Integrative approaches of DNA methylation patterns according to age, sex, and longitudinal changes

Jeong-An Gim (vitastar@korea.ac.kr)
Korea University Guro Hospital https://orcid.org/0000-0001-7292-2520

Research Article

Keywords: Aging, Biological age, DNA methylation, Sex differences, Longitudinal study

Posted Date: March 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1427963/v1

Abstract

Background: In humans, age-related DNA methylation has been studied in blood, tissues, buccal swabs, and fibroblasts, and changes in DNA methylation patterns according to age and sex have been detected. To date, approximately 137,000 samples have been analyzed from 14,000 studies and the infor-mation has been uploaded to the NCBI GEO database.

Methods: A correlation between age and methylation level and longitudinal changes in methylation levels was revealed in both sexes. Here, 20 public datasets derived from whole blood were analyzed using the Illumina BeadChip.

Results: Of the 20 datasets, nine were from a longitudinal study. All data had age and sex as common variables. Comprehensive details of age-, sex-, and longitudinal change-based DNA methylation levels in the whole blood sample were elucidated in this study. ELOVL2 and FHL2 showed the maximum correlation between age and DNA methylation. The methylation patterns of genes related to mental health differed according to age.

Conclusions: Based on 20 public DNA methylation datasets, methylation levels according to age and longitu-dinal changes by sex were identified and visualized using an integrated approach. The results highlight the molecular mechanisms underlying the association of sex and biological age with changes in DNA methylation, and the importance of optimal genomic information management.

Highlights

- Changes in methylation levels according to age and longitudinal changes by sex were revealed and visualized for 20 publicly available DNA methylation datasets.
- Methylation levels according to age and longitudinal changes by sex were identified and visualized using an integrated approach.
- Age correlations, sex differences, and longitudinally changed patterns are summarized as one summarized plot.

Background

Epigenetic modifications, such as DNA methylation, play important roles in development, aging, and disease [1–5]. DNA methylation levels showed age-related differences and has been used as a methylation score to predict chronological aging [3, 6–8]. Several diseases accelerate the change in DNA methylation levels with aging, and the disease-associated risk scores have been elucidated. Studies have been conducted on changes in DNA methylation according to neurological disorders, such as Alzheimer's disease (AD), posttraumatic stress disorder (PTSD), and suicidal ideation [9–13]. Smoking, lifestyle, and socioeconomic factors can be attributed to varying DNA methylation patterns [7, 14, 15]. Because of the large number of CpG sites (450,000) in one-sample analysis, it is relatively easy to discover biological features and explain complex traits. DNA methylation level is also related to X chromosome inactivation, and different patterns have been observed depending on sex [15–17]. Among many factors related to DNA methylation, age and sex are the most important. Therefore, studying patterns of DNA methylation with respect to these two factors is important.

Advanced technologies, such as the Illumina HumanMethylation BeadChip, which can evaluate DNA methylation levels, have been used in cohort studies of twins or individuals with chronic diseases [18–20]. In addition, a longitudinal study reported the change in the DNA methylation profile in one individual [21–25]. The Korean Genome and Epidemiology Study (KoGES) was conducted to analyze population health trends for personalized and preventive medicine. KoGES collected longitudinal follow-up clinical data using DNA methylation analysis [26]. Cohort-based longitudinal studies have sufficiently explained DNA methylation changes due to age and chronic diseases; however, integrated approaches and comprehensive analysis of these results is warranted.

DNA methylation analysis results derived from various cohorts have been deposited in public databases such as NCBI GEO, and information related to cohort characteristics, including age and sex, are publicly available. To date, approximately 137,000 samples from 14,000 studies have been analyzed and uploaded to the NCBI GEO database [27]. From NCBI GEO, 20 datasets derived from whole blood that were analyzed using the Illumina HumanMethylation BeadChip were selected for integrative analysis. To confirm the methylation status associated with aging and longitudinal changes according to sex, comprehensive landscapes showing the correlation between age and methylation level, differences between sexes, and longitudinal changes in methylation levels were visualized in this study.

Materials And Methods

Public datasets

The “GSExxxxx_series_matrix.txt.gz” file was downloaded from the “Series Matrix File(s)” section of NCBI GEO. In some cases, the downloaded file contained subject information and beta value; otherwise, a beta value was provided separately in the “Supplementary file” section. Sample information was retrieved from the “GSExxxxx_series_matrix.txt.gz” files, and “Sample_characteristics” strings were selected using the “filter” function of “dplyr” R package. The beta value and sample information files were opened using the “fread” function of “data.frame” R package. All beta values and sample information have been deposited as “dataframe” in R.

Batch effect correction

To correct the batch effects with respect to the time difference in longitudinal study, "ComBat" function of "sva" R package was utilized [28]. The two time-points, baseline (BL) and follow-up (FU) were the parameters used as the default of the “ComBat” function.

Statistical analysis

For information on the age of the subjects, the correlation coefficient and p-value with the subject's beta value were obtained using the “cor.test” function, which is the default function of R. Statistical significance of age-correlated CpG sites was visualized as Manhattan plots. The “manhattan” function of “qqman” R package was used, and input data were -log10 treated p-values of each CpG site. FCs between two sexes, and two time-points (BL and FU) with p-value were retrieved using the “t.test” R default function.

To confirm the longitudinal changes between the two time-points, the FU matrix was divided by BL. The overall change in pattern was retrieved by the inverse of the coefficient of variation (COV), which was obtained by dividing the standard deviation (SD) by the average of each CpG. The higher the methylation level of the CpG site between the two time-points and the greater the similarity in the pattern, the higher the **inverse of the COV** observed.

Visualization

In the volcano plot, the fold change (FC) between two sexes and time point is displayed on the x-axis, the -log10 treated p-value is displayed on the y-axis, and the points are the CpG sites analyzed in each study. Statistically significant genes that were hypermethylated in males and females are shown in blue on the upper right, and red on the upper left, respectively. Similarly, genes hypermethylated in FU and BL are indicated in blue on the upper right and red on the upper left, respectively. For Volcano plot, four R's default functions “with,” “plot,” “abline,” “subset,” and “points” were used.

Heatmaps were visualized using the “pheatmap” R package. Two sexes or time-points were provided as a column annotation bar. FCs, -log10 treated p-values, and chromosomal location (autosomes and sex chromosomes) of each CpG site are provided as row annotation bars. Selected CpG sites were labelled as gene symbols, and annotation numbers (cgxxxxxxxx) were labelled when the CpG sites were in the intergenic regions.

Results

Analysis platform

Twenty datasets derived from whole blood that were analyzed using the Illumina HumanMethylation BeadChip were included in this study (Table 1). From the 20 datasets, nine datasets were longitudinal studies, and 11 datasets were analyzed simultaneously. All datasets included age, sex, and other information, such as disease, treatment and clinical laboratory data. The average, SD, minimum, and maximum values of each study were recorded at the time of collection. CpG probe accession numbers (Illumina) starting with “cg” were retrieved from the beta-value matrix. Missing values and *duplicate* values of all samples were excluded to avoid errors in the statistical analysis.

Table 1
All datasets used in this study.

