Age-Related Patterns of DNA Methylation Changes

Kevin Chen[1], Wenshu Wang[2], Hari Naga Sai Kiran Suryadevara[3], Gang Peng[4*]

[1]DeBakey High School, Houston, TX, USA; [2]Herricks High School, New Hyde Park, NY, USA; [3]Bioinformatics Core, Pennington Biomedical Research Center, Louisiana State University, Baton Rouge, LA, USA; [4]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA

*Corresponding Author

Abstract

DNA methylation is a dynamic process that adds methyl groups to DNA molecules, which plays an important role in gene expression regulation, genome stability, and aging. Epigenetic clocks developed over the past decade have used DNA methylation values at specific CpG sites combined with factors such as environmental influences and lifestyle factors to predict biological age with remarkable accuracy. However, these epigenetic clocks, built using conventional penalized regression models, have several limitations: (1) they assume consistent linear changes in methylation with age, even though methylation patterns may vary at different ages; (2) the selected CpGs in these clocks often lack specific biological significance; and (3) there is minimal overlap among the selected CpGs across different clocks, suggesting that only a small subset of a larger group of age-correlated CpGs is used to build an effective clock. Using data from 4,899 samples across 22 GEO datasets, we first analyzed the methylation patterns of 1,868 CpGs from 9 widely used epigenetic clocks and observed the expected consistently linear patterns in a subcluster. Next, we applied our own CpG selection method, which does not assume a lifelong linear correlation between DNA methylation and age. we divided the lifespan into overlapping age windows— [0, 20], [5, 25], [10, 30] …, [55, 75], [60, 80]— and identified CpGs that showed strong correlations with age within each specific window. Most CpGs were selected in young and old age windows, while few were selected in the middle-aged windows, indicating fast DNA methylation changes during child development and aging. Among the 12,903 unique CpGs selected within any of the windows, we performed clustering and uncovered from main patterns: (1) increase before 20 years old, (2) decrease before 20 years old, (3) increase before 20 and then decrease after 60, and (4) decrease before 20 and then increase after 60. Comparing this to the clock CpGs, both sets contained CpGs that decreased during adolescence. When examining gender differences, we noted that female samples had notably slowed methylation changes in old age windows compared to male samples, possibly due to hormonal changes during menopause. These findings suggest that age-related methylation changes are more complex than previously thought, with implications for refining epigenetic clocks and redefining the way researchers study aging. Future research should explore the biological mechanisms behind these non-linear and sex-specific patterns.

Background

DNA methylation is a biological process in which methyl groups are attached to DNA molecules, most commonly at CpG dinucleotides, and is functionally linked to gene expression regulation, cellular differentiation, and genome stability[1,2]. Over the past three decades, the relationship between DNA methylation and aging has been extensively explored[3-6], particularly with the development of DNA methylation clocks designed to estimate biological age[7-14]. Rather than focusing on individual CpGs, these epigenetic clocks are typically constructed from dozens or even hundreds of CpGs selected through feature selection methods such as lasso and elastic net[15,16]. Epigenetic clocks have demonstrated superior accuracy in predicting biological age compared to telomere length, transcriptomic, proteomic, and metabolomic predictors, as well as composite biomarkers[17]. The discrepancy between epigenetic age and chronological age, known as age acceleration, has been associated with mortality and the risk of various diseases[14,18-22], and it is frequently used as an indicator of individual health status. Studies have shown that a healthy lifestyle, including regular physical activity and a diet rich in lean meats, can mitigate epigenetic age acceleration[23,24].

Although epigenetic clocks are widely used and perform well in various aging-related studies, their underlying biological mechanisms remain largely unknown[25]. While the epigenetic age estimated by most clocks shows a strong correlation with chronological age[26,27], there is minimal overlap in the CpGs used across different clocks (**Supplementary Figure 1A**), making it challenging to determine the biological functions of the selected CpGs. Given that epigenetic clocks with distinct sets of CpGs yield similar performance, it suggests that many CpGs are correlated with age. These age-related CpGs may cluster into groups, with each group exhibiting similar patterns of age-related changes. Although different clocks use various subsets of age-related CpGs from these groups to calculate epigenetic age, they achieve comparable results because the CpGs within each group are highly correlated.

In this study, we aimed to identify patterns of DNA methylation changes associated with age. we downloaded DNA methylation data for 4,899 samples from Gene Expression Omnibus (GEO), analyzed CpGs from nine epigenetic clocks, and identified four unique patterns of age-related methylation changes. we then extended our analysis to additional CpGs and observed similar patterns across the full sample set, as well as within male and female groups separately.

Results

Investigation into Past Epigenetic Clock CpGs Suggests a Larger Set of Age-Related CpGs

We collected a total of 5056 samples from 22 publicly available datasets on the Gene Expression Omnibus (GEO), with inclusion criteria being non-cancer and profiled with either the Illumina 450K or Illumina EPIC arrays. we only analyzed samples ranging from 0 years to 80 years due to a lack of samples above 80 years, reducing our dataset to 4899 samples (2,158 males, 2642 females, 99 unknown gender) with mean age of 37.68 ± 25.19 years (**Supplementary Figure 2A, Supplementary Table 1**). The distribution of ages was relatively the same between male and female samples, but overall, it was not normally distributed; there were significantly more samples between 0 years and 20 years as well as between 40 to 65 years compared to between 20 to 40 years (**Supplementary Figure 2A**). The ethnic composition of the samples was 28.2% White, 5.7% Hispanic, 3.4% Black, 1.1% Asian, and 61.8% Undisclosed (**Supplementary Table 1**). The datasets encompassed a wide variety of tissues – blood, skin, adipose, buccal, immune (CD4+), airway epithelial, hepatocytes, and saliva – only interested in tissue-independent age-related patterns of DNA methylation (**Supplementary Table 2**). The datasets were not limited to healthy individuals; some included specific diseased conditions. For example, GSE61256 featured samples from obese individuals, GSE114134 included samples from those with allergies, GSE53740 contained samples from patients with neurological disorders such as progressive supranuclear palsy (PSP) and frontotemporal dementia (FTD), GSE106648 involved individuals with multiple sclerosis (MS), GSE32148 included those with Crohn's Disease, and GSE80261 featured samples from patients with fetal alcohol syndrome (**Supplementary Table 2**). 20 out 22 datasets profiled methylation using the Illumina 450K BeadChip, and the remaining two datasets – GSE114134 and GSE124076 – used the Illumina EPIC BeadChip; platform differences were corrected using the mLiftOver package. Because beta value matrices were directly accessed from the GEO datasets, differing preprocessing procedures led to slightly different sets of CpGs across all datasets. To account for this, only 367,013 CpG sites present in at least 75% (17 out of 22) datasets proceeded down the pipeline (**Supplementary Figure 2B**). Since each dataset contained samples in varying age ranges, we did not perform any inter-dataset batch effect correction that could potentially remove the genuine effects of age. The first step in our study was to investigate a set of 1,451 CpGs (present in >75% of datasets) taken from nine highly cited clocks: Hannum, Horvath, Levine (PhenoAge), Lu, McEwen (PedBE), Teschendorff, Shireby, Belsky. Hannum and Horvath were the pioneer epigenetic clocks, built in 2013, with

later clocks branching off and specializing (**Supplementary Table 3**). We observed minimal overlap among CpGs selected between the clocks, attributing this to the diverse tissue types and populations between each study, as well as the variability inherent in the penalized regression process used to build these clocks (**Supplementary Figure 1A**). Through this, we hypothesize that these clocks draw their features from a much larger set of age-related CpGs. To ensure a degree of robustness in our dataset, we plotted four CpGs that were common across multiple clocks (cg16867657: McEwen, Lu, Hannum, Shireby; cg06144905: McEwen, Horvath, Levine, Shireby; cg09809672, cg19722847: Wu, Horvath, Hannum, Levine) and observed consistent linear changes with age (**Supplemental Figure 1B**).

