bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

R methylCIPHER: A Methylation Clock Investigational Package for Hypothesis-Driven Evaluation & Research

Kyra L. Thrush[1,2], Albert T. Higgins-Chen[3,4], Zuyun Liu[5], Morgan E. Levine[2*]

[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

[2] Altos Labs, San Diego Institute of Science, San Diego, CA, USA

[3] Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA

[4] VA Connecticut Healthcare System, West Haven, CT, USA

[5] Center for Clinical Big Data and Analytics of the Second Affiliated Hospital and Department of Big Data in Health Science School of Public Health, Zhejiang University School of Medicine, The Key Laboratory of Intelligent Preventive Medicine of Zhejiang Province, Hangzhou, Zhejiang, China

* Corresponding Author

Abstract

Background: Epigenetic clocks are promising tools for the study of aging in humans. The clocks quantify biological aging above and beyond chronological age, demonstrate systematic associations with risk factors that accelerate aging, and predict age-related morbidity and mortality. There is interest in using them as surrogate endpoints in intervention studies. However, the large number of clocks, decentralized publication and explosive popularity in the last decade has made for poor accessibility and standardization. This has hampered the abilities of new researchers to conduct truly hypothesis driven research—whether by not knowing about the best available clocks for a given question, or by systematically testing many or all as they become available.

Results: We report a centralized R package which can be installed and run locally on the user's machine, and provides a standardized syntax for epigenetic clock calculation. The package includes a set of helper functions to assist with navigating clock literature and selecting clocks for analysis, as well as affording the user with the details of clock

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

calculation. We describe each clock's resilience to missing CpG information, combined with functionality to assess the need for imputation in the user's own data. Furthermore, we demonstrate that while CpGs may not be shared among clocks with similar outputs, many clocks have highly correlated outputs.

Conclusions: Due to the previous decentralization of epigenetic clocks, gathering code and performing systematic analysis, particularly in protected datasets, has required significant information gathering effort. Here, we offer an R package with standardized implementation and potential for future growth and clock incorporation to assist with hypothesis driven investigation of aging as measured by epigenetic clocks. We show the potential of this package to drive the user to think globally about signals captured by epigenetic clocks, as well as to properly identify the potential and limitations of each clock in their current research.

Keywords

Background

Epigenetic clocks are promising tools, often discussed as future surrogate biomarkers for studies of aging and longevity. These clocks have been extensively reviewed; for their phenotypic associations [1, 2]; to understand the mechanisms of epigenetic aging [3, 4]; [5][6] Our current intent is to instead provide a practical overview of the categories, training methods, and applications of existing epigenetic clocks. As epigenetic clock research gathers further momentum in the study of aging, it is increasingly clear that a centralized toolkit to introduce the epigenetic clocks is essential. Such a toolkit must satisfy, in our estimation, a handful of requirements: It

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

must (1) organize thematically and systematically the existing epigenetic clocks to minimize the risks of multiple testing and publication bias; (2) provide functionality to allow the researcher to perform not only *pro forma* analyses, implicating epigenetic clocks in a disease or dataset of interest, but push researchers to glean further biological insight as to the associations found; (3) be complete in its access of epigenetic clocks, while still giving editorial insight such as to make use of them navigable; (4) be sufficiently flexible so as to allow future advances to be made equally accessible.

Because of the relative ease of training epigenetic clocks and of DNA methylation (DNAm) collection, as well as the numerous age-related CpGs in the genome, there are currently numerous human epigenetic clocks available in the literature (Figure 1). The earliest such clocks utilize 1-10 highly age-associated CpGs in regression models, and these remain useful as low-cost assays [7–11]. However, the advent of large scale, streamlined collection of DNA methylation data on lllumina Beadchip methylation array technologies, as well as the adoption of elastic-net penalized regression to the training method, led to a new generation of clocks that can capture genome-wide aging signals. The first of these were trained to predict chronological age with high accuracy, including the Hannum blood [12] and Horvath multi-tissue [13] clocks, and have since expanded to include additional clocks [14–17]. For those studying development and gestation, significant effort has been spent to create reliable gestational and pediatric age clocks [18–22]. Similar approaches led to the generation of mitotic clocks, so-called for their presumed ability to track the rate of mitotic divisions and project cancer risk [23–25]. The telomere length estimator

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DNAmTL, which is highly correlated with cellular replication rate, can be included in this category as well [26].

In addition to clocks that predict discretely measurable aspects of age and cell turnover, efforts have been made to capture heterogeneity in aging as meaningful biological signal. These [27][28], and DunedinPACE[29] are trained to predict individuals; degree or rate of biological change with time, especially those changes that contribute to age-related morbidity and mortality risk. It has also been found that DNAm can be used to predict various traits, lifestyles or exposures that are not necessarily related to aging [30].

Finally, there has been recent interest in “bespoke” clocks designed for particular tissues, diseases, data types, or applications. For example, one ongoing challenge for epigenetic clocks is that most human clocks were trained using primarily whole blood[2]. Multi-tissue clocks[13]s conserved across tissues and may ignore tissue-specific aging changes [6]. Thus, tissue-specific, bespoke clocks have been developed, including for skin-and-blood [16], brain cortex [31], skin [32], and the scAge framework for predicting epigenetic age from single cell methylation [33] have been reported. Also, since clocks are often trained in large aging cohorts, it is possible they may miss aging patterms that occur in small subsets of the population, as in the context of rare diseases. Due to the increasing abundance of non-blood tissue DNAm, new methods for collection [34], additional approaches to clock-training [35, 36], and emerging age-related diseases, we predict that the number of bespoke clocks will see a dramatic increase in the next few years.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Yet the challenge remains: Which clock should be used? The variety of epigenetic clocks can be useful for investigating many different aspects of aging. But selecting the appropriate clock(s) for a study requires navigating a decentralized body of nuanced literature. The choice of clock may be impacted by the phenotype that they were trained to predict or the context they were trained for. However, the differences between clocks can be subtle, amounting to differences in training data composition or procedure, such as age ranges [22], or preselection of CpGs [23].

This clock selection problem creates concerns regarding the integrity and interpretability of aging studies. In particular, there are two consequences we would hope to avoid. The first is exclusive, repeated testing of the best cited and most reported aging clocks—namely the Horvath multi-tissue, Hannum, Levine PhenoAge and GrimAge DNA methylation clocks. [1]. This aligns well with the plethora of publications reporting the associations of acceleration of the Horvath multi-issue, PhenoAge, and GrimAge predictors with biological changes[37–39], and disease risk or mortality[40–45]. While this produces some standardization in the field, these clocks are not necessarily the optimal choice in all cases. If a researcher instead has all available clocks at their disposal and then applies a hypothesis-driven selection of clocks, an alternative, lesser-known clock may indeed be the optimal choice. The second unintended consequence could be that individuals test many clocks as clocks are published or the researcher becomes aware of them, and only significant results tend to be noted and published.

The variety of clocks and their decentralized distribution also creates practical obstacles for aging research. Researchers wishing to apply epigenetic clocks must first

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mine the literature for their options, identify one (or multiple) clocks to test, locate and download the data to do so, and ensure that the calculation is properly performed across their samples. This process creates substantial logistical barriers for researchers. Clocks that are published along with public code and data should be applauded. Furthermore, a few existing platforms can calculate multiple clocks. These include the online Horvath calculator (http://dnamage.genetics.ucla.edu) and EstimAge (https://estimage.iac.rm.cnr.it). However, these require data to be uploaded to a third party server, which is prohibited for protected datasets and limits researchers' access to the underlying details of calculation. There also exists the methylclock Bioconductor package, which is currently limited to chronological age clocks, gestational age clocks, DNAmTL and PhenoAge. In summary, the sheer number and variety of clocks creates two primary challenges that impede use by the broader scientific community: (1) the selection of the most appropriate clock(s) for the scientific question or hypothesis at hand; (2) access to the many clocks. We address this by providing a centralized resource in which individuals can explore, investigate, and calculate any and all clock(s) appropriate for a research question from a project's inception. This is a necessary improvement for the field, as it allows for systematic study of epigenetic clocks, which in turn advances future understanding of their underlying relationships and biological significance.