Study no.	Subject no. (Female no., %)	Age (Average ± SD) [Min, Max]	Healthy subjects (Female no., %)	Disease (Female no., %)	Sources CpG no.	Platform	References
Longitudinal study includes normal samples with continuous age							
BL KoGES	446 (220, 49.33%)*	52.24 ± 8.41 [40, 69]	All samples were considered as normal.		403,129	GPL13534	[26]
FU KoGES	50 (21, 42.00%)*	44.94 ± 4.81 [40, 63] (Base) 52.90 ± 4.83 [48, 71] (FU)	All samples were considered as normal.		431,651	GPL13534	[26]
GSE61151	92 (92, 100.00%)	53.20 ± 8.62 [35, 77] (Base) 59.20 ± 8.63 [41, 83] (FU)	All samples were considered as normal.		484,949	GPL13534	[33]
GSE74548	87 (47, 54.02%)	70.92 ± 2.96 [65, 75] (Base) 72.92 ± 2.96 [67, 77] (FU)	Placebo group: 43 subjects.	Supplementation group: folic acid and vitamin B12 for 44 subjects.	485,512	GPL13534	[32]
GSE130748	20 (11, 55.00%)*	75.47 ± 2.39 [71, 79] (Base) 80.00 ± 2.50 [76, 84] (FU)	All samples were considered as normal.		866,836	GPL21145	[51]
GSE140038	72 (72, 100.00%)*	56.68 ± 10.17 [36, 77] (Base) 57.07 ± 10.28 [36, 77] (FU)	NA	All subjects were breast cancer patients.	865,859	GPL23976	[29]
GSE142512	174 (79, 45.40%)	4.0626 ± 3.1680 [0.7255, 12.2272] (Unique) 2.3500 ± 2.7512 [0.5859, 15.0773] (Base) 8.7060 ± 4.4064 [1.1010, 22.7930] (FU)	Healthy controls: 199 subjects.	T1D: 196 subjects.	375,020 664,614	GPL13534 (n = 184), GPL23976 (n = 211)	[19, 35]
GSE143411	10 (1, 10.00%)**	58.30 ± 4.85 [49, 66] (Base) 63.50 ± 4.90 [54, 71] (FU)	NA	All subjects were CLL patients.	364,108	GPL13534	[30]

CB; cord blood. CD; Crohn's disease. CLL; chronic lymphocytic leukemia. FU; follow-up. GPL13534; Illumina HumanMethylation450 BeadChip. GPL21145; Illumina Infinium MethylationEPIC BeadChip. GPL23976; Illumina Infinium HumanMethylation850 BeadChip. KoGES: Korean genome and epidemiology study. NA; not applicable. PBB; Polybrominated biphenyl. PD; Parkinson's disease PD; Parkinson's disease. T1D; type 1 diabetes. *Not matched between BL and FU subjects. **Because the female subject was only one, correlation analysis was performed in the male subjects.

Study no.	Subject no. (Female no., %)	Age (Average ± SD) [Min, Max]	Healthy subjects (Female no., %)	Disease (Female no., %)	Sources CpG no.	Platform	References
GSE150643	120 (73, 60.83%)	11.376 ± 1.017 [9.188, 13.807] (Base) 13.330 ± 1.034 [11.150, 15.850] (FU)	All samples were considered as normal.		797,603	GPL21145	[52]
GSE161476	54 (54, 100.00%)	40.61 ± 13.30 [19, 69] (Base) 42.37 ± 13.27 [21, 70] (FU)	NA	Lupus patients, 54 female subjects of 229 samples.	582,738	GPL21145	[21]
Large cohort study includes normal samples with continuous age							
GSE30870	40 (0, 0%)	93.15 ± 4.31 [89, 103] (Only nonagenarians)	Male newborns 20 subjects.	Male nonagenarians 20 subjects.	485,577	GPL13534	[53, 54]
GSE40279	656 (338, 51.52%)	64.04 ± 14.74 [19, 101]	All samples were considered as normal.		473,034	GPL13534	[40]
GSE51388	60 (24, 40.00%)	34.52 ± 12.27 [23, 74]	All samples were considered as normal.		362,822	GPL13534	[55]
GSE55763	2,711 (871, 32.13%)	51.02 ± 10.09 [23.7, 75.0]	All samples were considered as normal.		431,906	GPL13534	[36, 37]
GSE69270	184 (111, 60.33%)	44.22 ± 3.25 [40, 49]	All samples were considered as normal.		408,148	GPL13534	[56, 57]
GSE72774	508 (227, 44.69%)	69.58 ± 11.22 [35.1, 91.9]	Healthy controls: 219 subjects.	PD: 289 subjects.	484,673	GPL13534	[58–60]
GSE72775	335 (138, 41.19%)	70.22 ± 10.30 [36.5, 90.5]	All samples were considered as normal.		484,915	GPL13534	[42, 61, 62]
GSE87571	729 (388, 32.13%)	47.40 ± 20.94 [14, 94]	All samples were considered as normal.		450,282	GPL13534	[34]
GSE111629	572 (249, 43.53%)	69.05 ± 11.50 [35, 92]	Healthy controls: 237 subjects.	PD: 335 subjects.	484,643	GPL13534	[44, 59, 60]
GSE112611	402 (170, 32.13%)	13.45 ± 3.26 [4.50, 20.78]	Healthy controls: 74 subjects.	CD: 328 subjects.	504,790	GPL21145	[63]

CB; cord blood. CD; Crohn's disease. CLL; chronic lymphocytic leukemia. FU; follow-up. GPL13534; Illumina HumanMethylation450 BeadChip. GPL21145; Illumina Infinium MethylationEPIC BeadChip. GPL23976; Illumina Infinium HumanMethylation850 BeadChip. KoGES: Korean genome and epidemiology study. NA; not applicable. PBB; Polybrominated biphenyl. PD; Parkinson's disease PD; Parkinson's disease. T1D; type 1 diabetes. *Not matched between BL and FU subjects. **Because the female subject was only one, correlation analysis was performed in the male subjects.

Study no.	Subject no. (Female no., %)	Age (Average ± SD) [Min, Max]	Healthy subjects (Female no., %)	Disease (Female no., %)	Sources CpG no.	Platform	References
GSE116339	679 (399, 58.76%)	53.92 ± 12.92 [23.00, 88.46]	Healthy controls were considered as PBB-153 exposure < 1: 520 subjects	PBB exposure subjects were considered as PBB-153 exposure > 1: 159 subjects.	763,746	GPL21145	[31]

CB; cord blood. CD; Crohn's disease. CLL; chronic lymphocytic leukemia. FU; follow-up. GPL13534; Illumina HumanMethylation450 BeadChip. GPL21145; Illumina Infinium MethylationEPIC BeadChip. GPL23976; Illumina Infinium HumanMethylation850 BeadChip. KoGES: Korean genome and epidemiology study. NA; not applicable. PBB; Polybrominated biphenyl. PD; Parkinson's disease PD; Parkinson's disease. T1D; type 1 diabetes. *Not matched between BL and FU subjects. **Because the female subject was only one, correlation analysis was performed in the male subjects.

As KoGES is a community-based cohort and does not include patients of acute diseases, all subjects were considered healthy [26]. In three studies, all samples were from breast cancer, chronic lymphocytic leukemia (CLL), and lupus. GSE140038 included 72 female subjects with breast cancer [29], and nine male CLL patients were enrolled in GSE143411 [30]. GSE161476 included 54 female subjects with lupus [21]. In GSE140038, it was unclear which subject was connected between the two time-points. Therefore, serial changes could not be confirmed, and only a *t*-test could be performed between two time-points. Healthy subjects and patients were separated and analyzed separately. In the GSE116339 dataset, subjects with a total PBB exposure of 1 were considered healthy [31]. The GSE74548 study was divided into a folic acid and vitamin B12 treatment group and a non-treated group; the treatment group was excluded [32].