Variation in Selected age-associated CpGs (aaCpGs) Suggests Nonlinear Methylation Patterns

All further analysis proceeded with a reduced version of our dataset: a matrix B_A where $B_{i,j}$ is the average β value of CpG *i* at age *j*. we divided ages into 13 sliding windows ([0, 20], [5, 25], [10, 30] …, [55, 75], [60, 80]) and selected aaCpGs for each window independently with the following criterion: (i) |correlation coefficient between beta value and age| > 0.6 AND (ii) absolute mean beta change > 0.25. we tried various correlation coefficient and absolute beta change cutoffs, aiming for a final set of aaCpGs that is ~5% of the complete set of 450K array CpGs. we acknowledge the inherent information loss resulting from analyzing an averaged dataset rather than all samples, so we implemented an additional quality control filter (iii) large variation at less than 2 age points in any sliding window, defined as having standard deviation $D_{i,j}$ larger than $0.5\sqrt{B_{i,j}(1 - B_{i,j})}$, where $B_{i,j}$ is the mean beta value at CpG site *i* and age *j*. The first pass of the filter over our data notably resulted in 6,350 aaCpGs selected in the window [5, 25], 7,940 aaCpGs in [10, 30], 1313 aaCpGs in [50, 70], 2813 aaCpGs in [55, 75], and no aaCpGs between 25 years and 65 years (**Supplementary Table 4**). we desired representation of all sliding windows in our subsequent clustering step, so we modified the filtering to select the top 100 CpGs by absolute beta change out of the top 1000 CpGs by correlation coefficient in sliding windows with less than 100 aaCpGs initially. we ended up with 13,361 unique aaCpGs selected for all samples, again with remarkably more aaCpGs selected in the adolescent age windows [5, 25], [10, 30] and old age windows [50, 70], [55, 75] (**Table 1**). Ensuring that the aaCpGs weren't stochastically selected, an UpSet plot showed moderate overlap between the CpGs selected in adolescent sliding windows ([5, 25], [10, 30]), as well as old age sliding windows ([50 70], [55, 75]) (**Figure 2B**). From this, we hypothesized that there were two main periods of rapid methylation changes: adolescence and old age, with very little change in midlife. Major changes in adolescence are expected, due to the prior poor performance of

universal clocks (Hannum, Horvath) on young samples which led to the development of specialized clocks such as McEwen and Wu. However, there is limited evidence of significant methylation changes in older age groups, and no epigenetic clocks have been specifically developed for them. we explore these ideas further in the clustering analysis. we followed the same filtering procedures for B_M and B_F, average beta matrices for only male samples and female samples, respectively. 22,545 aaCpGs were selected for male samples, and 8,463 for female samples, once more with the majority of aaCpGs being in adolescence and old age windows (**Table 1**). UpSet plots again revealed overlap in aaCpGs for adolescent and old age windows (**Supplementary Figure 3**). Given that males are known to age faster than females, it is unsurprising that more aaCpGs were identified in male samples compared to female samples. 5,591 aaCpGs were present in all samples aaCpGs, male samples aaCpGs, and female samples aaCpGs, but only 38 of those were found among the previously investigated clock CpGs (**Figure 2A**). we attribute this to the stringent filtering criteria applied within a smaller sliding window, whereas epigenetic clocks were designed to select CpGs that exhibit gradual changes across the entire age range. As a result, our method prioritizes CpGs with stronger correlations within narrower age windows, reducing the likelihood of overlap with the broader, more gradual age-correlated CpGs used in clock models. This further supports our hypothesis of a wider set of age-related CpGs that may not have consistently linear relationships with age.

Clustering Analysis Reveals Novel Aging-Related Methylation Patterns

We performed clustering, first partitioning using k-means then hierarchically clustering on each subcluster, to identify age-related patterns of methylation changes. Since clock CpGs are already extensively validated throughout literature, we used them as a reference to compare to the patterns discovered in our aaCpGs. When clustering on matrices of the aaCpGs, we removed overlapping clock CpGs to remove any bias present. we scaled the matrices, aiming to preserve patterns of methylation changes with age regardless of the baseline mean methylation. To decide on the number of clusters, we examined patterns in four, five, and six clusters. Since there were no new patterns visualized in five or six clusters, we proceeded with analyses on four clusters (**Supplementary Figure 5, Supplementary Figure 6**). Clustering the 1,451 clock CpGs revealed the following patterns: cluster 1 – consistently increasing; cluster 2 – decreasing, increasing, then decreasing again (a periodic trend); cluster 3 – decreasing in adolescence, then steady; cluster 4 – no clear pattern (**Figure 3A**). In cluster 1, consistently increasing methylation with age was an expected finding due to the nature of the selection process for past clocks, including the well-established age biomarker cg16867657. Cluster 3, decreasing in

adolescence, was significantly enriched with CpGs from the McEwen (PedBE) clock (adjusted p-value: 0.01), also an expected finding (**Figure 2C**). The unclear patterns in cluster 4 can likely be attributed to noise from the penalized regression feature selection process.

In our selected aaCpGs, we observed clear patterns: cluster 1 – increase in adolescence then steady; cluster 2 – slight increase in adolescence then decrease in old age; cluster 3 – decrease in adolescence then steady; cluster 4 – slight decrease in adolescence then increase in old age (**Figure 2B**). aaCpGs selected in the windows [5, 25] and [10, 30] were significantly enriched in clusters 1 and 3, and aaCpGs selected in all other windows were significantly enriched in clusters 2 and 4 (**Figure 2D)**. When examining gender differences, we found notably more extreme increases and decreases in methylation at later ages in males compared to females (**Figure 3A-B**). This can possibly be attributed to menopause, a major hormonal change that leads to a cascade of side effects in the female body. Males are established to have a shorter lifespan, so a consistently higher epigenetic age acceleration (EAA) could be caused by these more extreme changes in methylation. Diving deeper into the inter-age analysis, we calculated EAA for all samples and observed elevated EAA for males across all sliding windows (**Figure 4C**). Medium effect sizes were present for midlife to old age sliding windows [25, 45] (Cohen's D: 0.54), [30, 50] (Cohen's D: 0.7), [35, 55] (Cohen's D: 0.78), [40, 60] (Cohen's D: 0.79), [45, 65] (Cohen's D: 0.69), and [50, 70] (Cohen's D: 0.56) (**Figure 4C**).

Genomic Distribution of aaCpGs supports a Whole-Genome Perspective of Aging

We conducted pathway enrichment analysis for our aaCpGs, and found a few pathways of interest, including some developmental and immune regulatory pathways (**Supplementary Figure 4**). To explore this further, we plotted the genomic distribution of all 28 million CpGs in our genome (truncated to 1 million randomly selected CpGs because of computational limitations), the starting set of 367,013 CpGs, and our selected aaCpGs, and found that the distribution was extremely similar (**Figure 2D**). From this, we can conclude that whatever enrichment we observed was due to already-present bias in CpG distribution and Illumina 450K probe bias. We believe that aging is a whole genome process, as evidenced by the multitude of CpGs changing dynamically.

Discussion

In this study, we identified approximately 13,000 CpGs associated with age across all samples. When samples were divided into male and female groups, more age-associated CpGs (aaCpGs) were detected in males than in females (**Table 1**). Although few aaCpGs overlapped

between the two groups (**Figure 2A, Table 1**), most aaCpGs selected within each age window exhibited the same direction of correlation with age for both males and females (**Figure 2C**). The correlation between DNA methylation and age was generally weaker in females than in males, explaining the higher number of aaCpGs observed in males. This may also account for the faster epigenetic age acceleration seen in males compared to females (**Figure 4C**)[8,26]. The aaCpGs detected in male and female were consistent with each other. This consistency supported the validity of our findings. While males and females exhibit similar patterns of DNA methylation change with age, the rate of change is slower in females.

For aaCpGs selected in each age window, most overlapping aaCpGs were found in adjacent windows. Notably, some overlapping aaCpGs were also observed between the younger age windows (5-30 years) and older age windows (50-75 years). When clustering aaCpGs, we identified two clusters of aaCpGs with methylation changes occurring in both early and later life stages. The CpGs in each cluster displayed similar patterns of change with age. These changes are nonlinear instead of the linear that was assumed in most epigenetic clocks (linear models), aligning with recent findings with multiple omics data[28]. However, we only found very large DNA methylation change after 60 years old. To estimate biological age, it may not be necessary to focus solely on specific CpGs included in traditional epigenetic clocks. Instead, CpGs from these clusters could serve as effective markers. For instance, if an individual shows low methylation levels in certain CpGs within cluster 2 in Figure 3B, they are likely over 60 years old.