There is currently no true standard format or resource for the researcher to publish and distribute clocks. Here we present a consolidated resource that applies a standardized format to the calculation of epigenetic clocks, establishes a repository for the fitted values of existing clocks, and provides a few helper functions for the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

exploration of appropriate clocks, their CpGs, and inter-clock correlations. Further, this package can be installed and run on a local machine, eliminating the need for uploading of potentially protected data, and affords near-immediate results to the researcher, regardless of the number or identity of clocks they choose to select. To further facilitate accessibility of the epigenetic clocks, we have also provided a thorough tutorial walking through use of the package and questions to be addressed on the github page for this package (github.com/MorganLevineLab/methylCIPHER).

Implementation

This should include a description of the overall architecture of the software implementation, along with details of any critical issues and how they were addressed.

The current package is implemented using the R programming language and distributed via installation from Github (github.com/MorganLevineLab/methylCIPHER). This distribution allows us to provide a flexible, regularly updated, and community driven package. Not only can we push regular updates to users as new clocks are added, but the research community can rapidly suggest new clocks, helper functions, or improvements to code. Users wishing to generate their own independent Github R packages during the publication process of novel clocks can be imported by this package, or be referred to in the online Github based wiki and tutorial. The functions of this package are represented in the schema in Figure 2.

To calculate epigenetic clocks in methylCIPHER (functions of the form *calc[Clock-Name]*), the user provides a labeled data matrix of pre-processed methylation Beta values obtained via the Illumina HumanMethylation450 Beadchip or Infinium MethylationEPIC kit from Illumina (San Diego, CA). These are commonly referred to as

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

450k and EPIC arrays respectively. Preprocessing and normalization of methylation data is typically performed within R using the *minfi* [46], *wateRmelon* [47], or *SeSAMe* [48] packages, however methylCIPHER functions regardless of normalization protocol. For more details regarding the effects of choice of normalization, refer to Ori et al. [49]. The user must simply have an object of *matrix* or *data frame* class, with named columns corresponding to the Illumina CpG names, and cells containing methylation beta ratios between 0 and 1.

We recommend that the methylation data object have named rows corresponding to unique sample identifiers. Most analyses will benefit from a corresponding "phenotype" data frame with sample identifiers; sample metadata; demographics; health outcomes; age; sex; and other traits of interest. While optional for individual clock calculations, this typically assists the researcher with downstream analyses. Without this data frame, clocks can be computed using only a single function at a time, with output to a vector object. However, the "pheno" dataframe provides a central location to append multiple clocks to if using the multi-clock wrapper functions *calcUserClocks* or *calcCoreClocks*.

Of note, for some analyses (for example, calculation of IEAA or EEAA [50]) estimates of blood cell composition are necessary. To obtain such estimates, individuals may want to use the Houseman method [51] of cell-type deconvolution from the *minfi* package. However, local methods for predicting cell composition of blood can be effectively run only when preprocessing occurs from raw methylation files (i.e. IDAT). If access is limited to preprocessed methylation beta values, as in some publicly available datasets, the Horvath Online calculator can predict blood component proportions. We hope to provide users with this convenient functionality in the future. However, we have

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

provided a formatting function *formatHorvathOnline* which allows the user to quickly generate the input files for the Horvath Online calculator. This allows for both calculation of blood composition estimates and GrimAge [28].

The current version of the *R/methylCIPHER* package has been tested on both Windows and Mac computer systems, running R v3.6+. It can run on most modern personal computers, requiring less than 16 GB of RAM (and in many cases less than 8 GB) for methylation datasets containing hundreds of samples at once. The functions run efficiently and will provide near-immediate results within seconds or minutes.

Results & Discussion

The R package *methylCIPHER* provides both seasoned and casual users of epigenetic clocks with the tools necessary for thorough, hypothesis-driven research using existing epigenetic clocks. We provide a comprehensive listing of human epigenetic clocks that use; (1) linear approaches; and (2) CpGs found in the commonly utilized Illumina 450k and EPIC arrays. This broad set of clocks can be searched through the function *getClockOptions(),* which allows users to explore their options. We have also provided convenient referencing of the source papers for each of the clock calculation functions, using *citeMyClocks*. This accepts a group or list of functions at once, which in turn allows readers to quickly refer to the original clock papers and understand the underlying principles of their training. This process of information gathering is shown graphically in the top right region of Figure 2.

Due to the variable performance of experimental designs, and a multitude of existing pipelines for quality control, users may find missing probes or DNAm values in their data. This can be seen as missing probe columns in the final normalized beta

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

value matrix, columns of *NA* values, or sporadic *NA* values in sample/ probe pairs. If the probes are missing entirely from the beta value matrix, this can impact the decision to implement specific clocks. Therefore, *getClockProbes* provides the user with a table to determine what portion of probes are available for the various clock options, so that a clearly informed decision can be made. Alternatively, they may find columns of all NA values, which can be removed using *removeNAcol.* Sporadic missing values for select probe/sample pairs can either be mean imputed across the matrix samples (using function *meanimpute*), or imputation can be done later within the clock calculation functions if a mean vector is provided as a reference containing the necessary CpGs. However, most clocks can be calculated without imputation by simply ignoring those CpGs in the resultant weighted regression values for such samples. The effect of doing so varies by clock.

In figure 3 we visualize the degree of information contained in individual CpGs. The model contribution of each CpG was estimated by multiplying the absolute regression value in each clock for each site by its standard deviation in the Framingham Heart Study (FHS) offspring cohort [52], and plotted against the standard deviation alone. These results are shown for the Hannum (Figure 3A), Horvath Multi-Tissue (Figure 3B), and PhenoAge (Figure 3C) clocks (additional clocks in supplemental materials, figures S1-S2). With the CpGs plotted in this manner, we can see that each CpG in the Hannum clock tends to have higher weight, evident from the higher mean contribution represented by the blue horizontal dashed line. Further, PhenoAge employs CpGs with higher standard deviation than the other clocks, but due to lower weight in the clock regression, these tend to have dampened model contributions. These plots

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

help to conceptualize the effect of using mean imputation on model CpGs, as utilizing mean imputation removes signal in individual samples that may reflect meaningful inter-individual differences.

To functionalize this effect, we simulated the effects of increasing amounts of missing array methylation probes for a given sample. We performed 1000 iterations each of randomly drawing 0.01, 5, 10, and 20% of CpGs within each clock. Then, we found the distribution of *Clock Years* information contained within each potentially missing sample. Because we are approximating information lost, the *Clock Years* measurement is defined by the absolute value of the sampled CpGs' clock regression coefficient, multiplied by the standard deviation of these CpGs. The result is an approximation of the information in years of measurement lost for a sample or set of samples for which that proportion of CpGs's methylation is unavailable. We note that because *Clock Years* is defined as absolute values for each CpG, oppositely weighted CpGs in a clock do not counteract each other. Thus, the information lost may reflect a larger difference than the shift in actual clock value. As each clock selected is of varying size, the percentage of missing CpGs varies in the absolute number lost. Here, we demonstrate that *Clock Years* information lost appears strongly associated with clock size: Bigger clocks show a larger amount of information lost, despite removal of the same proportion of sites (Figure 3D-F). However, as is demonstrated in the disparate axis scaling between clocks, there must be far more CpGs missing to exert a similar effect on the larger clocks than Hannum.

