate information, N=527) are variables that can be used to adjust for technical variation. Each plate contains 12 slides with 8 samples per slide. Each sample is uniquely identified by slide and array.

Additionally, DNA methylation is different by cell type, so the cell composition of the sample can affect the resulting data. We have added selected cell data from corresponding complete blood count and flow cytometry analyses from the VBS 2016 project to help assist with cell type composition adjustment. Note that the variables included with the release represent cell type percentages, not counts. Users can also estimate cell type composition using the DNAm data itself. More information on cell type estimation methods can be found in the reference section (Houseman, 2012).



Picture courtesy of Scott Ratliff

Age at data collection, gender, self-reported race and Hispanic ethnicity are also included as phenotype data with the released files.

3. Principal Components

We found test-statistic inflation was an issue when conducting trial EWAS studies of some phenotypes (i.e. high inflation factors when examining Quantile-quantile (*QQ*) plots). While there are many methods to control for this, one is to use principal components (PCs) to try to capture unmeasured, possibly technical, variation. We have constructed and released 20 PCs with the DNAm data to assist with this kind of model adjustment.

However, it is important to note that some PCs covary with important phenotype variables that a user might not want to adjust for in their model (e.g. gender or race). We used feature selection methods to identify covariates represented by each of the first 20 principal components. Features considered include age, years of education, Activities of Daily Living, Instrumental Activities of Daily Living, BMI, self-rated health, cell composition, number of chronic conditions, wealth (quartiles), census region, coupled status, current smoker, ever smoker, cohort, gender, race, Hispanic ethnicity, sample plate, row, and slide. Variables were standardized and data was split into testing and training datasets. Four machine learning regression algorithms were utilized, Lasso, Random Forest, Elastic Net, and Gradient Boosting. Grid search was utilized for parameter tuning and models were fit to the training data. Models were applied to the testing data and permutation importance was utilized to mitigate tendency of models to favor high dimensional features. For each PC, the top 5 features were examined to identify common important features between the models. The results are summarized in the table below.

Covariates identified for released PCs

PC	Covariate Identified	Notes
1	Sample plate + slide effects	plates 137, 102, slides 202702240149, 202172220168
2	Sample plate + slide effects	137, 136, 191, 189, 106
3	Gender + black race + cell composition	
4	Gender	
5	Sample plate + slide effects	Plates 191, 189, 192, 109, slides 202702240149, 202253670031
6	Sample plate effect	136, 137, 118, 117, 128
7	Cell composition + age + black race	
8	Black race	
9	Sample plate + slide effects	Plates 191, 108, 189, 106 slides 202184900146, 202194900152
10	Black race + cell composition	
11	Black race + cell composition + age	
12	Sample plate & slide effects	Plates 128, 136, 107, 125 Slide 202178770032
13	Age + plate effects	

14	Age + plate effects	
15	Sample plate effect	110, 131, 140, 191
16	Hispanic + mix of other covariates	Row 8
17	Hispanic	
18	Hispanic + age + plate effects	
19	Sample plate effects	141, 132, 143, 140
20	Sample plate effects	137, 112, 113, 111

III. Acknowledgements

We would like to acknowledge the help of Em Arpawong, Jonah Fisher, Eric Klopack, Helen Meier, Trey Smith, and Sithara Vivek for their review of the data and this documentation report.

IV. If You Need to Know More

This document is intended to serve as a brief overview to the HRS DNA methylation data products. If you have questions or concerns that are not adequately covered here or on the HRS website, or if you have any comments, please contact us. We will do our best to provide answers.

A. HRS Internet Site

Health and Retirement Study public release data and additional information about the study are available on the Internet. To access public data or to find out more about restricted data products and procedures, visit the HRS website.

B. Contact Information

If you need to contact us, you may do so by one of the methods listed below.

Internet: Help Desk at the HRS Web site (http://hrsonline.isr.umich.edu)

E-mail: hrsquestions@umich.edu

C. Citing this Document

Please include the following citation in any research reports, papers, or publications based on these data along with the citation for the reference epigenetic clock:

In text: "The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (NIA U01AG009740) and is conducted by the University of Michigan."

In references: "Faul JD, Crimmins E, Munro S, Thyagarajan B, Weir DR. HRS DNA Methylation Data – VBS 2016. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan; Oct 2023."

V. References

Houseman, E.A., Accomando, W.P., Koestler, D.C. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012; 13, 86. https://doi.org/10.1186/1471-2105-13-86

López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013 Jun 6;153(6):1194-217. doi: 10.1016/j.cell.2013.05.039. PMID: 23746838; PMCID: PMC3836174.