The distribution of aaCpGs closely resembles the distribution of CpGs across the entire genome, except for a region on chromosome 6p (**Figure 2D**). This difference arises because the aaCpGs were selected from a set of 367,013 trimmed CpGs, with an enrichment on chromosome 6p due to the design of the 450K array. Overall, these aaCpGs appear to be randomly distributed across the genome, without specific enrichment in any region. On average, approximately one out of every 30 CpGs (all samples: 367,013/12,903 ≈ 28; males: 367,013/22,086 ≈ 17; females: 367,013/8,053 ≈ 46) shows a strong association with age. This finding suggests that DNA methylation changes associated with aging occur throughout the genome rather than in any specific locations. A limitation of this study is that this observation is based on the 367,013 trimmed CpGs; further validation with whole-genome DNA methylation sequencing (such as WGBS or Nanopore) is necessary to confirm these results.

When we clustered the CpGs used in epigenetic clocks along with aaCpGs, these aaCpGs grouped into several distinct clusters, each exhibiting unique patterns of DNA methylation

changes with age. Notably, within the aaCpGs, we identified two clusters with significant methylation changes occurring before age 20, while two other clusters showed substantial changes both before age 20 and after around age 60 (**Figure 3**). In contrast, clustering the CpGs from epigenetic clocks did not reveal clusters with large changes after age 60, as observed in the aaCpGs. This discrepancy may be due to limitations in the training datasets and statistical models used for epigenetic clocks, which often contain fewer samples over age 70 (as seen in our data, **Supplementary Figure 2**), leading to greater weight being applied to age ranges with more samples during regression. This bias is also reflected in the distribution of epigenetic age acceleration (EAA), which centers around zero for ages 20-60 but shifts to less than zero beyond age 65 (**Figure 4C**). Most epigenetic clocks focus on DNA methylation changes over a broad age range, particularly from 20 to 100 years, which may result in the exclusion of CpGs with methylation changes occurring within shorter, specific age periods.

In summary, we identified CpGs that are highly correlated with age, with more detected in males, likely due to faster epigenetic age acceleration. While the aaCpGs showed minimal overlap between sexes, similar age-related methylation patterns between sexes validated our findings. The aaCpGs were distributed broadly across the genome, with clustering revealing distinct age-related patterns: some showed significant changes before age 20, while others changed before 20 and after 60. These findings suggest that age-related methylation changes occur genome-wide and across different life stages. These findings may enhance our understanding of the relationship between biological age and DNA methylation and could contribute to the development of a more accurate epigenetic clock.

Methods

Data Availability

All datasets used in this study were publicly available on the Gene Expression Omnibus (GEO). Dataset accessions: GSE50660, GSE90124, GSE40279, GSE67705, GSE61256, GSE73103, GSE124366, GSE114134, GSE89253, GSE53740, GSE85568, GSE106648, GSE51057, GSE124076, GSE32148, GSE80261, GSE137495, GSE36054, GSE138279, GSE94734, GSE137502, GSE50759. we selected them with the following criterion: 1) large (>100) sample size, age metadata present, dataset already processed into beta value.

Data Extraction

Methylation data and patient metadata were systematically extracted and organized using Python libraries GEOparse, pandas, gzip, and math from the Gene Expression Omnibus (GEO).

β Value Matrix Generation

All further analysis was conducted using a condensed version of the datasets. For reference, *B* is a matrix of methylation level of CpGs at different ages, where $B_{i,j}$ is the β value of CpG *i* at age *j*, j ∈ 0, 1, 2, 3, ⋯, 79, 80. If there were multiple samples including CpG *i* at age *j*, we used mean β value, $B_{i,j} = \frac{1}{K}\sum_{k=1}^{K}\beta_{i,j}$, assuming there are *K* samples at age *j* who have data for CpG *i*. $B_{i,j,k}$ is the beta value for CpG *i* at age *j*, for sample *k*. *D* is a variance matrix, where $D_{i,j}$ is the standard deviation of beta values of CpG *i* at age *j*, j ∈ 0, 1, 2, 3, ⋯, 79, 80, $D_{i,j} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\beta_{i,j} - B_{i,j})^2}$. we generated 6 matrices to proceed to the rest of the pipeline: B_A (mean beta matrix, all samples included), D_A (variance beta matrix, all samples included), B_M (mean beta matrix, only male samples included), D_M (variance beta matrix, only male samples included), B_F (mean beta matrix, only female samples included), D_F (variance beta matrix, only female samples included). we linearly interpolated missing values (0.5%) in the final matrices.

Dataset Harmonization

Platform differences between the EPIC and legacy 450k datasets were corrected for using the mLiftOver package. GSE89253 and GSE50759 were converted from M-values to -values using the equation $\beta = \frac{2^M}{2^M+1}$.

Epigenetic Clocks

We investigated CpG sites used across highly cited clocks: McEwen (PedBE)[10], Wu[29], Hannum[8], Horvath (Multi-Tissue Age Estimator)[7], Levine (PhenoAge)[11], Belsky (DunedinPACE)[13], Teschendorff (EpiTOC2)[30], Shireby (Cortical Clock)[31], and Lu (DNAmTL)[32], extracting the sites from the corresponding paper's supplementary material.

Filtering of CpG sites

We selected a custom set of age-associated CpG sites (aaCpGs) across all datasets using a sliding window approach. A complete set of aaCpGs is defined as all unique CpG sites selected across all age windows. From ages 0-80, i used the following sliding windows: [0, 20], [5, 25], [10, 30], …, [55, 75], [60, 80]. CpG sites passed our filtering criterion when having (i) |correlation coefficient between β value and age| > 0.6 AND (ii) absolute mean beta change > 0.25 AND (iii)

large variation at less than 2 age points in any sliding window. Large variation at CpG site *i* and age *j* can be defined as having standard deviation $D_{i,j}$ larger than $0.5\sqrt{B_{i,j}(1 - B_{i,j})}$, where $B_{i,j}$ is the mean beta value. If the number of aaCpGs for an age window was less than 100, a secondary method was employed. aaCpGs had to fulfill the variation criteria (iii above), the top 1000 sites by correlation coefficient were taken, then the top 100 sites by absolute beta change were selected. Different sets of aaCpGs were taken for B_A, B_M, and B_F.

CpG Clustering

Clustering was performed using an integrated k-means and hierarchical clustering approach from R ComplexHeatmap 2.20.0. Default parameters were used with euclidean distance and complete linkage. Beta values were scaled to ensure age-related patterns were preserved while disregarding their original methylation ranges. The data was initially partitioned into *n* clusters using the k-means, with the number of clusters determined by the parameter *row_km*. Hierarchical clustering was conducted independently within each cluster with the clusters then being agglomerated. The optimal number of clusters, c, was determined through an iterative process. Beginning with c = 4 clusters, i incrementally increased n by one and evaluated the resulting cluster configurations. If the additional cluster revealed a distinguishable pattern, i proceeded by setting c = c + 1. This process was continued until no further discernible patterns emerged, at which point the iteration was terminated. Clusters were visualized by plotting all points $B_{i,j,c}$ for CpG sites *i* at age *j* in cluster *c*, with the overall trend of the cluster visualized by fitting a local polynomial regression (R splines 4.4.1). Data points at age 0 were consistently outliers, possibly due to varying embryonic methylation patterns, so they were excluded from the plots. Results for clock CpG sites were visualized alongside our aaCpGs, and results for male samples were visualized alongside results for female samples.

EAA Calculations

EAA was calculated using R methylCIPHER 0.2.0, using the function calcPedBE for samples aged 0 years to 20 years and the function calcHorvath1 for samples aged 21 years to 80 years. EAA was compared between genders in each of the 13 age windows using the cohens_d function from R rstatix 0.7.2.

Cluster Enrichment Analysis

We used Fisher's exact test to assess i) the enrichment of clock CpGs in each cluster and ii) the enrichment of CpGs selected in each age window in each cluster. P-values were adjusted using the Benjamini-Hochberg method to identify clusters with significant feature enrichment.

Genomic Distribution and Pathway Analysis

Pathway enrichment analysis was conducted using the gometh function from R missMethyl 1.38.0. The circus plot showing the distribution of CpGs was generated using R karyoploteR 1.30.0, with CpGs annotated to their genomic coordinates with R GenomicRanges 1.56.2.