The user can choose, based on what is most appropriate for their hypothesis and data available, a set of CpG-based DNA methylation clocks to calculate. Then, either

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

using the manual functions of the form *calc[Clock-Name]*, or a user-specific character vector list of clocks, input to *calcUserClocks*, the appropriate values can be generated. These are ideally output by binding to an existing "phenotype" data frame supplied by the user containing relevant sample metadata.

To calculate each clock, an RData object is accessed containing CpG identity and weight information as supplied by the original authors. Each of these objects can be accessed using *data("[Clock-Name]_CpG(s)")*. A central repository where these data objects are readily accessible has two advantages. First, it makes the details of clock calculation transparent to the user. Second, it facilitates studies of clock CpG identities and their biological underpinnings.

For example, methylCIPHER allowed us to quantify the overlapping CpG identities of clocks within some of the categories identified in Figure 1. It is often discussed within the field that CpGs—at least those on the Illumina array technologies— change in a concerted, multi-collinear manner with age. This has motivated some of our prior approaches to clustering clock CpGs to ascertain underlying biological signals or changes [53, 54]. We find that the vast majority of CpGs selected by clocks do tend to be unique to those clocks, though a small subset of methylation sites are common across many clocks within categories (Figure 4A-B). Other observatons, such as the fact that EpiTOC2 is a subset of the original EpiTOC sites, or that DNAmTL is entirely unique in its CpG selections, are immediately obvious from this analysis (Figure 4B). However, despite sparse overlap in clock CpGs, it remains that these clocks' sex-adjusted age accelerations (i.e., residuals of regressing clock values onto sample age and sex) are typically quite correlated (Figure 4C).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In fact, it was recently reported that numerous combinations of CpGs can be used to train epigenetic clocks across the epigenome, a concept which arises primarily from this noted redundancy [55]. To further investigate the CpG identities of the clocks as distributed by the original authors in our data files, we repetitively retrained multiple well known epigenetic clocks. We performed this analysis in both a chronological aging clock (Hannum) and a tissue specific aging clock (Horvath Skin & Blood). We first generated an experimental design matrix (Figure 5A), which consists of 19x200 sample cells. Each cell is the result of one of 19 bootstrapped samples of the training data individuals drawn with replacement (for up to 6 draws of the same individual), and 1 of 200 versions of 10,000 CpGs drawn without replacement from the available probes on the 450K array. The concept of bootstrapped samples for model training was inspired by the original Hannum training method [12]. We used the original Hannum training dataset [12], and the publicly available training datasets (supplemental materials, table 1A) from the original Horvath Skin & Blood clock publication.

Given each experimental matrix cell, we retrain the epigenetic clocks in the sampled data, by applying elastic net penalized regression on chronological age. Elastic net regression was performed with 10-fold cross validation and a 0.5 ratio of LASSO and Ridge regression. All models were subsequently evaluated in an independent test dataset consisting of either whole blood methylation data [56] or skin and fibroblast datasets (supplemental materials, table 1B).

This analysis enabled comparison across randomly selected subsets of individuals and/or CpGs on retrained clocks. The correlations of predicted age and true age measures were visualized as density plots organized by sample bootstrap, for both

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hannum (Figure 5B) and Horvath Skin & Blood (Figure 5C). Training sample correlations were verified normal distributions with high degree of correlation, while independent test sample correlations for these models retained relatively high correlations. Medians of each density plot (small vertical bars) are compared to the original clock's correlation to age in the test dataset (large vertical bar). As is demonstrated in the case of Hannum, while some CpG subsets produce models that have correlation as low as 0.89 with chronological age, this is still a relatively high correlation. Thus, while access to some CpGs improves model performance, the improvement is modest. Furthermore, even when lists of CpGs are partially overlapped, the degree of overlap in selected CpGs is lower than expected (Figure 3D). Therefore, while some CpGs may contain relatively important information for age prediction, there may be significant redundancy in DNA methylation, allowing high model correlation and performance even with low shared identity of CpGs. It is also important to note that the original Hannum methylation age clock used CpG preselection [12] whereas here we are performing unsupervised selection. This may account for the overall slightly reduced performance. However, we find it more promising that the spread of resultant correlations in test data is relatively small, retaining correlations above 0.8 for the vast majority of models. This high redundancy amongst CpGs may largely explain why different clocks trained to predict the same outcome can have such sparse overlap in their composition.

Similar results are found for a tissue specific aging clock, Horvath Skin & Blood (Figure 3C,E). Here, we see that despite training and testing with half the sample size of the original clock, more than 50% of the models outperform the original clock's

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

prediction in the test dataset. It has been demonstrated that there are measurable differences in age related DNA methylation changes between tissues [57, 58]. Consequently, there may exist particular sets of CpGs which are essential to the function of for tissue specific aging clocks [49]. However, our results suggest that the majority of the CpGs have significant redundancy even for use in a tissue specific age predictor. Again, the selected CpGs do not have significant overlap between lists, despite them having significant overlap within a list across sample bootstraps (Figure 5E). Therefore, there may be many CpGs to select from for age prediction in tissue-specific contexts, though given the limitations of the current training method, they may not be truly tissue-specific merely single-tissue trained.

Beyond enabling characterization and investigation of the existing epigenetic clocks' mechanics, the current package enables efficient comparison of desired clocks with important biological phenotypes, biomarker data, or other sample metadata (Figure 6). Due to the rapid calculation of DNAm-based clocks in new data with automatic appending of results to existing phenotype/ sample metadata, researchers can quickly search for associations between outcomes or biomarkers, and clock scores.

Typically, the user will want to residualize the clock scores rather than using raw scores, to assess the effects of age acceleration. We have left this step up to the user, as details of correction for batch, sex, race, or other features tend to be dataset, and use specific. For instructions on how to approach calculating age residuals, please refer to the included tutorial or wiki on our GitHub distribution repository.

We find that we can rapidly uncover interesting results, such as univariate associations between the cluster of accelerated mitotic clocks and cardiovascular

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

outcomes in the Framingham heart study (FHS) data (Figure 6A). Alternatively, we find that the acceleration of BMI and Alcohol clocks by McCartney et al. [30] well surpass the univariate associations found by other clock accelerations to cancer types and stages in a subset of TCGA data (Figure 6B).

In nearly all cases, we do not recommend that the user calculate all clocks available to them, as this introduces significant multiple testing. We have previously shown that clocks can be driven by similar information content [53], highlighting the potential utility of clustered-clock approaches. Typically the researcher would be best served by selecting a hypothesis-driven subset of clocks: This would look like a single clock tested from within a cluster (e.g. one mitotic clock) or calculating and reporting all clocks and their agreement. However, we aim to show the ease with which few or all clocks can be assessed using the *methylCIPHER* package, as well as to provide resources to guide those decisions according to the high degree of correlation between clocks, particularly relative to those of the traits of interest.

Conclusions

The current software is an important compendium of clocks currently distributed through a wide variety of means. This reduces impediments to users, both in gathering the data for calculation, and ensuring reproducible and accurate calculation. Further, prior decentralized reporting and distribution of epigenetic clocks has led to the potential for researchers to inadvertently conduct significant multiple testing, potentially without proper correction: This can even occur throughout the course of a project in which the researcher becomes aware of, and iteratively tests, additional epigenetic clocks.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Through the provision of standardized clock calculation functions, and tools to rapidly investigate options available to the user, we aim to improve uptake of epigenetic clocks while enhancing the reproducibility of DNAm clock-based studies in the future. Further, as the current package is installed locally to one's personal computer or computing cluster, it is possible to rapidly calculate several epigenetic clocks, even in protected data. We intend to expand the clocks contained in this package in the future: (1) The addition of future human DNAm (regression) based clocks to the present researcher's toolkit will be essential; (2) Availability of mammalian arrays [34] will spur the use of similarly implemented epigenetic clocks for nonhuman vertebrates, and should also be included here; (3) The online tutorial provided in the currently discussed Github repository will be further expanded according to developing practice and user suggestions for standard features of epigenetic clock-based disease and trait analysis. Due to their different operating requirements and less standardized implementation, other forthcoming methods of epigenetic clock calculation are unlikely to be housed within this package, but we will direct users to their own sources using our wiki and tutorial pages. These include, but are not limited to, deep learning-based clock approaches [35, 59], the next generation of low-noise clocks referred to as PC Clocks [36], and single cell epigenetic clocks approaches [33].