In the GSE61151 dataset, four samples (two subjects) were excluded out of the 188 female samples [33], *because the samples had unclear subject information*. The GSE87571 dataset included 732 samples, and samples with no information (n = 3) were excluded from the analysis. [34] In the case of GSE142512 analysis, three or more time-points were used [19, 35], and the first and last results were used (time-points ≥ 3) in our analysis. In most FU studies, subjects enrolled at baseline (BL) were included in the final *analysis*. Unmatched BL and FU subjects were excluded from the longitudinal analysis. In the case of KoGES, a FU study was conducted on 50 of the 446 BL samples [26]. Therefore, a separate row is included in Table 1, and 50 BL subjects are indicated together with the FU studies.

Figure 1 displays the average age of the subjects in each cohort in proportion to the x-axis for the 20 datasets. The interval between the BL and FU studies was reflected in the longitudinal study. Cohorts that comprised only patients are shown on a green background, and cohorts containing both the normal and disease groups are shown on a yellow background. The rest are displayed on a white background. A comparative analysis was performed on samples that exactly matched the subjects between the BL and FU studies.

Correlation analysis of DNA methylation and age at a given point in time

For each of the 20 cohorts, the correlation between age and total CpG sites was analyzed 51 times. The correlation coefficient and p-value were obtained using R's default function "cor.test." A total of 2,343,070 CpG sites satisfying p-value < 0.01 were obtained from each of the 51 analyses. These sites were divided into 1,191,273 male- and 1,151,797 female-specific CpG sites and provided as two Manhattan plots. CpG sites, which showed a statistically significant correlation with age, were observed in most of the genomic regions (Fig. 2). Next, the genomic region of the CpG site that had a statistically significant correlation with age was determined. In males, two or more CpG sites beyond the red horizontal line with p-value

$< 10^{-200}$ were observed on chromosomes 2 and 6. In females, and patterns of significant correlation with age were observed on chromosomes 2 and 6, although a relatively lower statistical significance was observed.

Manhattan plots for each dataset were obtained from 51 analyses (Figure S1). Two guide horizontal lines are indicated in blue and red at 50% and 80% of the maximum value of log10 (p-value), respectively. The top 100 and bottom 100 CpG sites were selected by sorting them in the order of decreasing correlation coefficients. These data were merged according to cohort information, sex, and negative or positive correlation, and finally, 10,200 CpG sites were identified (Table S1). From these 10,200 CpG sites, 6,868 CpG sites were unique, whereas 3,332 CpG sites were common in two or more cohorts. The CpG sites that showed a statistically significant correlation with age in males and females were the *FHL2* gene on chromosome 2 and the *ELOVL2* gene on chromosome 6. A higher correlation of the two genes was found in GSE55763 that had the largest number of samples; GSE87571 had the next largest number. *ELOVL2* and *FHL2* were commonly observed in 11 male and 15 female cohorts. A probe on the Y chromosome in the female datasets was considered sample contamination or false positive.

Differently methylated patterns between two sexes at a given point in time

A total of 21 *t*-tests were performed on 15 cohorts to identify significant genes between the two sexes, and FC and p-values were obtained. Volcano plots are presented based on these analyses, and genes significantly different between the sexes are displayed in different colors (Figure S2). As the number of samples increased, the p-value tended to decrease, and in large cohorts, such as GSE55763 [36, 37], as evident from a point located at the end of the y-axis with a low p-value.

A heatmap was generated to visualize the beta value for each subject by selecting significantly different methylated CpG sites according to sex (Figure S3). The two sexes are presented as column annotation bars indicated on the x-axis, and they were well distinguished by unsupervised k-means clustering. In the y-axis direction, FC and -log10 treated p-values for each CpG site are presented as row annotation bars. If the gene symbol was annotated at the CpG site, it was presented on the label; otherwise, the CpG probe annotation number was presented. The thresholds for selecting different methylated CpG sites are listed in Table 2, and 1,070 CpG sites are listed in Table S2.

Table 2
Thresholds for selecting different methylated CpG sites between two genders.

Study no.	Time point	CpG sites no.	FC <	FC >	-log10(PV)
GSE40279	One time	44	-10	10	100
GSE51388	One time	58	-0.2	0.2	5
GSE55763	One time	56	-0.1	0.1	200
GSE69270	One time	47	-0.55	0.4	150
GSE72774	One time	56	-0.55	0.4	160
GSE72775	One time	54	-0.6	0.5	200
GSE74548	BL	57	-0.45	0.35	25
GSE74548	FU	57	-0.45	0.35	29
GSE87571	One time	50	-0.6	0.4	300
GSE111629	One time	45	-0.45	0.35	200
GSE112611	One time	59	-0.5	0.5	35
GSE116339	One time	46	-0.55	0.4	300
GSE130748	BL	46	-0.45	0.35	18
GSE130748	FU	35	-0.45	0.35	15
GSE142512	One time	60	-5	4	10
GSE142512	BL	44	-5	4	33
GSE142512	FU	60	-5	4	37
GSE150643	BL	53	-0.1	0.1	10
GSE150643	FU	54	-0.1	0.1	10
KoGES	BL	53	-0.55	0.35	280
KoGES	FU	36	-0.6	0.5	20

BL; baseline. FU; follow-up. FC; fold change. PV; *p*-value.

Ratios between two time-points of the longitudinal study

To quantify the difference between BL and FU in the longitudinal study, the FU value was divided by BL for the CpG sites. To evaluate the degree and consistency of the pattern of change between the two groups, the SD and average FU/BL were calculated for each CpG site. The COV, calculated as SD/average, was then obtained, and the reciprocal of the COV was calculated. The reciprocal of COV is expressed as a Manhattan plot. The reciprocal of the COV was higher, with a lower SD and a higher average (high amount of change) at each CpG site.

A total of 7,284,406 CpG sites were retrieved, and Manhattan plots for the COV reciprocal for all 14 pairs are presented for each dataset. Selected 27,839 CpG sites with |COV reciprocal| > 0.6 are listed in Table S3. Then, COV reciprocals are provided as two Manhattan plots according to sex (Figure S4). A total of 6,172,646 and 1,111,760 COV reciprocals were used for the men and women, respectively. The larger the positive value on the y-axis, the more hypermethylated the CpG

sites in FU compared to BL. Alternatively, a negative value indicated a hypomethylated CpG site in the second analysis (Fig. 3).

Differently methylated patterns between two time-points in longitudinal study

Fourteen *t*-tests were performed on the nine cohorts to identify the significant genes between the two time-points, and FC and p-values were obtained. A volcano plot was generated based on these results, and significantly different genes are displayed in different colors (Figure S5).

By selecting significantly different methylated CpG sites according to the two time-points, a heat map was generated to visualize the beta value for each subject (Figure S6). The two time-points are presented as column annotation bars indicated on the x-axis, and they were distinguished by unsupervised k-means clustering. In the y-axis, FC and -log10 treated p-values for each CpG site are presented as row annotation bars. If the gene symbol was annotated at the CpG site, it was presented on the label; otherwise, the CpG probe annotation number starting with “cg” was presented. The thresholds for selecting different methylated CpG sites are listed in Table 3, and 784 CpG sites are listed in Table S4.

Table 3
Threshold for selecting different methylated CpG sites between two time points.