Figure 1: Visual Workflow

A) DNA methylation data acquisition of our complete cohort. 8 different dataset types were collected, with blood being the most prevalent. B) Data filtering and preparation workflow. C) Matrix generation equation. Two different types of reduced matrices were generated: mean beta and standard deviation. D) 1 – Major clocks from which CpGs were selected from, 2 – Custom aaCpG filtering process. CpGs must fulfill 3 criteria: high correlation coefficient, high absolute beta change, and <2 age points with high standard deviation of beta values. E) Clustering workflow: First division using k-means, then intra-subcluster hierarchical clustering, and finally smooth figure plotting.

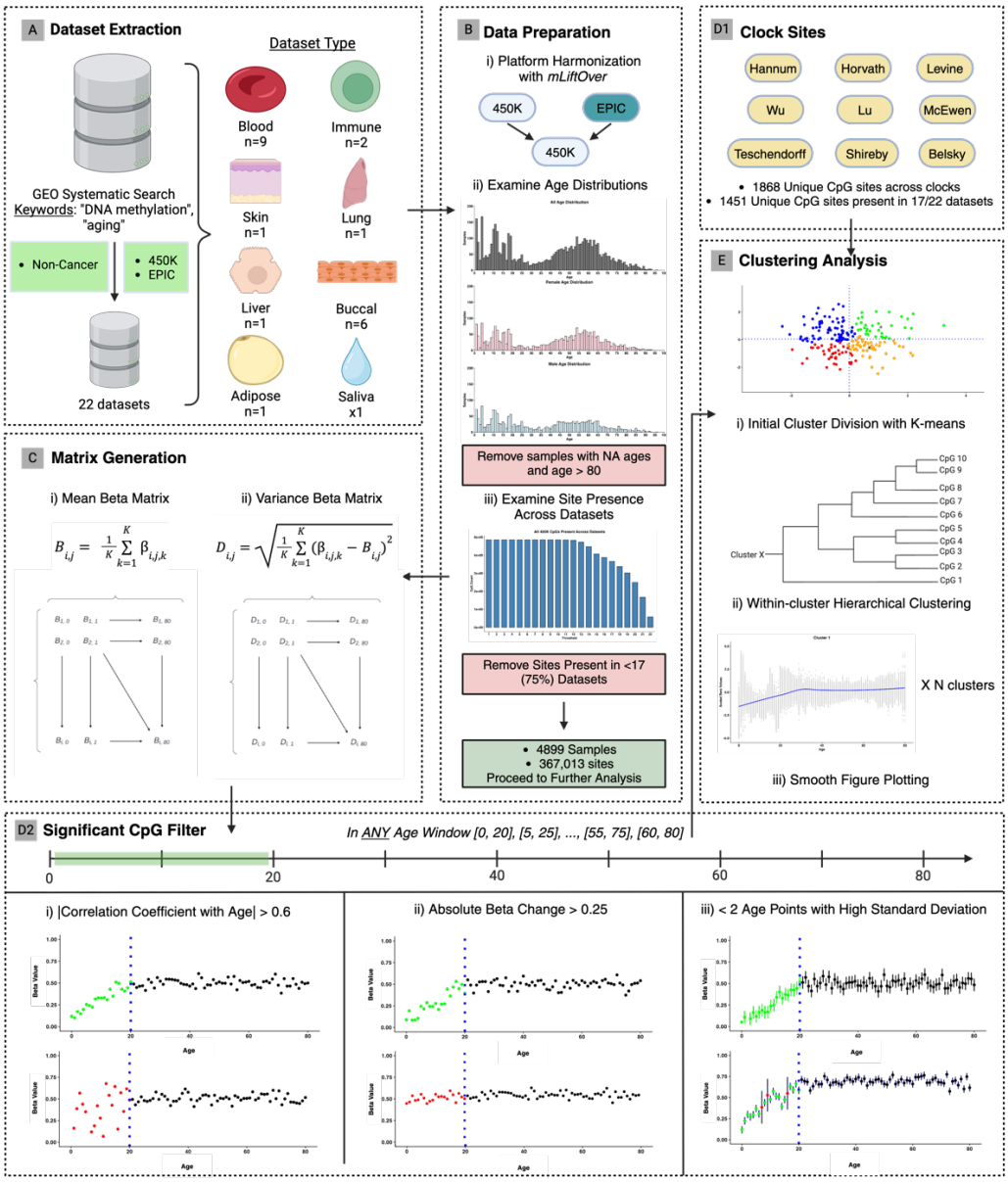


Figure 2: Filter Visualization

A) UpSet plot showing overlaps between selected aaCpGs, male aaCpGs, female aaCpGs, and clock CpGs. Intersection sizes are displayed on the top bar plot, and individual set sizes are displayed on the right-side bar plot. B) UpSet plot showing overlaps between aaCpGs selected in each of the 13 age windows. Intersection sizes are displayed on the top bar plot, and individual set sizes are displayed on the right-side bar plot C) Scatterplots visualizing filter results in each age window, x-axis: correlation of average beta values with age, male samples only; y-axis: correlation of average beta values with age, female samples only. D) Circos plot showing the distribution of all CpGs in the genome (randomly selected 1 million out of 28 million, outer ring), our 367,013 trimmed CpGs (middle ring), and our aaCpGs (inner ring)

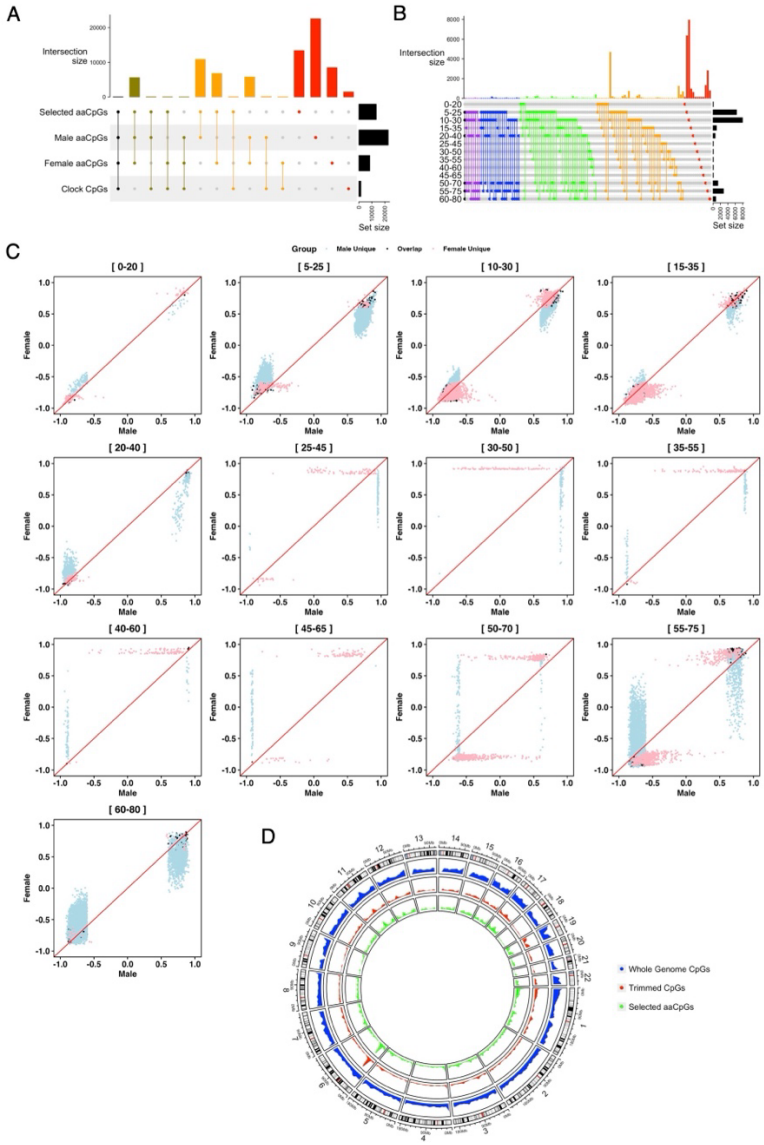


Figure 3: Heatmaps and Smooth Figures for Clock CpGs vs aaCpGs

A) Heatmap of clock CpGs with ages on the x-axis, dendrogram visualizing the clustering process on the left annotation, annotation of rows to corresponding clocks on the right (left). Scatterplot of chronological age on the x-axis by scaled average beta value (y-axis), blue smooth line plotted with local polynomial regression (right). B) Heatmap of aaCpGs, ages on the x-axis, dendrogram visualizing the clustering process on the left annotation (left) Scatterplot of chronological age on the x-axis by scaled average beta value (y-axis), blue smooth line plotted with local polynomial regression (right). C) Enrichment of clock CpGs in each cluster (refer to A), significant enrichment (adjusted p-value < 0.05, Fischer's exact test, Bonferroni correction) colored red for overrepresentation and blue for underrepresentation. D) Enrichment of aaCpGs selected in each window for each cluster (refer to B), significant enrichment (adjusted p-value < 0.05, Fischer's exact test, Bonferroni correction) colored red for overrepresentation and blue for underrepresentation.