Availability and Requirements

Project name: methylCIPHER

Project home page: github.com/MorganLevineLab/methylCIPHER

Operating system(s): tested on Mac OS 11+, with and without M1 chip, Windows

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Programming Language: R 3.6.1+

Other requirements:

License: TBD

Any restrictions to use by non-academics: TBD

List of abbreviations

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data necessary for clock calculation is provided through the distributed package. Datasets used for figures 4 and 6 were provided by the Framingham Heart Study (FHS) Offspring Study Cohort and The Cancer Genome Atlas (TCGA). The data used for Figures 4C and 6A in the current publication are based on the use of study data downloaded from the dbGaP web site, under phs000724. TCGA data (Figure 6B) can be downloaded via the GDC Data Portal (https://portal.gdc.cancer.gov).

Competing interests

The authors declare no competing interests.

Funding

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Authors' contributions

KLT was responsible for conception, packaging of code, and writing the present manuscript. AHC, ZL, and MEL contributed code and suggestions for clocks to include within the package, and reviewed the manuscript.

Acknowledgements

All figures were created in whole or in part using Lucidchart.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Oblak L, van der Zaag J, Higgins-Chen AT, Levine ME, Boks MP. A systematic review of biological, social and environmental factors associated with epigenetic clock acceleration. Ageing Research Reviews. 2021;69:101348.
2. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature Reviews Genetics. 2018;19:371–84.
3. Seale K, Horvath S, Teschendorff A, Eynon N, Voisin S. Making sense of the ageing methylome. Nature Reviews Genetics. 2022;0123456789.
4. Kabacik S, Lowe D, Fransen L, Leonard M, Ang SL, Whiteman C, et al. The relationship between epigenetic age and the hallmarks of ageing in human cells. Nature Aging. 2022. https://doi.org/10.1038/s43587-022-00220-0.
5. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. Mol Cell. 2018;71:882–95.
6. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: Challenges and recommendations. Genome Biology. 2019;20:1–24.
7. Bocklandt S, Lin W, Sehl ME, Sánchez FJ, Sinsheimer JS, Horvath S, et al. Epigenetic Predictor of Age. PLOS ONE. 2011;6:e14821.
8. Vidal-Bralo L, Lopez-Golan Y, Gonzalez A. Simplified assay for epigenetic age estimation in whole blood of adults. Frontiers in Genetics. 2016;7 JUL:1–7.
9. 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 Biology. 2014;15.
10. Garagnani P, Bacalini MG, Pirazzini C, Gori D, Giuliani C, Mari D, et al. Methylation of ELOVL2 gene as a new epigenetic marker of age. Aging Cell. 2012;11:1132–4.
11. Zbieć-Piekarska R, Spólnicka M, Kupiec T, Makowska Z, Spas A, Parys-Proszek A, et al. Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. Forensic Science International: Genetics. 2015;14:161–7.
12. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda SV, et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Molecular Cell. 2013;49:359–67.
13. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology. 2013;14:3156.
14. 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.
15. Zhang Y, Wilson R, Heiss J, Breitling LP, Saum KU, Schöttker B, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. Nature Communications. 2017;8.
16. 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 ex vivo studies. Aging. 2018;10:1758–75.
17. 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 Medicine. 2019;11:54.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18. Bohlin J, Håberg SE, Magnus P, Reese SE, Gjessing HK, Magnus MC, et al. Prediction of gestational age based on genome-wide differentially methylated regions. Genome Biology. 2016;17:1–9.
19. Knight AK, Craig JM, Theda C, Bækvad-Hansen M, Bybjerg-Grauholm J, Hansen CS, et al. An epigenetic clock for gestational age at birth based on blood methylation data. Genome Biology. 2016;17:1–11.
20. Mayne BT, Leemaqz SY, Smith AK, Breen J, Roberts CT, Bianco-Miotto T. Accelerated placental aging in early onset preeclampsia pregnancies identified by DNA methylation. Epigenomics. 2017;9:279–89.
21. Lee Y, Choufani S, Weksberg R, Wilson SL, Yuan V, Burt A, et al. Placental epigenetic clocks: Estimating gestational age using placental DNA methylation levels. Aging. 2019;11:4238–53.
22. McEwen LM, O'Donnell KJ, McGill MG, Edgar RD, Jones MJ, MacIsaac JL, et al. The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. Proc Natl Acad Sci U S A. 2020;117:23329–35.
23. Yang Z, Wong A, Kuh D, Paul DS, Rakyan VK, Leslie RD, et al. Correlation of an epigenetic mitotic clock with cancer risk. Genome Biology. 2016;17:1–18.
24. Youn A, Wang S. The MiAge Calculator: a DNA methylation-based mitotic age calculator of human tissue types. Epigenetics. 2018;13:192–206.
25. Teschendorff AE. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. Genome Medicine. 2020;12:56.
26. Lu AT, Seeboth A, Tsai PC, Sun D, Quach A, Reiner AP, et al. DNA methylation-based estimator of telomere length. Aging. 2019;11:5895–923.
27. 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.
28. 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.
29. Belsky DW, Caspi A, Corcoran DL, Sugden K, Poulton R, Arseneault L, et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. Elife. 2022;11:e73420.
30. 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.
31. Shireby GL, Davies JP, Francis PT, Burrage J, Walker EM, Neilson GWA, et al. Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. Brain. 2020;:1–13.
32. Boroni M, Zonari A, Reis de Oliveira C, Alkatib K, Ochoa Cruz EA, Brace LE, et al. Highly accurate skin-specific methylome analysis algorithm as a platform to screen and validate therapeutics for healthy aging. Clinical Epigenetics. 2020;12:105.
33. Trapp A, Kerepesi C, Gladyshev VN. Profiling Epigenetic Age in Single Cells. Nature Aging. 2021;1:1189–201.
34. Arneson A, Haghani A, Thompson MJ, Pellegrini M, Kwon S bin, Vu H, et al. A mammalian methylation array for profiling methylation levels at conserved sequences. Nature Communications. 2022;13:1–13.
35. Galkin F, Mamoshina P, Kochetov K, Sidorenko D, Zhavoronkov A. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. Aging Dis. 2021;12:1252–62.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