Study no.	Gender	CpG sites no.	FC <	FC >	PV
GSE74548	F	61	-0.04	0.04	2
GSE74548	M	53	-0.07	0.07	2
GSE130748	F	53	-0.095	0.095	2
GSE130748	M	50	-0.12	0.12	2
GSE140038	F	44	-0.055	0.055	5
GSE142512	F	49	-1.1	1.1	4
GSE142512	M	62	-1.2	1.2	5
GSE143411	M	44	-0.2	0.2	1
GSE150643	F	55	-0.037	0.037	3
GSE150643	M	55	-0.033	0.033	2
GSE161476	F	66	-0.35	0.35	2.5
KoGES FU	F	71	-0.075	0.075	8
KoGES FU	M	73	-0.06	0.06	6
BL; baseline. FU; follow-up. FC; fold change. PV; *p*-value.					

Comprehensive visualization

All datasets were merged according to genomic location, correlation coefficients, FCs, and p-values. Enrichment terms were then extracted using correlation coefficients and FCs to select functions related to age and sex differences. KEGG terms, upset plots, and network analyses were performed.

Two KEGG enrichment analyses were performed and enrichment terms were visualized using KEGG analysis (Figure S7) [38]. Correlation coefficients were retrieved from 20 integrated cohorts, and the top 20 KEGG terms were visualized as dot plots. The reciprocals of the COVs were used as input data for KEGG enrichment analysis, and the top 20 KEGG terms were

listed. Four of the top five terms were common: “human papillomavirus infection,” “MAPK signaling pathway,” “calcium signaling pathway,” and “Rap1 signaling pathway.”

Upset plots and networks were visualized based on the age-correlated CpG sites (Fig. 4) and the sites correlated with other three conditions (Figure S8). Among the age-correlated genes, differentially methylated CpG sites were enriched in neurologic terms. Then, differently methylated regions of the two sexes were used to input features of the “pathfindR” package [39]. Ten terms with hypomethylated genes in female samples were revealed in an upset plot (Figure S8a), and five nodes were revealed in the network analysis (Figure S8b). The reciprocals of the COVs were used as input features, and the upset plot and network analysis revealed ten terms and two nodes (Figure S8c and S8d). Ten terms and one node (poor school performance) were detected and the longitudinal differences in differently methylated regions are shown in Figures S8e and S8f.

Four analyses were integrated into two peaks and two heatmaps in one circ*o*s plot (Fig. 5). Age correlations, sex differences, and longitudinally changed patterns are summarized as four tracks. In the outer peak of the Circos plot, positive and negative correlations between age and beta value of CpG sites are indicated as outer red and inner blue peaks, respectively. In the inner peak of the Circos plot, the reciprocals of the COVs were visualized as peaks. The yellow and green colors represent hypermethylation and hypomethylation in FU, respectively, when compared to BL.

The FC between the two groups is presented as the second and fourth heatmaps from the outside. In most genomic regions, hypermethylated CpG sites were observed in males, and a hypermethylated pattern on the sex chromosome was observed in females. When comparing the two time-points, hypermethylation and hypomethylation in FU compared to BL are displayed in pink and dark green, respectively. Overall, hypomethylation was observed in FU, and 14 regions showed hypermethylation, including chromosomes 1, 2, 4, and 10.

Discussion

Comprehensive analysis of DNA methylation patterns by integrating public datasets

This study aimed to provide a comprehensive landscape of the correlation between age and DNA methylation level, differences between the sexes, and longitudinal changes in DNA methylation levels. To the best of my knowledge, this is the first study to reveal and visualize patterns of DNA methylation changes based on age, sex, and time. The human aging rate has been explained by DNA methylation and established in terms of the aging clock [40] or epigenetic drift [5]. According to the aging clock model, the epigenetic clock accelerates under the influence of PTSD, menopause, and Down syndrome [13, 41, 42]. This study further elaborates on the aging clock model and its outcomes can be utilized for disease prediction and health management according to the age and sex of the subject.

DNA methylation patterns have been studied according to sex and age, and global hypomethylation has been detected in aged samples. In female samples, global hypomethylation patterns were detected except for the sex chromosome. When comparing the genomic imprinting between the sexes, the sex chromosomes have a different pattern from the autosome [16, 17]. This study also detected global hypomethylation in females, and the methylation levels decreased over time (Fig. 5). The approaches used in this study can help to reveal and visualize group-specific methylation patterns in future studies.

DNA methylation analysis was included in the FU data in the cohort study, which will remain publicly available. DNA methylation patterns have been observed in FU studies for various diseases [21, 43, 44], and a methodology has been presented. In this study, a comparative analysis using the *t*-test, is presented. Here, the reciprocal of COV was used as the degree of change to confirm the longitudinal change. This method can be used to track time-dependent changes in DNA methylation patterns.

DNA methylation patterns of normal controls

As the cost of genomic analysis per sample remains high, public datasets of normal individuals or patients with similar diseases are valuable resource. DNA methylation patterns based on different groups were confirmed in this study. Due to limited resources, it was difficult to collect normal samples. Thus, here, normal samples from large cohorts were merged and processed together (Fig. 1 and Table 1). A large cohort study for normal subjects will continue [23, 26, 45], and a more elaborate batch effect correction tool will be presented [28, 46]. The methylation patterns of the normal controls were compared with those of the patient samples, and patient-specific DNA methylation patterns were elucidated.

DNA methylation is the reversible addition of a methyl group to a nucleotide, which can be altered by environmental and lifestyle factors [14, 15, 43]. DNA methylation is a dynamic process that also occurs in a chronological manner and is influenced by multiple factors. For normal subjects included in the cohort, lifestyle factors including diet and exercise could not be completely controlled. Similarly, among normal subjects, those who may have had an undiagnosed disease, or a transitional stage of a chronic disease, may have been included. Therefore, in this study, an effort was made to observe the overall DNA methylation pattern by integrating several healthy subjects. DNA methylation pattern-based selection criteria for normal/healthy individuals should be developed for future studies.

Limitations of this study

This study has several limitations. First, the two time-points were not constant. The KoGES study was conducted approximately eight years apart. Usually, analyses are performed at approximately 5-year interval, whereas here, a 2-year interval was used. As the number of FU datasets increase, a cohort with a longer FU period will be created. The methods used in this study can be used to analyze samples obtained during long FU periods. Second, the number of subjects included in the cohort differed, and only age and sex were used as variables. Moreover, classification criteria subdivided according to race or lifestyle were not used. A cohort design is necessary to subdivide the subject enrollment criteria, such as lifestyle and race. Third, no gene set enrichment analysis (GSEA) was performed. GSEA is not only applied to NGS or microarray-based gene expression data [47], but also to DNA methylation analysis [48]. In this study, GSEA was excluded because it was difficult to combine three or more study groups analyzed on different platforms and at different times. In the future, integrated GSEA can be performed using a more sophisticated algorithm for data integration.

Further aspects: genomic information management

As the analyzed libraries have become more sophisticated, omics data processing and visualization are feasible in R or Python. The computer processing speed and storage space have increased. In this scenario, integrated insights are applicable to the processing of omics data. Thus, it is possible to understand the mechanisms of aging according to sex and to improve health in old age by suggesting preventive or therapeutic strategies. Identifying and maintaining the optimal DNA methylation state for healthy aging would enable proper genome information management [43, 49].

Illumina chips for DNA methylation analysis have been developed from 27k to 450k and 850k [50]. Currently, computer performance and R-or Python-based library analysis is sufficient. However, appropriate methods would be required to quickly process the beta values for several CpG sites. The method presented in this study can be used to identify factors associated with a healthy state. Future studies on DNA methylation changes should be performed for various conditions for a higher number of subjects to generate more comprehensive datasets.