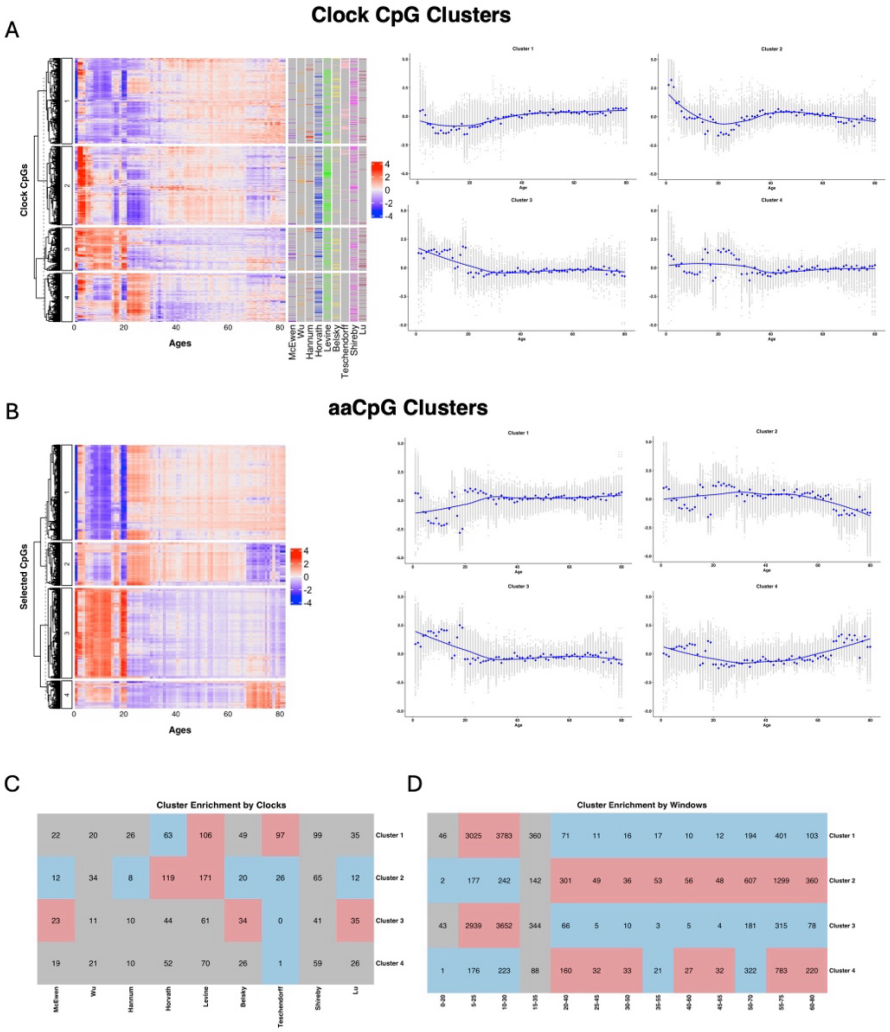


Figure 4: Heatmaps and smooth figures for male aaCpGs vs female aaCpGs

A) Heatmap of male aaCpGs, ages on the x-axis, dendrogram visualizing the clustering process on the left annotation (left). Scatterplot of chronological age on the x-axis by scaled average beta value (y-axis), blue smooth line plotted with local polynomial regression (right). B) Heatmap of female aaCpGs, ages on the x-axis, dendrogram visualizing the clustering process on the left annotation (left) Scatterplot of chronological age on the x-axis by scaled average beta value (y-axis), blue smooth line plotted with local polynomial regression (right). C) (left) Scatter plot of chronological age vs. EAA. (right) Box plot of EAA distribution differences between male and female, stratified by age window (x-axis).

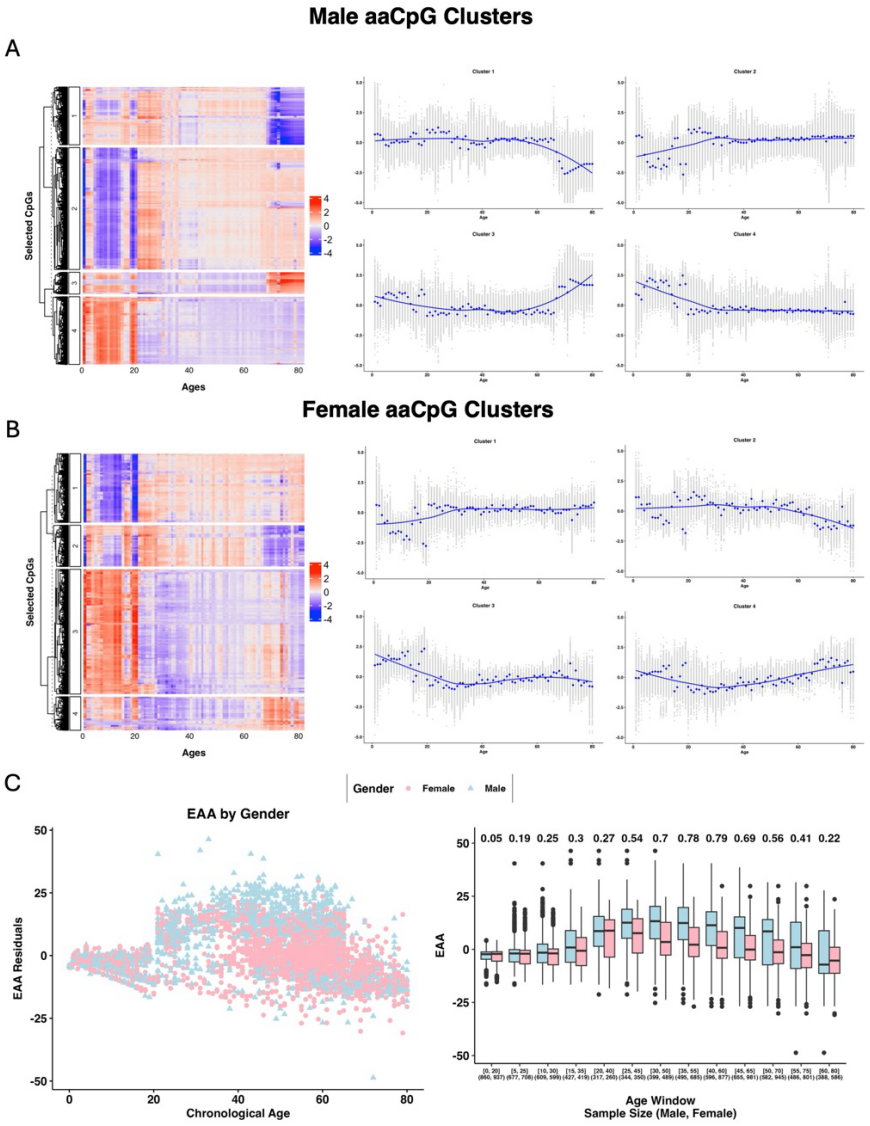


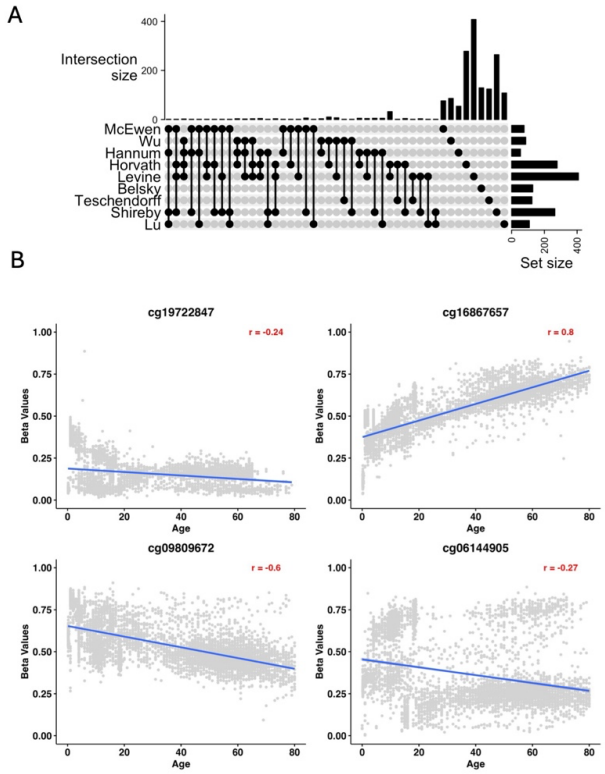
Table 1: Final Filter Results

Table displaying the number of aaCpGs selected in each age window, stratified by all samples, male sample, female sample, and male/female overlap.