508 36. Higgins-chen AT, Thrush KL, Wang Y, Kuo P-L, Wang M, Minteer CJ, et al. A computational
509 solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and
510 longitudinal tracking. bioRxiv. 2021.
511 https://doi.org/https://doi.org/10.1101/2021.04.16.440205.
512 37. Levine ME, Lu AT, Chen BH, Hernandez DG, Singleton AB, Ferrucci L, et al. Menopause
513 accelerates biological aging. Proc Natl Acad Sci U S A. 2016;113:9327–32.
514 38. Breitling LP, Saum KU, Perna L, Schöttker B, Holleczek B, Brenner H. Frailty is associated with
515 the epigenetic clock but not with telomere length in a German cohort. Clinical Epigenetics.
516 2016;8:1–8.
517 39. Simpkin AJ, Cooper R, Howe LD, Relton CL, Davey Smith G, Teschendorff A, et al. Are
518 objective measures of physical capability related to accelerated epigenetic age? Findings from a
519 British birth cohort. BMJ Open. 2017;7.
520 40. Beynon RA, Ingle SM, Langdon R, May M, Ness A, Martin RM, et al. Epigenetic biomarkers of
521 ageing are predictive of mortality risk in a longitudinal clinical cohort of individuals diagnosed
522 with oropharyngeal cancer. Clinical Epigenetics. 2022;14:1–13.
523 41. McCrory C, Fiorito G, Hernandez B, Polidoro S, O'Halloran AM, Hever A, et al. GrimAge
524 Outperforms Other Epigenetic Clocks in the Prediction of Age-Related Clinical Phenotypes and
525 All-Cause Mortality. Journals of Gerontology - Series A Biological Sciences and Medical Sciences.
526 2021;76:741–9.
527 42. Han LKM, Aghajani M, Clark SL, Chan RF, Hattab MW, Shabalin AA, et al. Epigenetic aging in
528 major depressive disorder. American Journal of Psychiatry. 2018;175:774–82.
529 43. Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock.
530 Journal of Infectious Diseases. 2015;212:1563–73.
531 44. Marioni RE, Shah S, McRae AF, Ritchie SJ, Muniz-Terrera G, Harris SE, et al. The epigenetic
532 clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936.
533 International Journal of Epidemiology. 2015;44:1388–96.
534 45. Perna L, Zhang Y, Mons U, Holleczek B, Saum KU, Brenner H. Epigenetic age acceleration
535 predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. Clinical
536 Epigenetics. 2016;8:1–7.
537 46. Fortin JP, Triche TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina
538 HumanMethylationEPIC array with minfi. Bioinformatics. 2017;33:558–60.
539 47. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to
540 preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14.
541 48. Zhou W, Triche Jr TJ, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA
542 methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Research.
543 2018;46:e123–e123.
544 49. Ori APS, Lu AT, Horvath S, Ophoff RA. A systematic evaluation of 41 DNA methylation
545 predictors across 101 data preprocessing and normalization strategies highlights considerable
546 variation in algorithm performance. bioRxiv. 2021;:2021.09.29.462387.
547 50. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of
548 Parkinson's disease patients. Aging. 2015;7:1130–42.
549 51. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al.
550 DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics.
551 2012;13.

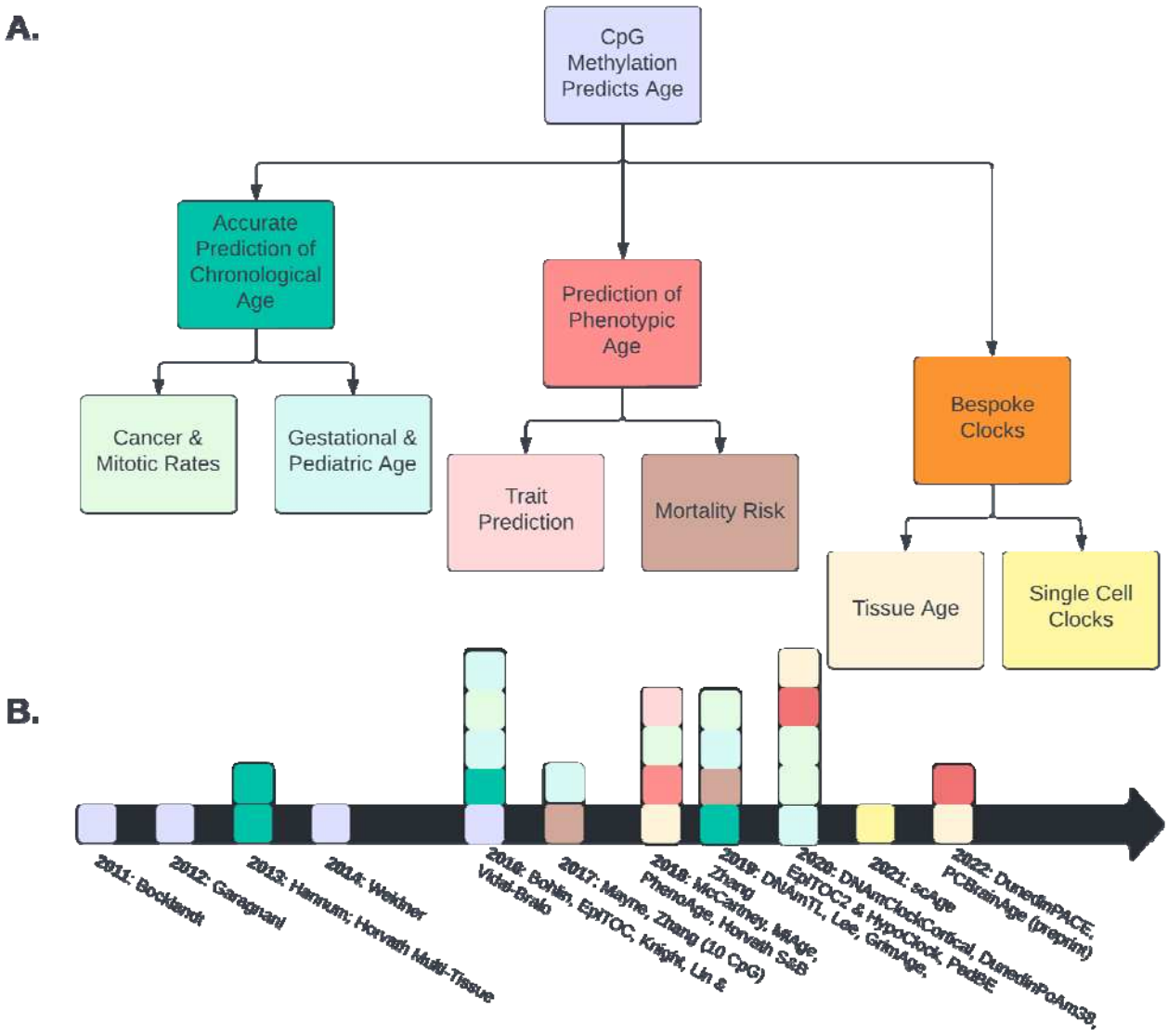
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

552 52. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An Investigation of
553 Coronary Heart Disease in Families: The Framingham Offspring Study. American Journal of
554 Epidemiology. 1979;110:281–90.
555 53. Liu Z, Leung D, Thrush K, Zhao W, Ratliff S, Tanaka T, et al. Underlying features of epigenetic
556 aging clocks in vivo and in vitro. Aging Cell. 2020; March:1–11.
557 54. Levine M, Higgins-Chen AT, Thrush K, Minteer CJ, Niimi P. Clock Work: Deconstructing the
558 Epigenetic Clock Signals in Aging, Disease, and Reprogramming. bioRxiv.
559 2022;:2022.02.13.480245.
560 55. Porter HL, Brown CA, Roopnarinesingh X, Giles CB, Georgescu C, Freeman WM, et al. Many
561 chronological aging clocks can be found throughout the epigenome: Implications for
562 quantifying biological aging. Aging Cell. 2021;20:1–13.
563 56. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MPM, van Eijk K, et al. Aging effects on DNA
564 methylation modules in human brain and blood tissue. Genome Biol. 2012;13:R97–R97.
565 57. Maegawa S, Hinkal G, Kim HS, Shen L, Zhang L, Zhang J, et al. Widespread and tissue specific
566 age-related DNA methylation changes in mice. Genome Research. 2010;20:332–40.
567 58. Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, Brooks JD, et al. Differential
568 DNA methylation with age displays both common and dynamic features across human tissues
569 that are influenced by CpG landscape. Genome Biology. 2013;14:1.
570 59. de Lima Camillo LP, Lapierre LR, Singh R. AltumAge: A Pan-Tissue DNA-Methylation
571 Epigenetic Clock Based on Deep Learning. bioRxiv. 2021;:2021.06.01.446559.
572