Conclusions

In this study, changes in methylation levels according to age and longitudinal changes by sex were revealed and visualized for 20 publicly available DNA methylation datasets. As an integrated approach, the analysis methods and visualization strategies will form the basis of methylation-based data analysis in the future.

Authors’ contributions

The author defined the research hypotheses and study design. We performed a database search and selected studies. The author performed the data analysis and biological interpretation of the results. The authors have written the final manuscript.

Declarations

Authors’ contributions

The author defined the research hypotheses and study design. We performed a database search and selected studies. The author performed the data analysis and biological interpretation of the results. The authors have written the final manuscript.

Funding

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C0012); the National Research Foundation (NRF) funded by the Ministry of Education (grant number: NRF-2020R1I1A1A01052701).

Competing interests

The author declares that I have no competing interests.

References

1. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST. Age-associated DNA methylation in pediatric populations. Genome Res. 2012;22(4):623–32.
2. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol. 2013;20(3):274–81.
3. Kołodziej-Wojnar P, Borkowska J, Wicik Z, Domaszewska-Szostek A, Połosak J, Cąkała-Jakimowicz M, Bujanowska O, Puzianowska-Kuznicka M. Alterations in the Genomic Distribution of 5hmC in In Vivo Aged Human Skin Fibroblasts. Int J Mol Sci. 2021;22(1):78.
4. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD. **Global epigenomic reconfiguration during mammalian brain development**. Science 2013, 341(6146).
5. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013;22(R1):R7–15.
6. Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. Epigenetics. 2020;15(1–2):1–11.
7. Rauschert S, Melton PE, Heiskala A, Karhunen V, Burdge G, Craig JM, Godfrey KM, Lillycrop K, Mori TA, Beilin LJ. Machine learning-based DNA methylation score for fetal exposure to maternal smoking: development and validation in samples collected from adolescents and adults. Environ Health Perspect. 2020;128(9):097003.
8. Zhu T, Gao Y, Wang J, Li X, Shang S, Wang Y, Guo S, Zhou H, Liu H, Sun D. CancerClock: A DNA methylation age predictor to identify and characterize aging clock in pan-cancer. Front Bioeng Biotechnol. 2019;7:388.
9. Bhak Y, Jeong H-o, Cho YS, Jeon S, Cho J, Gim J-A, Jeon Y, Blazyte A, Park SG, Kim H-M. Depression and suicide risk prediction models using blood-derived multi-omics data. Translational psychiatry. 2019;9(1):1–8.
10. Jeremian R, Chen Ya, De Luca V, Vincent JB, Kennedy JL, Zai CC, Strauss J. Investigation of correlations between DNA methylation, suicidal behavior and aging. Bipolar Disord. 2017;19(1):32–40.

11. McCartney DL, Stevenson AJ, Walker RM, Gibson J, Morris SW, Campbell A, Murray AD, Whalley HC, Porteous DJ, McIntosh AM. Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. Alzheimer's & Dementia: Diagnosis Assessment & Disease Monitoring. 2018;10:429–37.
12. Mehta D, Bruenig D, Lawford B, Harvey W, Carrillo-Roa T, Morris CP, Jovanovic T, Young RM, Binder EB, Voisey J. Accelerated DNA methylation aging and increased resilience in veterans: the biological cost for soldiering on. Neurobiol stress. 2018;8:112–9.
13. Wolf EJ, Logue MW, Hayes JP, Sadeh N, Schichman SA, Stone A, Salat DH, Milberg W, McGlinchey R, Miller MW. Accelerated DNA methylation age: associations with PTSD and neural integrity. Psychoneuroendocrinology. 2016;63:155–62.
14. Zhao W, Ammous F, Ratliff S, Liu J, Yu M, Mosley TH, Kardia SL, Smith JA. Education and lifestyle factors are associated with DNA methylation clocks in older African Americans. Int J Environ Res Public Health. 2019;16(17):3141.
15. Hughes A, Smart M, Gorrie-Stone T, Hannon E, Mill J, Bao Y, Burrage J, Schalkwyk L, Kumari M. Socioeconomic position and DNA methylation age acceleration across the life course. Am J Epidemiol. 2018;187(11):2346–54.
16. El-Maarri O, Becker T, Junen J, Manzoor SS, Diaz-Lacava A, Schwaab R, Wienker T, Oldenburg J. Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. Hum Genet. 2007;122(5):505–14.
17. Zhang FF, Cardarelli R, Carroll J, Fulda KG, Kaur M, Gonzalez K, Vishwanatha JK, Santella RM, Morabia A. Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. Epigenetics. 2011;6(5):623–9.
18. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, Sommer IE, Kahn RS, Ophoff RA. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PLoS ONE. 2009;4(8):e6767.
19. Johnson RK, Vanderlinden LA, Dong F, Carry PM, Seifert J, Waugh K, Shorrosh H, Fingerlin T, Frohnert BI, Yang IV. Longitudinal DNA methylation differences precede type 1 diabetes. Sci Rep. 2020;10(1):1–13.
20. Svane AM, Soerensen M, Lund J, Tan Q, Jylhävä J, Wang Y, Pedersen NL, Hägg S, Debrabant B, Deary IJ. DNA methylation and all-cause mortality in middle-aged and elderly Danish twins. Genes. 2018;9(2):78.
21. Coit P, Ortiz-Fernandez L, Lewis EE, McCune WJ, Maksimowicz-McKinnon K, Sawalha AH. **A longitudinal and transancestral analysis of DNA methylation patterns and disease activity in lupus patients**. JCI insight 2020, 5(22).
22. Grant CD, Jafari N, Hou L, Li Y, Stewart JD, Zhang G, Lamichhane A, Manson JE, Baccarelli AA, Whitsel EA. A longitudinal study of DNA methylation as a potential mediator of age-related diabetes risk. Geroscience. 2017;39(5–6):475–89.
23. Tharakan R, Ubaida-Mohien C, Moore AZ, Hernandez D, Tanaka T, Ferrucci L. Blood DNA methylation and aging: A cross-sectional analysis and longitudinal validation in the InCHIANTI study. The Journals of Gerontology: Series A. 2020;75(11):2051–5.
24. Wang D, Liu X, Zhou Y, Xie H, Hong X, Tsai H-J, Wang G, Liu R, Wang X. Individual variation and longitudinal pattern of genome-wide DNA methylation from birth to the first two years of life. Epigenetics. 2012;7(6):594–605.
25. Wang Y, Pedersen NL, Hägg S. Implementing a method for studying longitudinal DNA methylation variability in association with age. Epigenetics. 2018;13(8):866–74.
26. Kim Y, Han B-G, Group K. Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. Int J Epidemiol. 2017;46(2):e20–0.
27. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41(D1):D991–5.
28. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3.