Final Filter Results				
Age Window	# Significant CpG sites			
	All	Male	Female	Male/Female Overlap
[0, 20]	100	152	100	9
[5, 25]	6,350	13,224	1,588	1383
[10, 30]	7,940	9,151	5,195	3519
[15, 35]	941	639	1,224	249
[20 ,40]	604	1,154	196	97
[25, 45]	100	100	100	0
[30, 50]	100	100	100	0
[35, 55]	100	100	100	2
[40, 60]	100	100	100	3
[45, 65]	100	100	100	1
[50, 70]	1,313	295	857	28
[55, 75]	2,813	5409	1,688	1041
[60, 80]	766	6144	213	181
Total Unique Sites	13,361	22,545	8,463	5176

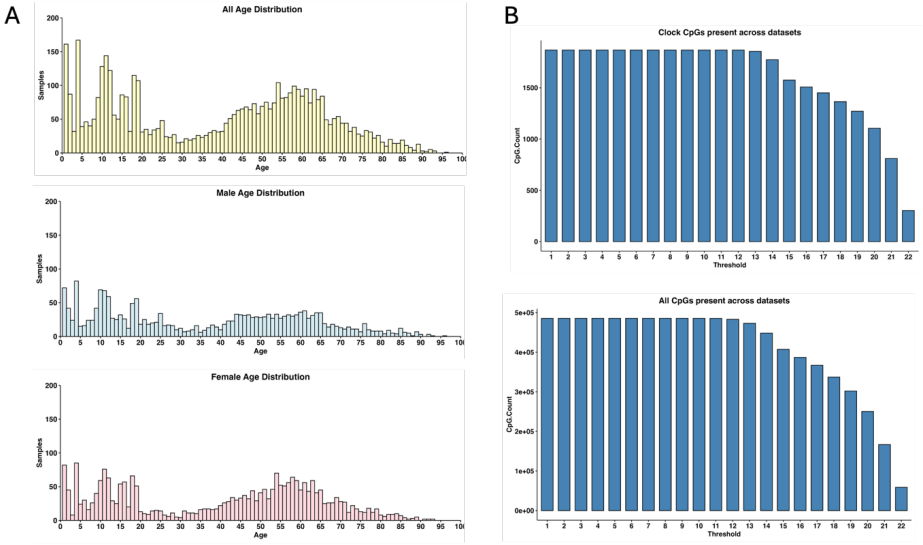
Supplementary Figure 1: Past Epigenetic Clock Investigation

A) Upset plot of overlap in clock CpGs. B) Beta values of CpGs present in multiple clocks



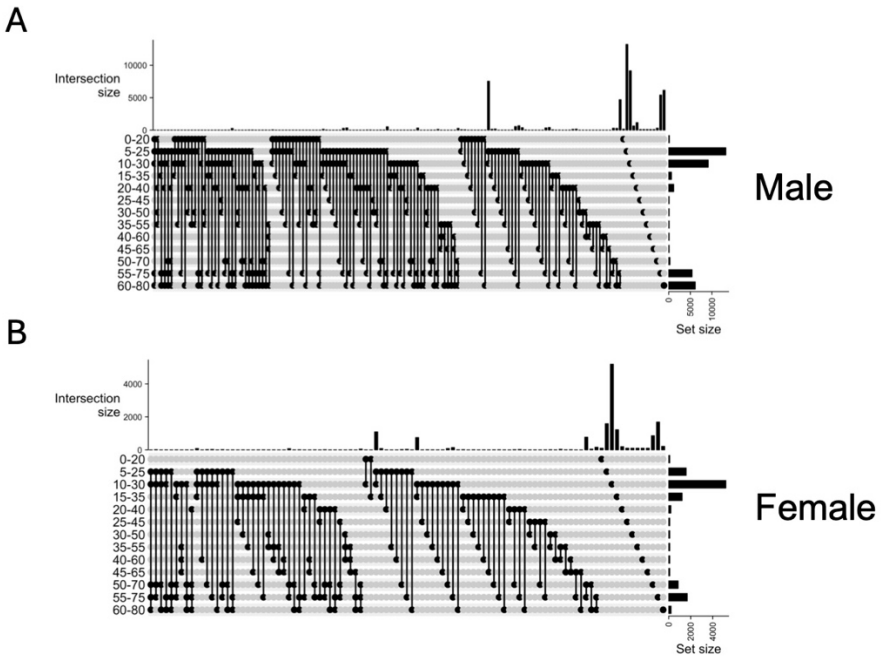
Supplementary Figure 2: Sample and Dataset Characterization

A) Histograms showing sample age distribution. B) Bar plots showing CpGs present across all datasets



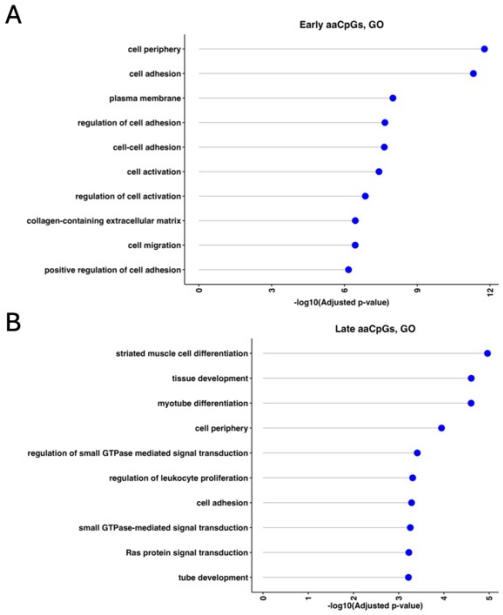
Supplementary Figure 3: Inter-Window aaCpG Overlap for Males and Females

A) UpSet plot showing overlaps between aaCpGs selected in each of the 13 age windows for male samples. B) UpSet plot showing overlaps between aaCpGs selected in each of the 13 age windows for female samples. Intersection sizes are displayed on the top bar plot, and individual set sizes are displayed on the right-side bar plot