573 ## Figures, Tables, and Additional Files

574

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



575

576 **Figure 1: The Evolution and Diversity of Epigenetic Clocks.** (**A**) Epigenetic clocks are semantically
577 organized into key categories, and (**B**) individual clocks categorized and highlighted along a timeline.
578 Note that years with multiple clocks are colored in blocks from top to bottom according to the alphabetical
579 list.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

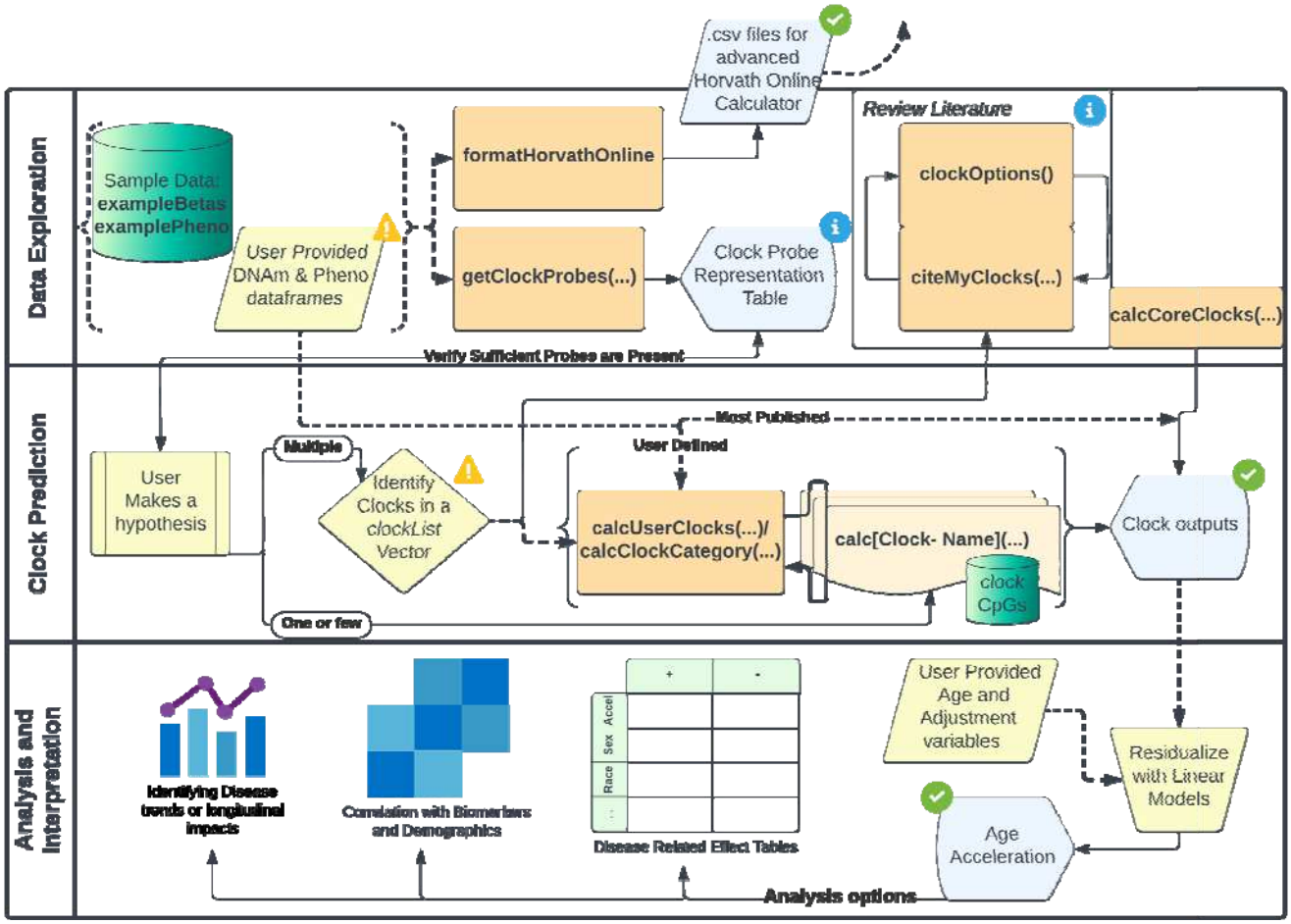


Figure 2: methylCIPHER Function Schema. A representational schematic of the functionality contained within methylCIPHER. Data is connected as inputs with dashed lines. User inputs are colored in yellow, with objects required to be supplied by the user highlighted with "!". Orange rectangles indicate functions exported by methylCIPHER for the user to run. Blue pentagonal boxes indicate outputs, with green checkmarks as endpoints. Green cylindrical objects are RData objects stored and accessible from within the package.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