29. Sehl ME, Carroll JE, Horvath S, Bower JE. The acute effects of adjuvant radiation and chemotherapy on peripheral blood epigenetic age in early stage breast cancer patients. NPJ breast cancer. 2020;6(1):1–5.
30. Zapatka M, Tausch E, Öztürk S, Yosifov DY, Seiffert M, Zenz T, Schneider C, Blöhdorn J, Döhner H, Mertens D: **Clonal evolution in chronic lymphocytic leukemia is scant in relapsed but accelerated in refractory cases after chemo (immune) therapy**. *Haematologica* 2020.
31. Curtis SW, Cobb DO, Kilaru V, Terrell ML, Kennedy EM, Marder ME, Barr DB, Marsit CJ, Marcus M, Conneely KN. Exposure to polybrominated biphenyl (PBB) associates with genome-wide DNA methylation differences in peripheral blood. Epigenetics. 2019;14(1):52–66.
32. Kok DE, Dhonukshe-Rutten RA, Lute C, Heil SG, Uitterlinden AG, van der Velde N, van Meurs JB, van Schoor NM, Hooiveld GJ, de Groot LC. The effects of long-term daily folic acid and vitamin B 12 supplementation on genome-wide DNA methylation in elderly subjects. Clin epigenetics. 2015;7(1):1–14.
33. Flanagan JM, Brook MN, Orr N, Tomczyk K, Coulson P, Fletcher O, Jones ME, Schoemaker MJ, Ashworth A, Swerdlow A. Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. Cancer Epidemiol Prev Biomarkers. 2015;24(1):221–9.
34. Johansson Å, Enroth S, Gyllensten U. Continuous aging of the human DNA methylome throughout the human lifespan. PLoS ONE. 2013;8(6):e67378.
35. Vanderlinden LA, Johnson RK, Carry PM, Dong F, DeMeo DL, Yang IV, Norris JM, Kechris K. An effective processing pipeline for harmonizing DNA methylation data from Illumina's 450K and EPIC platforms for epidemiological studies. BMC Res Notes. 2021;14(1):1–7.
36. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan S-T, Afzal U, Scott J, Jarvelin M-R, Elliott P. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol. 2015;16(1):1–12.
37. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai P-C, Ried JS, Zhang W, Yang Y. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541(7635):81–6.
38. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49(D1):D545–51.
39. Ulgen E, Ozisik O, Sezerman OU. pathfindR: An R package for comprehensive identification of enriched pathways in omics data through active subnetworks. Front Genet. 2019;10:858.
40. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan J-B, Gao Y. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67.
41. Horvath S, Garagnani P, Bacalini MG, Pirazzini C, Salvioli S, Gentilini D, Di Blasio AM, Giuliani C, Tung S, Vinters HV. Accelerated epigenetic aging in Down syndrome. Aging Cell. 2015;14(3):491–5.
42. Levine ME, Lu AT, Chen BH, Hernandez DG, Singleton AB, Ferrucci L, Bandinelli S, Salfati E, Manson JE, Quach A: **Menopause accelerates biological aging**. *Proceedings of the National Academy of Sciences* 2016, **113**(33):9327–9332.
43. Chen R, Xia L, Tu K, Duan M, Kukurba K, Li-Pook-Than J, Xie D, Snyder M. Longitudinal personal DNA methylome dynamics in a human with a chronic condition. Nat Med. 2018;24(12):1930–9.
44. Chuang Y-H, Lu AT, Paul KC, Folle AD, Bronstein JM, Bordelon Y, Horvath S, Ritz B. Longitudinal epigenome-wide methylation study of cognitive decline and motor progression in Parkinson's disease. J Parkinson's disease. 2019;9(2):389–400.
45. Hannon E, Knox O, Sugden K, Burrage J, Wong CC, Belsky DW, Corcoran DL, Arseneault L, Moffitt TE, Caspi A. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. PLoS Genet. 2018;14(8):e1007544.
46. Price EM, Robinson WP. Adjusting for batch effects in DNA methylation microarray data, a lesson learned. Front Genet. 2018;9:83.

47. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**(43):15545–15550.
48. Maksimovic J, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. Genome Biol. 2021;22(1):1–26.
49. Kim S, Wyckoff J, Morris A-T, Succop A, Avery A, Duncan GE, Jazwinski SM. DNA methylation associated with healthy aging of elderly twins. Geroscience. 2018;40(5):469–84.
50. Dhingra R, Kwee LC, Diaz-Sanchez D, Devlin RB, Cascio W, Hauser ER, Gregory S, Shah S, Kraus WE, Olden K. Evaluating DNA methylation age on the illumina MethylationEPIC bead chip. PLoS ONE. 2019;14(4):e0207834.
51. Bartlett AH, Liang JW, Sandoval-Sierra JV, Fowke JH, Simonsick EM, Johnson KC, Mozhui K. Longitudinal study of leukocyte DNA methylation and biomarkers for cancer risk in older adults. Biomark Res. 2019;7(1):1–13.
52. Moore SR, Humphreys KL, Colich NL, Davis EG, Lin DTS, MacIsaac JL, Kobor MS, Gotlib IH. Distinctions between sex and time in patterns of DNA methylation across puberty. BMC Genomics. 2020;21:1–16.
53. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ: **Distinct DNA methylomes of newborns and centenarians**. *Proceedings of the National Academy of Sciences* 2012, **109**(26):10522–10527.
54. Simo-Riudalbas L, Diaz-Lagares A, Gatto S, Gagliardi M, Crujeiras AB, Matarazzo M, Esteller M, Sandoval J. Genome-wide DNA methylation analysis identifies novel hypomethylated non-pericentromeric genes with potential clinical implications in ICF syndrome. PLoS ONE. 2015;10(7):e0132517.
55. Zhang N, Zhao S, Zhang S-H, Chen J, Lu D, Shen M, Li C. Intra-monozygotic twin pair discordance and longitudinal variation of whole-genome scale DNA methylation in adults. PLoS ONE. 2015;10(8):e0135022.
56. Kananen L, Marttila S, Nevalainen T, Jylhävä J, Mononen N, Kähönen M, Raitakari OT, Lehtimäki T, Hurme M. Aging-associated DNA methylation changes in middle-aged individuals: the Young Finns study. BMC Genomics. 2016;17(1):1–12.
57. Mishra PP, Hänninen I, Raitoharju E, Marttila S, Mishra BH, Mononen N, Kähönen M, Hurme M, Raitakari O, Törönen P. Epigenome-450K-wide methylation signatures of active cigarette smoking: The Young Finns Study. Biosci Rep. 2020;40(7):BSR20200596.
58. Paul KC, Binder AM, Horvath S, Kusters C, Yan Q, Rosario ID, Yu Y, Bronstein J, Ritz B. Accelerated hematopoietic mitotic aging measured by DNA methylation, blood cell lineage, and Parkinson's disease. BMC Genomics. 2021;22(1):1–10.
59. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. Aging. 2015;7(12):1130.
60. Chuang Y-H, Paul KC, Bronstein JM, Bordelon Y, Horvath S, Ritz B. Parkinson's disease is associated with DNA methylation levels in human blood and saliva. Genome Med. 2017;9(1):1–12.
61. Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, Ritz BR, Chen B, Lu AT, Rickabaugh TM. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016;17(1):1–23.
62. Chuang Y-H, Quach A, Absher D, Assimes T, Horvath S, Ritz B. Coffee consumption is associated with DNA methylation levels of human blood. Eur J Hum Genet. 2017;25(5):608–16.
63. Somineni HK, Venkateswaran S, Kilaru V, Marigorta UM, Mo A, Okou DT, Kellermayer R, Mondal K, Cobb D, Walters TD. Blood-derived DNA methylation signatures of Crohn's disease and severity of intestinal inflammation. Gastroenterology. 2019;156(8):2254–65. e2253.