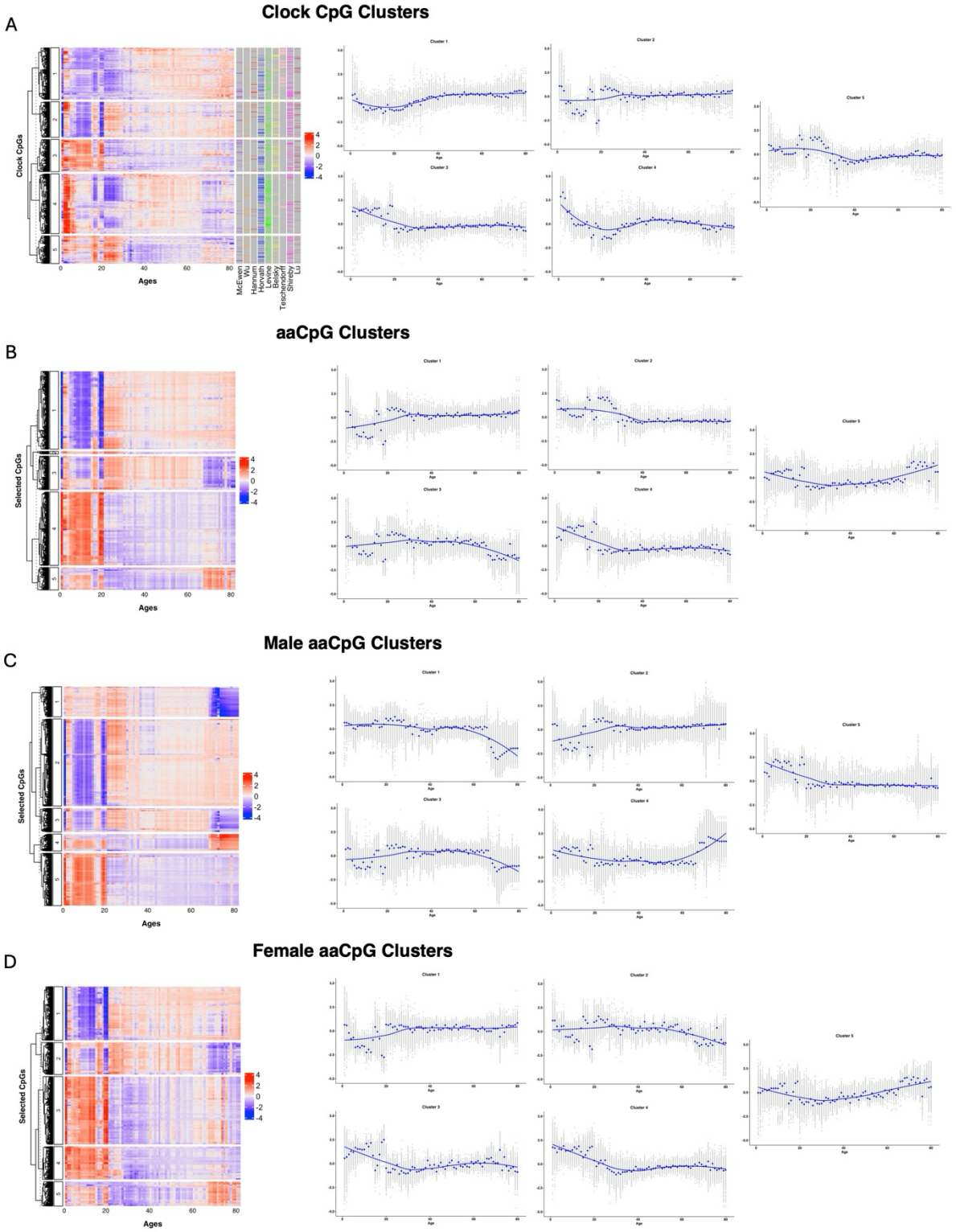


Supplementary Figure 4: Pathway Enrichment Analysis

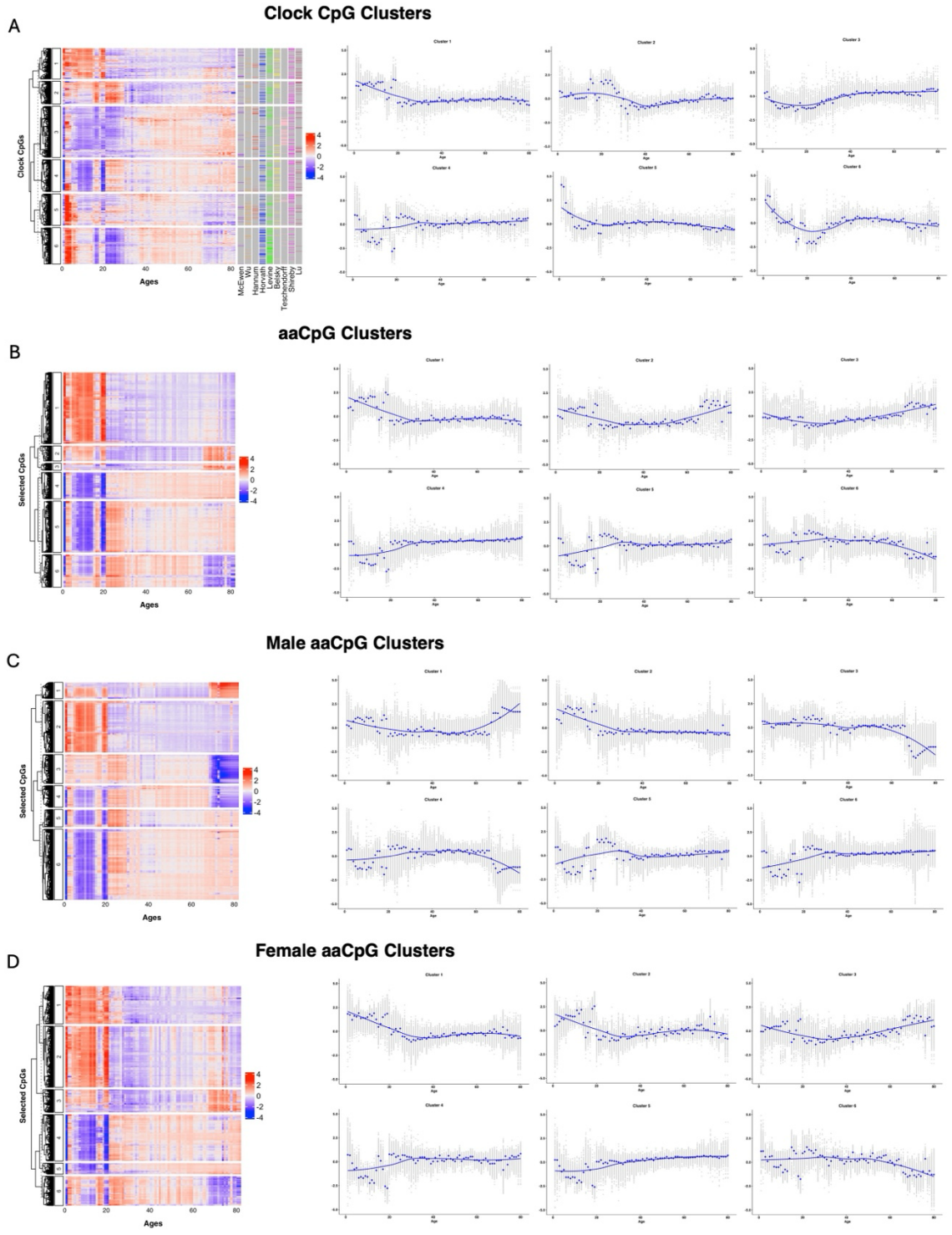
A) Lollipop plots for pathway enrichment analysis for aaCpGs changing in early ages. B) Lollipop plot for pathway enrichment analysis for aaCpGs changing in late ages.



Supplementary Figure 5: Heatmap and Smooth Figures for 5 Clusters



Supplementary Figure 6: Heatmap and Smooth Figures for 6 Clusters



Supplementary Table 1: Complete Sample Demographics

Table showing demographics of all samples in our datasets.

Variables		Mean ± SD or N (%)
		(n = 4899)
Age (years)		37.68 ± 25.19
Sex	Male	2158 (44.1)
	Female	2642 (53.9)
	Unknown	99 (2.0)
Race/Ethnicity	White	1381 (28.2)
	Black	166 (3.4)
	Hispanic	281 (5.7)
	Asian	54 (1.1)
	Undisclosed/Unknown	3017 (61.8)

Supplementary Table 2: Description of Datasets

Extra information characterizing our 22 datasets.