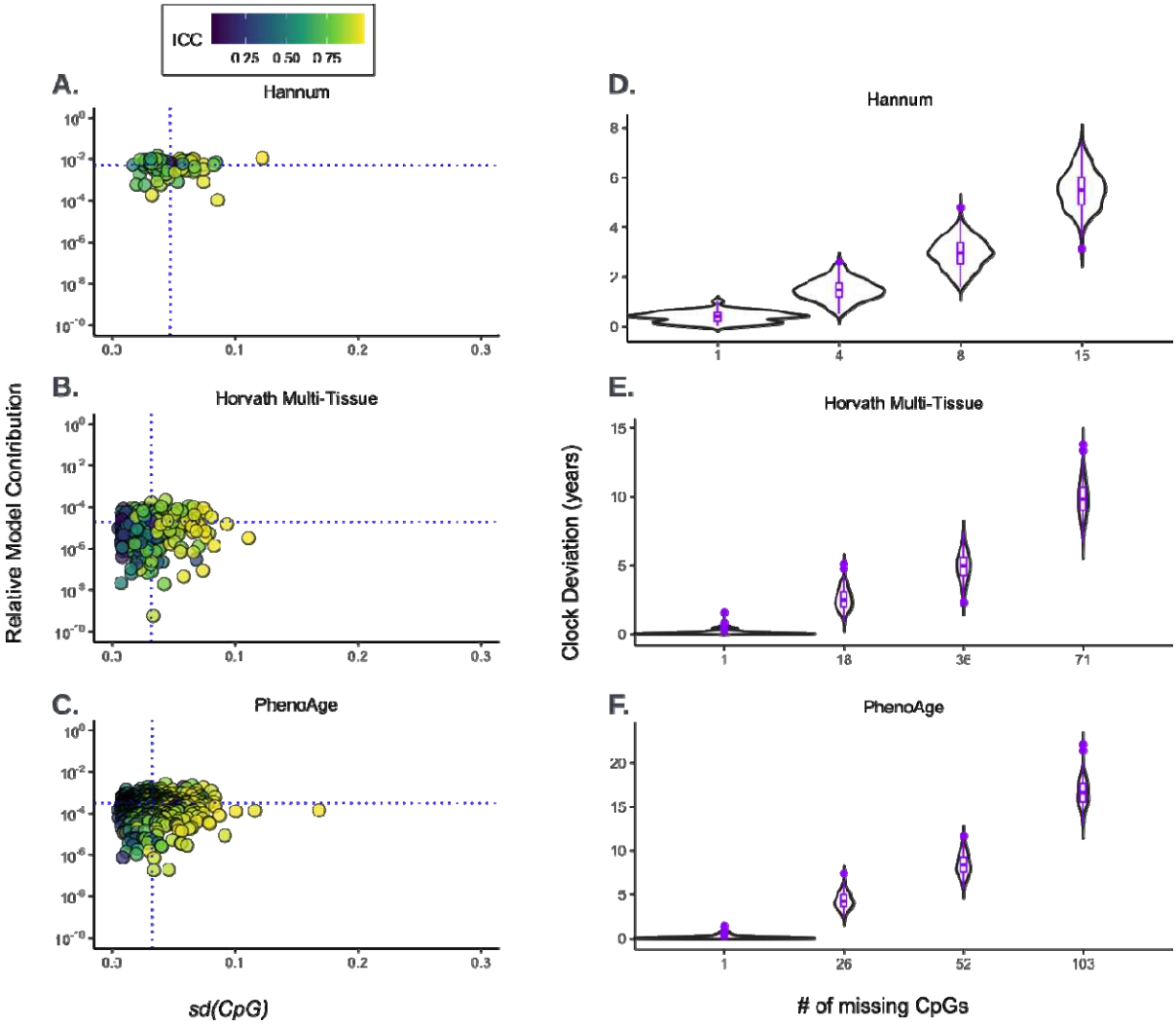
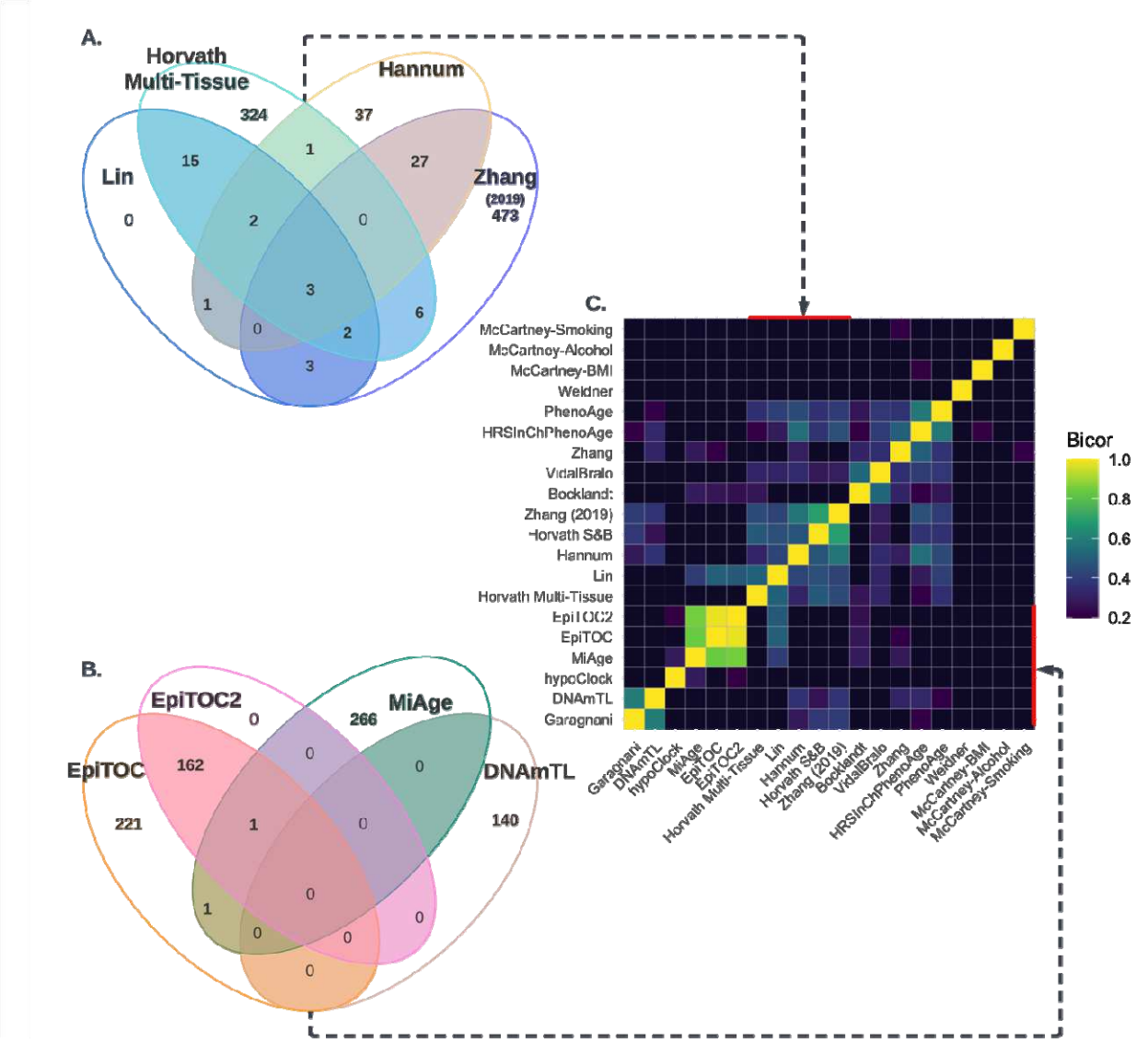


Figure 3: Impacts of Missing CpG Imputation on Key Clocks. Three representative clocks, Hannum (**A**, **D**), Horvath Multi-Tissue (**B**, **E**), and PhenoAge (**C**, **F**), were selected to assess the effects of imputation. Relative model contribution was calculated according to the absolute CpG weight in the regression model, multiplied by the standard deviation of the CpG in the Framingham Heart Study dataset. The model contributions of each CpG were plotted against the standard deviation of the CpGs, with the mean of each axis plotted as a blue crosshatch (**A**-**C**). Imputation of CpGs in the top right quadrant, with high standard deviation and high model contribution, will have a greater impact than CpGs in the other quadrants. Further, clocks whose CpGs extend further into this region will be more impacted by mean imputation effects. We repeatedly tested and plotted the effects of mean imputation on 0.1%,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

597 5%, 10% and 20% of CpGs selected at random (**D**-**F**). Due to varying sizes of the clocks, these

598 percentages represent varying numbers of missing CpGs for each of the clocks.



599

600 **Figure 4: Shared Signal Does Not Arise From Shared CpGs.** Epigenetic clocks were selected from the

601 categories of highly accurate chronological age clocks (A) and cancer & mitotic rate clocks (B) as defined

602 in Figure 1. Selection of CpG identity overlap was limited to 4 clocks per category, and in the case of