Supplementary Tables

Supplementary Tables S1-S4 are not available with this version

Figures

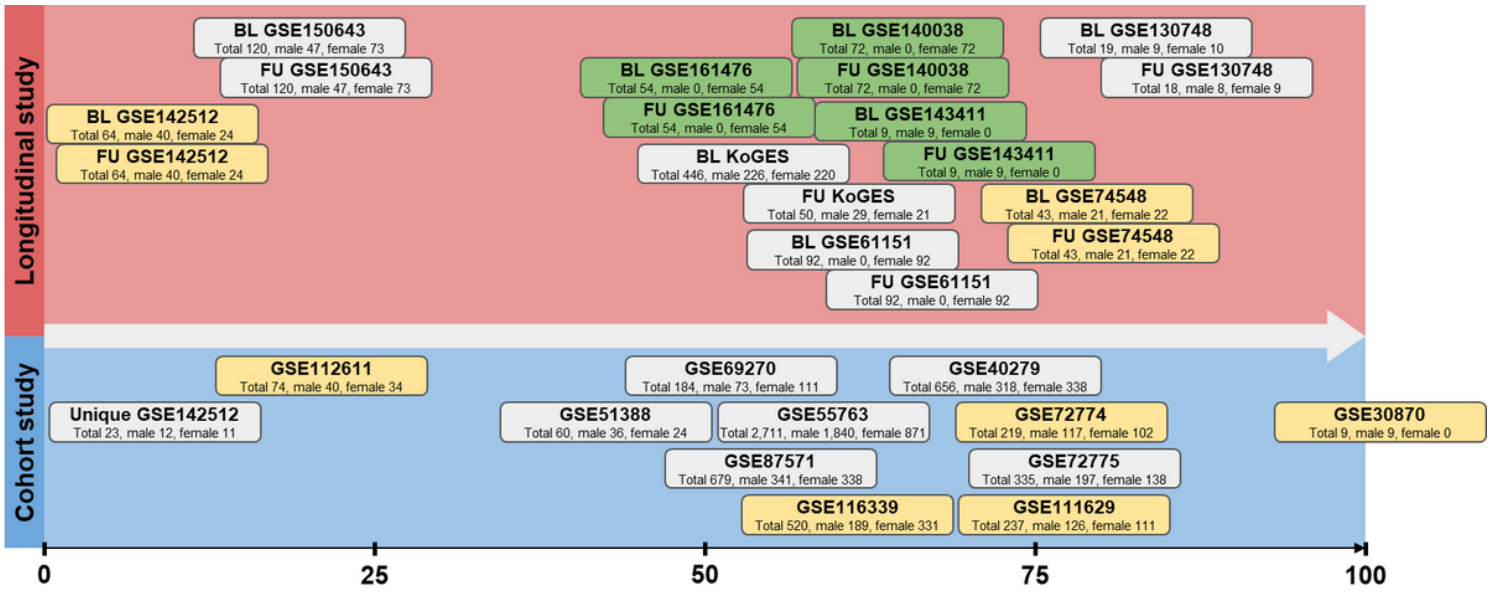

Figure 1

Twenty public datasets of DNA methylation analysis were used in this study. Left side of rectangle represents the average age at the position on the X-axis. Nine longitudinal studies are shown at the top, and 11 large cohort studies are shown at the bottom. White background indicates normal cohort, and yellow background indicates cohort containing disease samples. Green background indicates that cohort comprised patients. Total, male and female subjects included in the dataset are presented.

Figure 2

Manhattan plots of statistical significance for correlation between age and DNA methylation level. Correlation was analyzed for male (top) and female (bottom) samples, and -log10 was treated to p-value. Two guide horizonal lines (p-values of 10-50 and 10-200) are shown as blue and red lines, respectively.

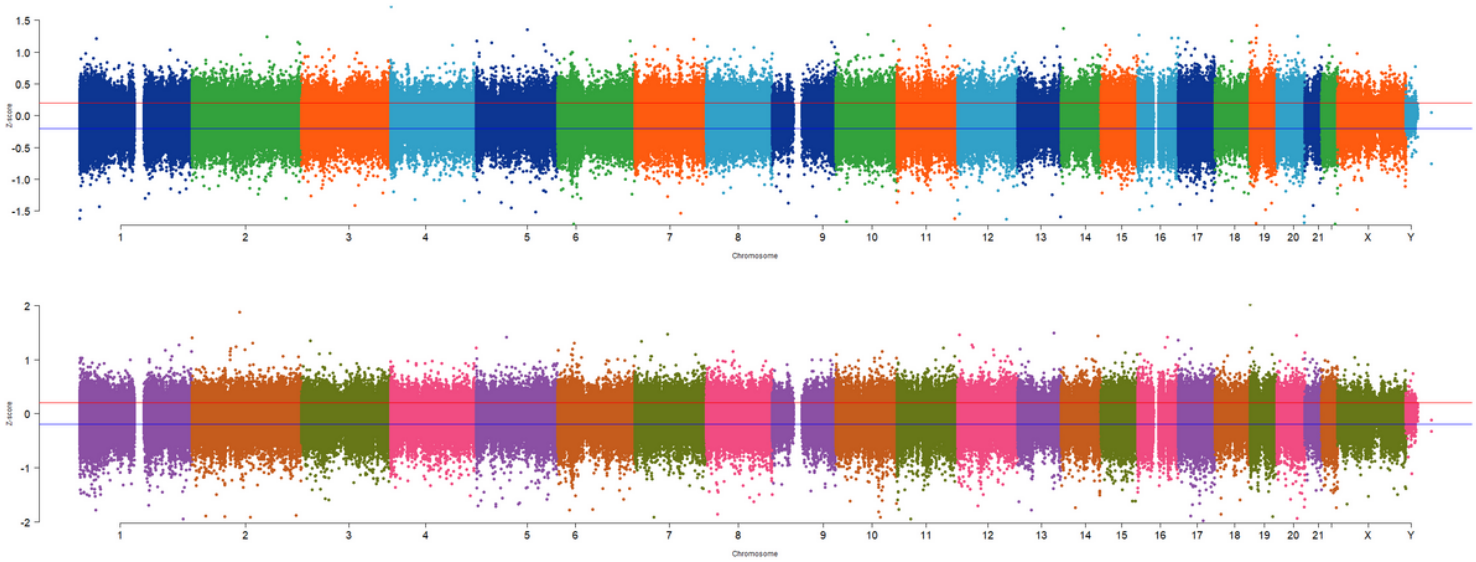

Figure 3

Manhattan plots of coefficient of variation (COV) reciprocals for each CpG sites from total pairs of baseline (BL) and follow-up (FU) datasets. In FU matrix, beta value for each CpG site was divided by that of BL matrix. COV (standard deviation (SD) / average) for each CpG site was calculated, and reciprocal of COV was displayed on y-axis of the Manhattan plot. Calculations were performed on the matrices corresponding to BL and FU in the two sexes.