Description of Datasets						
	Tissue Type	Sample Size	Mean Age (±SD)	Disease Status	Gender (F; M; Missing)	Bead Chip
GSE50660	Peripheral Blood	464	55 (±6.7)	None	137; 327; 0	450K
GSE90124	Skin	322	59 (±9.2)	None	322; 0; 0	450K
GSE40279	Whole Blood	656	64 (±15)	None	338; 318; 0	450K
GSE67705	Whole Blood	284	46 (±8.9)	None	2; 282; 0	450K
GSE61256	Adipose	271	50 (±14)	Obesity	77; 60; 0	450K
GSE73103	Whole Blood	355	20 (±4.8)	None	141; 214; 0	450K
GSE124366	Buccal Epithelial	215	7.1 (±3.1)	None	103; 112; 0	450K
GSE114134	CD4+ T cells (naïve)	205	1.8 (±1.1)	Allergic	109; 96; 0	EPIC
GSE89253	CD4+ T cells	261	11 (±4.4)	None	98; 38; 0	450K
GSE53740	Peripheral Blood	384	65 (±15)	PSP[1] & FTD[2]	155; 130; 99	450K
GSE85568	Airway Epithelial Cells	200	39 (±12)	Asthma	79; 36; 0	450K
GSE106648	Peripheral Blood	279	41 (±11)	Multiple Sclerosis	77; 202; 0	450K
GSE51057	Peripheral Blood	329	52 (±7.1)	None	329; 0; 0	450K
GSE124076	Hepatocytes	567	39 (±19)	None	25; 31; 0	EPIC
GSE32148	Peripheral Blood	48	22 (±18)	Crohn's Disease	25; 23; 0	450K
GSE80261	Buccal Epithelial	216	11 (±3.4)	Fetal Alcohol Syndrome	119; 97; 0	450K
GSE137495	Buccal Epithelial	145	0.3 (±0.1)	None	68; 77; 0	450K
GSE36054	Peripheral Blood (leukocytes)	192	4.6 (±4.1)	None	55; 79; 58	450K
GSE138279	Saliva	65	9.7 (±1.7)	None	34; 31; 0	450K
GSE94734	Buccal Epithelial	178	11 (±0.9)	None	84; 94; 0	450K
GSE137502	Buccal Epithelial	179	18 (±0.3)	None	102; 77; 0	450K
GSE50759	Buccal Epithelial	96	9.0 (±6.2)	None	38; 58; 0	450K

Supplementary Table 3: Clocks Investigated

Metadata for the nine clocks investigated in our study.

Clocks Investigated				
First Author (Year Published)	# CpGs Investigated	# CpGs present in 17/22 datasets (%)	Tissues derived	Age Range
Hannum (2013)	71	54 (76%)	Whole Blood	19-101
Horvath (2013)	353	278 (79%)	27 tissues	0-101
Levine (2018)	513	408 (80%)	Whole Blood	21-100
Wu (2019)	111	86 (77%)	Peripheral Blood	0.75-17.7
Lu (2019)	140	108 (77%)	Whole Blood	53-73
McEwen (2020)	94	76 (81%)	Buccal Epithelial	0-20
Teschendorff (2020)	163	124 (76%)	Whole Blood	19-101
Shireby (2020)	347	264 (76%)	Brain Cortex	1-108
Belsky (2022)	173	129 (75%)	Blood	26-45

Supplementary Table 4: Pre-Filter Results

Table showing the number of aaCpGs selected in each window before implementing the secondary filter for top 100 CpGs by absolute beta change.

Pre-Filter Results			
Age Window	# Significant CpG sites		
	All	Male	Female
[0, 20]	99	152	95
[5, 25]	6,350	13,224	1,588
[10, 30]	7,940	9,151	5,195
[15, 35]	941	639	1,224
[20 ,40]	604	1,154	196
[25, 45]	0	8	0
[30, 50]	0	0	1
[35, 55]	0	0	0
[40, 60]	0	0	0
[45, 65]	0	0	0
[50, 70]	1,313	295	857
[55, 75]	2,813	5409	1,688
[60, 80]	766	6144	213
Total Unique Sites	12,903	22,086	8,053

References

1 Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23-38 (2013).

2 Mattei, A. L., Bailly, N. & Meissner, A. DNA methylation: a historical perspective. *Trends Genet* **38**, 676-707 (2022). https://doi.org/10.1016/j.tig.2022.03.010

3 Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* **102**, 10604-10609 (2005). https://doi.org/10.1073/pnas.0500398102

4 Bollati, V. *et al.* Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech Ageing Dev* **130**, 234-239 (2009). https://doi.org/10.1016/j.mad.2008.12.003

5 Bell, J. T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet* **8**, e1002629 (2012). https://doi.org/10.1371/journal.pgen.1002629

6 McCartney, D. L. *et al.* An epigenome-wide association study of sex-specific chronological ageing. *Genome Med* **12**, 1 (2019). https://doi.org/10.1186/s13073-019-0693-z

7 Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013). https://doi.org/10.1186/gb-2013-14-10-r115

8 Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* **49**, 359-367 (2013). https://doi.org/10.1016/j.molcel.2012.10.016

9 Lu, A. T. *et al.* Universal DNA methylation age across mammalian tissues. *Nat Aging* **3**, 1144-1166 (2023). https://doi.org/10.1038/s43587-023-00462-6

10 McEwen, L. M. *et al.* The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proc Natl Acad Sci U S A* **117**, 23329-23335 (2020). https://doi.org/10.1073/pnas.1820843116

11 Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* **10**, 573-591 (2018). https://doi.org/10.18632/aging.101414

12 Knight, A. K. *et al.* An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol* **17**, 206 (2016). https://doi.org/10.1186/s13059-016-1068-z

13 Belsky, D. W. *et al.* DunedinPACE, a DNA methylation biomarker of the pace of aging. *Elife* **11** (2022). https://doi.org/10.7554/eLife.73420

14 Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303-327 (2019). https://doi.org/10.18632/aging.101684

15 Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267-288 (1996).

16 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301-320 (2005).

17 Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological Age Predictors. *EBioMedicine* **21**, 29-36 (2017). https://doi.org/10.1016/j.ebiom.2017.03.046

18 Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol* **16**, 25 (2015). https://doi.org/10.1186/s13059-015-0584-6

19 Xiao, C. *et al.* Association of Epigenetic Age Acceleration With Risk Factors, Survival, and Quality of Life in Patients With Head and Neck Cancer. *Int J Radiat Oncol Biol Phys* **111**, 157-167 (2021). https://doi.org/10.1016/j.ijrobp.2021.04.002

20 Jain, P. *et al.* Analysis of Epigenetic Age Acceleration and Healthy Longevity Among Older US Women. *JAMA Netw Open* **5**, e2223285 (2022). https://doi.org/10.1001/jamanetworkopen.2022.23285

21 Faul, J. D. *et al.* Epigenetic-based age acceleration in a representative sample of older Americans: Associations with aging-related morbidity and mortality. *Proc Natl Acad Sci U S A* **120**, e2215840120 (2023). https://doi.org/10.1073/pnas.2215840120
22 Fransquet, P. D., Wrigglesworth, J., Woods, R. L., Ernst, M. E. & Ryan, J. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. *Clin Epigenetics* **11**, 62 (2019). https://doi.org/10.1186/s13148-019-0656-7
23 Quach, A. *et al.* Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging (Albany NY)* **9**, 419-446 (2017). https://doi.org/10.18632/aging.101168
24 Fitzgerald, K. N. *et al.* Potential reversal of epigenetic age using a diet and lifestyle intervention: a pilot randomized clinical trial. *Aging (Albany NY)* **13**, 9419-9432 (2021). https://doi.org/10.18632/aging.202913
25 Kabacik, S. *et al.* The relationship between epigenetic age and the hallmarks of ageing in human cells. *Nature Aging*, 1-10 (2022).
26 Crimmins, E. M., Thyagarajan, B., Levine, M. E., Weir, D. R. & Faul, J. Associations of Age, Sex, Race/Ethnicity, and Education With 13 Epigenetic Clocks in a Nationally Representative U.S. Sample: The Health and Retirement Study. *J Gerontol A Biol Sci Med Sci* **76**, 1117-1123 (2021). https://doi.org/10.1093/gerona/glab016
27 Liu, Z. *et al.* Underlying features of epigenetic aging clocks in vivo and in vitro. *Aging Cell* **19**, e13229 (2020). https://doi.org/10.1111/acel.13229
28 Shen, X. *et al.* Nonlinear dynamics of multi-omics profiles during human aging. *Nat Aging* (2024). https://doi.org/10.1038/s43587-024-00692-2
29 Wu, X. *et al.* DNA methylation profile is a quantitative measure of biological aging in children. *Aging (Albany NY)* **11**, 10031-10051 (2019). https://doi.org/10.18632/aging.102399
30 Teschendorff, A. E. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. *Genome Med* **12**, 56 (2020). https://doi.org/10.1186/s13073-020-00752-3
31 Shireby, G. L. *et al.* Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. *Brain* **143**, 3763-3775 (2020). https://doi.org/10.1093/brain/awaa334
32 Lu, A. T. *et al.* DNA methylation-based estimator of telomere length. *Aging (Albany NY)* **11**, 5895-5923 (2019). https://doi.org/10.18632/aging.102173