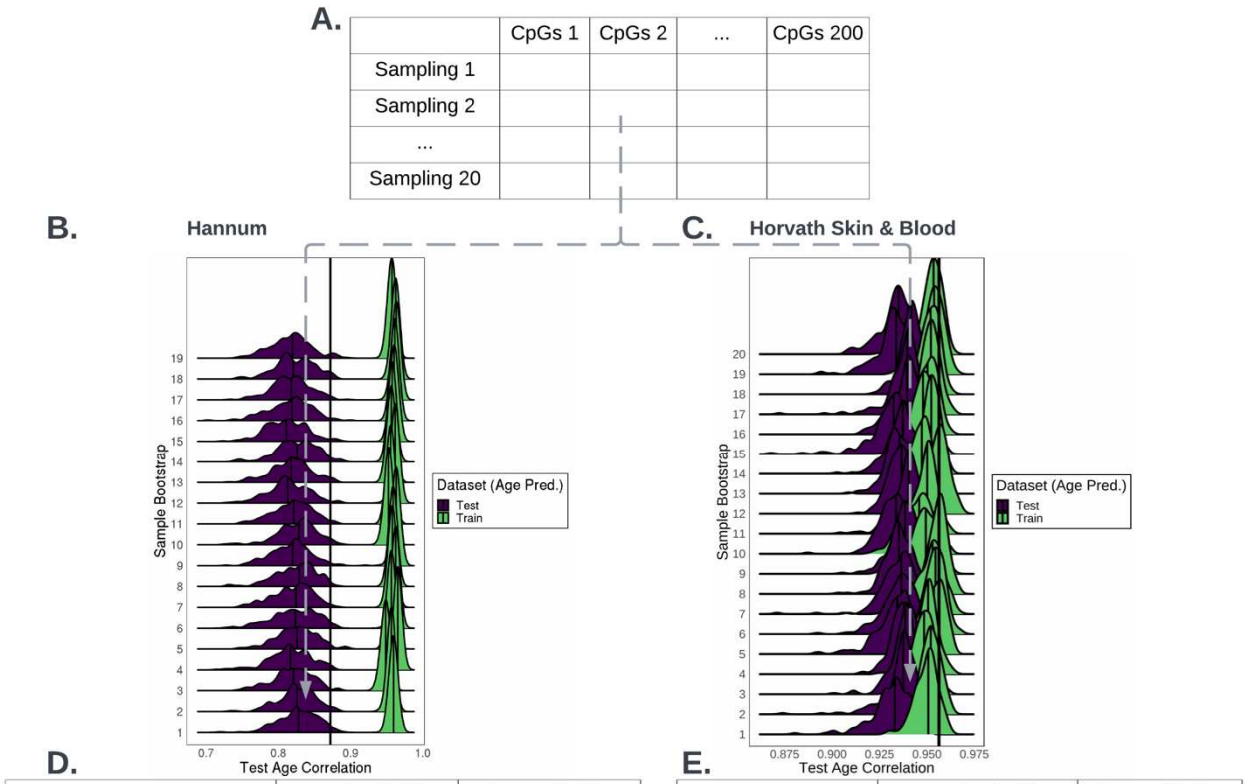
603 mitotic clocks hypoClock was not included as it was designed to select different CpGs from EpiTOC2. (C)

604 Sex-adjusted clock residuals were found in FHS and correlated according to biweight midcorrelation

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(thresholded at >0.2). While clock residuals can capture well-correlated information, their CpG overlaps within a cluster can be quite low.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



D.

	p-value (adj.)	Jaccard Index
Within Sample	0.11 ± 0.25	0.21 ± 0.12
Between Sample	$1x10^{-37}$ ± $7.23x10^{-36}$	0.17 ± 0.02

E.

	p-value (adj.)	Jaccard Index
Within Sample	0.15±0.34	0.28±0.19
Between Sample	$1.53x10^{-24}$ ± $1.23x10^{-22}$	0.20±0.03

Figure 5: Bootstrapping Methylation Data Demonstrates Methylation Redundancy. The experimental design matrix (**A**) demonstrates how both samples (19x) and CpGs (200 x 10k CpGs) were selected with random amounts of overlap in their sampled dimensions. Each sampling cell was used to retrain an elastic net regression model of the Hannum (**B**) or the Horvath Skin & Blood (**C**) clocks. The large vertical lines demonstrate the correlation of the originally developed clock in the whole test dataset. Smaller vertical lines in the purple density plots indicate the median age correlation of the within-sample bootstrap elastic net trials. As each experiment is allowed to overlap in CpGs to an extent, we use the p-value and Jaccard indices of modified gene set overlap tests to determine whether CpGs are repeatedly selected across bootstrapped CpG lists and within a given list across sampling sets, for both Hannum (**D**) and Horvath Skin & Blood (**E**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.13.499978; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

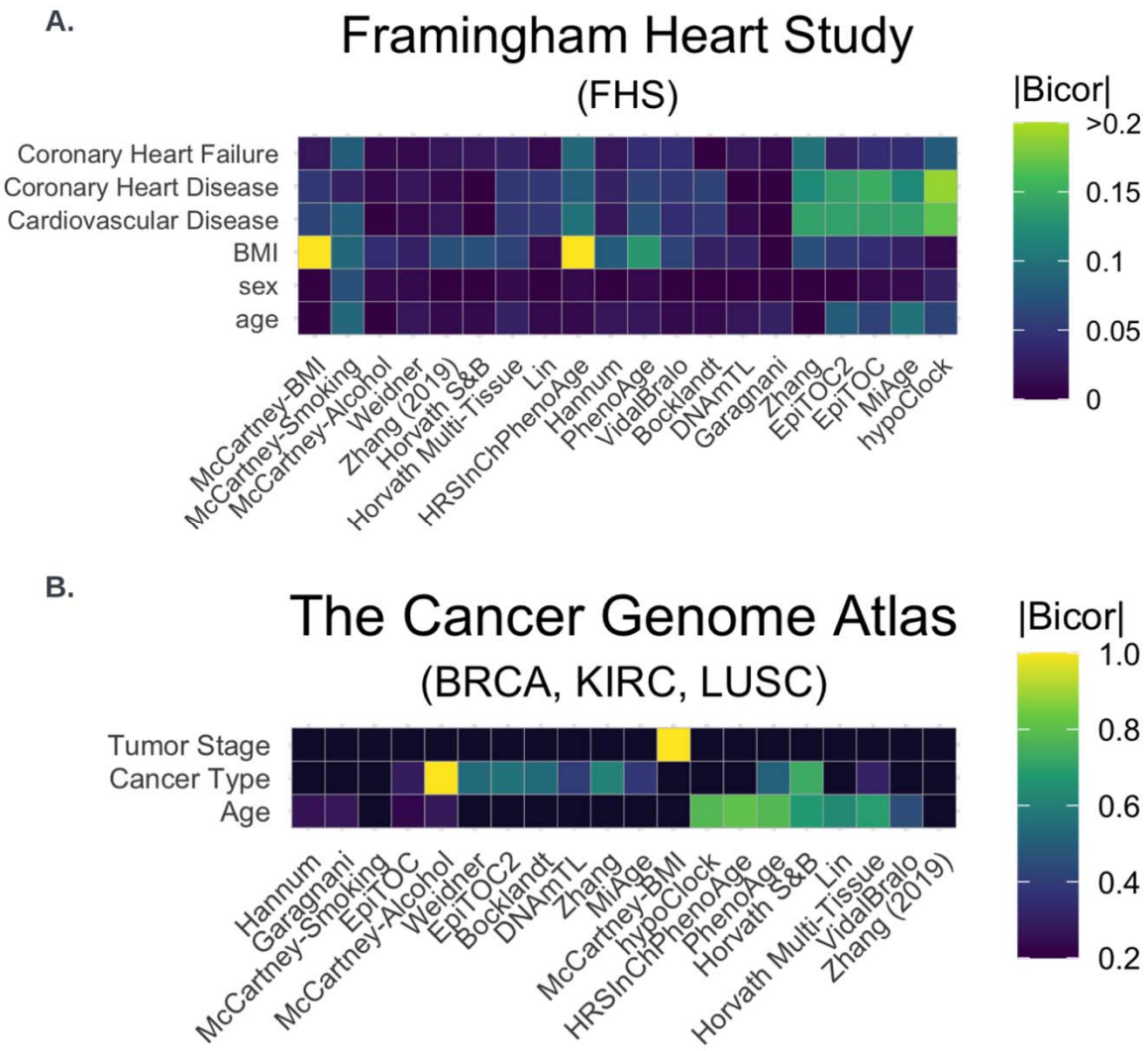


Figure 6: Clock Univariate Associations Capture Key Signal In Diverse Datasets. The available epigenetic clocks' simple age regression (regression of clock onto just age) were converted to z-scores across each dataset, as were samples metadata such as age, BMI, or sex. Additional sample traits were left as original variables as z-scores aren't realistic. The univariate associations were then described as the absolute biweight midcorrelation between the z-scored clock residuals and sample metadata. (A) In FHS data, there is clear associations between traits of interest and a cluster of mitotic clocks, whereas (B) in a few cancers in TCGA data, McCartney lifestyle/ trait predictors of Alcohol and BMI show the strongest correlations to traits of interest.