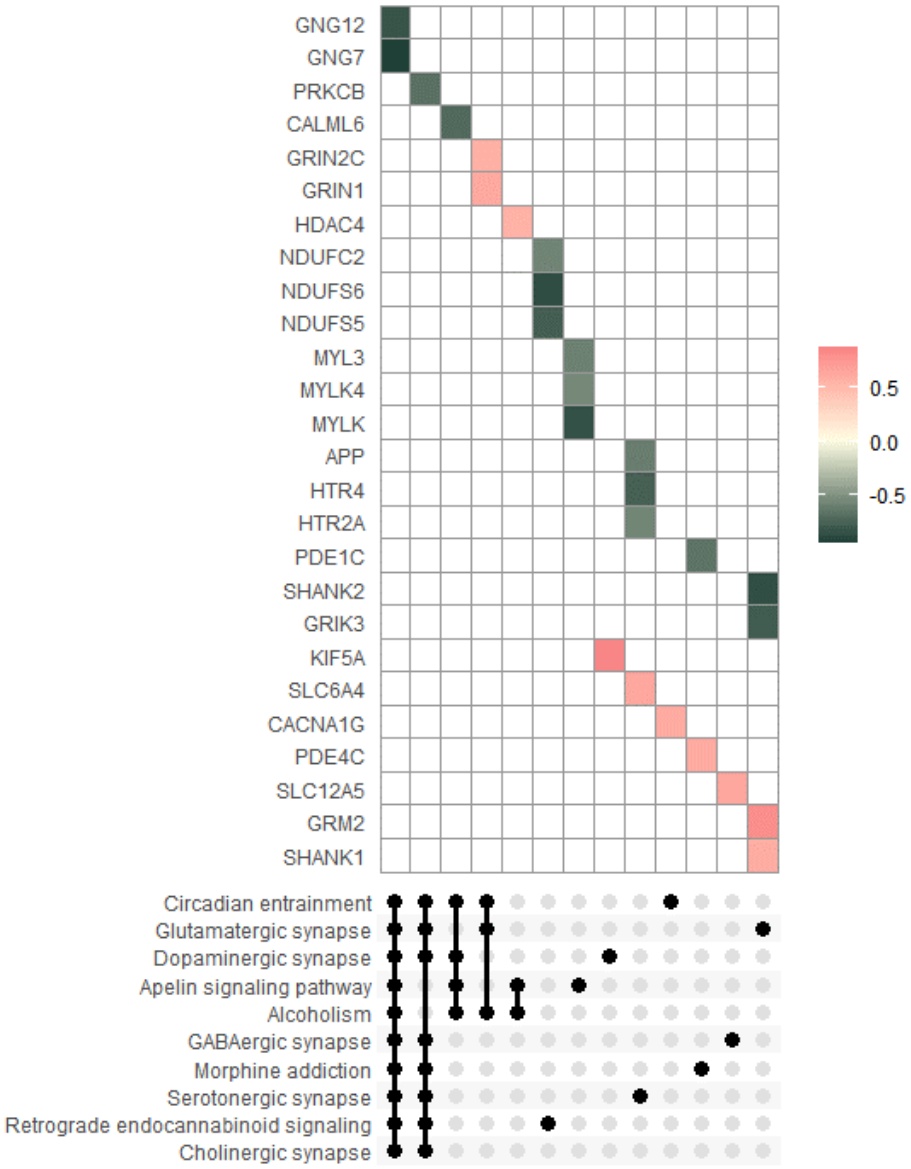

Figure 4

Comprehensive analysis of age-correlated genes using upset plot and network analysis. (a) Upset plot of age-correlated genes and ten enrichment terms. In the heatmap, red and green indicate positively and negatively correlated genes, respectively. (b) Network analysis of age-related genes.

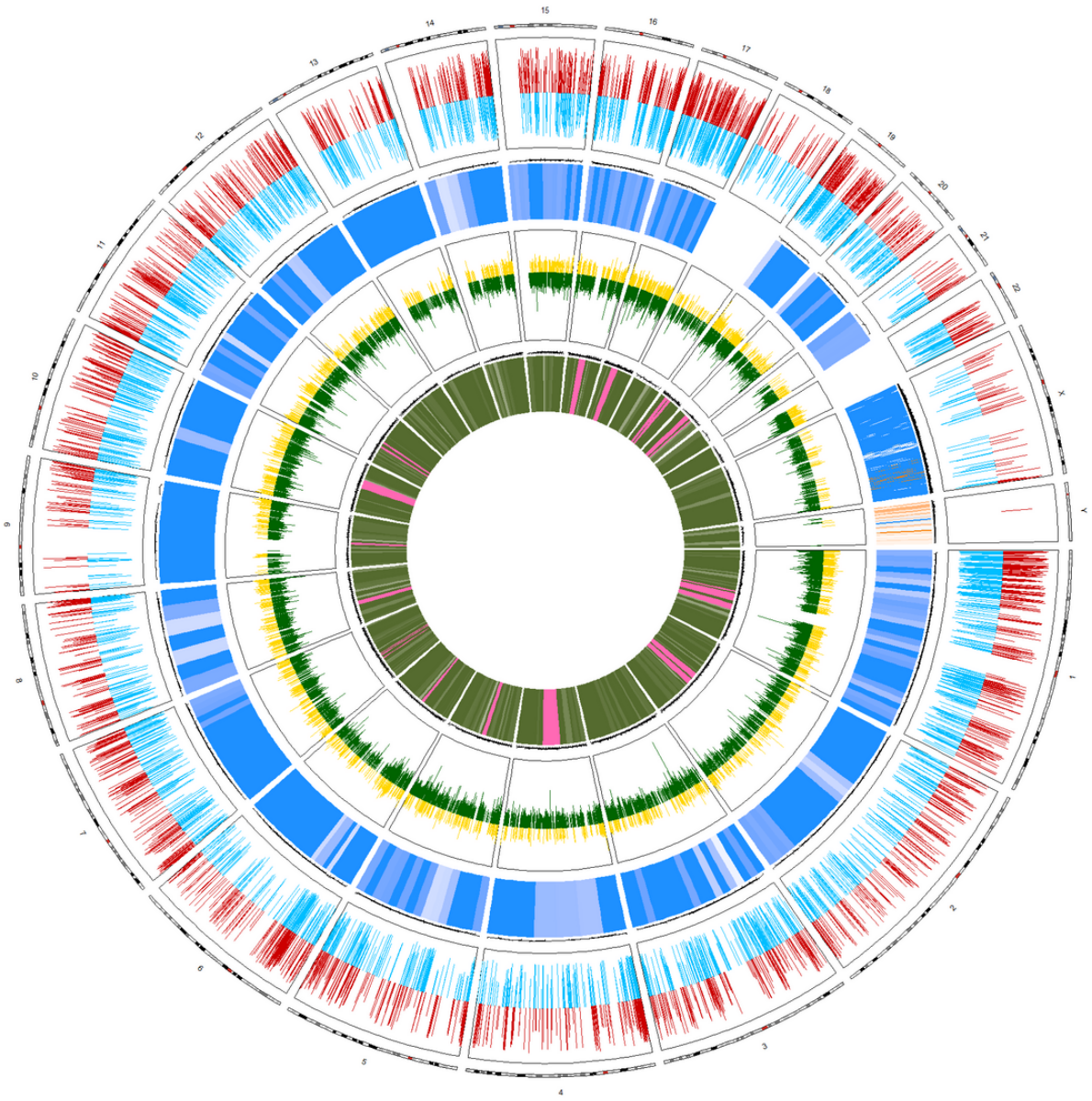

Figure 5

Circos plot indicates the genomic location of the four features (from outer to inner circles), age correlations of peak, differently methylated regions between two sexes of heatmap, ratios between two longitudinal times of peak, and differences between two longitudinal times of heatmap. Continuous variables are indicated as peaks (positively correlated peaks are shown in red and yellow, and negatively correlated peaks are shown in blue and green, respectively). Fold changes (FCs) indicated as heatmap, and the two sexes and two longitudinal time-points (baseline; BL and follow-up; FU) were compared. Hypermethylation with high FC in males and FU are shown in blue and pink, and hypermethylation with low FC in females and BL are shown in orange and green.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supportingmaterials11.docx