

The Multi-State Epigenetic Pacemaker enables the identification of combinations of factors that influence DNA methylation

Colin Farrell^{1,4}, Keshiv Tandon¹, Roberto Ferrari², Kalsuda Lapborisuth¹, Rahil Modi¹, Sagi Snir³, and Matteo Pellegrini^{1,4}

¹Dept. of Molecular, Cell and Developmental Biology;
University of California, Los Angeles, CA 90095, USA;;

²Dept. of Chemistry, Life Sciences and Environmental Sustainability, Laboratory of Molecular Cell Biology of the Epigenome (MCBE), University of Parma, Italy;

³Dept. of Evolutionary Biology, University of Haifa, Israel;

⁴Corresponding Authors; colinpfarrell@gmail.com, matteop@mcdb.ucla.edu

Abstract

Epigenetic clocks, DNA methylation based predictive models of chronological age, are often utilized to study aging associated biology. Despite their widespread use, these methods do not account for other factors that also contribute to the variability of DNA methylation data. For example, many CpG sites show strong sex-specific or cell type specific patterns that likely impact the predictions of epigenetic age. To overcome these limitations, we developed a multidimensional extension of the Epigenetic Pacemaker, the Multi-State Epigenetic Pacemaker (MSEPM). We show that the MSEPM is capable of accurately modeling multiple methylation associated factors simultaneously, while also providing site specific models that describe the per site relationship between methylation and these factors. We utilized the MSEPM with a large aggregate cohort of blood methylation data to construct models of the effects of age, sex and cell type heterogeneity on DNA methylation. We found that these models capture a large fraction of the variability at thousands of DNA methylation sites. Moreover, this approach allows us to identify sites that are primarily affected by aging and no other factors. An analysis of these sites reveals that those that lose methylation over time are enriched for CTCF transcription factor chip peaks, while those that gain methylation over time are associated with bivalent promoters of genes that are not expressed in blood. These observations suggest mechanisms that underlie age associated methylation changes and suggest that age associated increases in methylation may not have strong functional consequences on cell states. In conclusion, the MSEPM is capable of accurately modeling multiple methylation associated factors and the models produced can illuminate site specific combinations of factors that affect methylation dynamics.

1 Introduction

DNA methylation, the addition of a methyl group to the fifth carbon of the cytosine pyrimidine ring, is associated with the topological organization of the cellular genome, gene expression and the state of a cell. Within a population of cells the methylation pattern at certain sites can change

34 predictably with the age of the individual from which the cells are drawn.
35 This predictable nature of DNA methylation has led to the development
36 of accurate DNA methylation based predictive models for age and health,
37 termed epigenetic clocks. The difference between the predicted and the
38 expected epigenetic age given an individual's chronological age has been
39 interpreted as a measure of age acceleration[1], and has been associated
40 with mortality[2, 3] and other adverse health outcomes[4–8].

41 However, epigenetic clocks suffer from several limitations that limit
42 the interpretability of their predictions and the underlying mechanisms.
43 Epigenetic clocks are generally trained by using penalized regression based
44 methods that attempt to minimize the difference between the predicted
45 and observed value of age. As a result, as the error between predicted
46 and observed age is decreased, the associations between age acceleration
47 and mortality disappears[9]. Second generation epigenetic clocks attempt
48 to resolve this issue by fitting a measure of human health, rather than
49 age, and as a result these clocks are generally more sensitive to individual
50 health status[10–12]. However, while the response variable is modified in
51 these clocks the method used to fit the clock is largely the same. Epigenetic
52 clocks are generally trained using regularized regression models,
53 where the likelihood is maximized by minimizing the difference between
54 the observed and predicted response variable subject to the elastic net
55 penalty, λ_1 and λ_2 . Methylation sites that increase model error and are
56 influenced by other relevant factors such as smoking or obesity, may be
57 discarded during model fitting, thus limiting the ability of this approach
58 to account for the effects of these extraneous factors on epigenetic aging.

59 As an alternative to penalized regression based methods we previously
60 developed an evolutionary based model for epigenetic dynamics, the Epi-
61 genetic Pacemaker (EPM)[13, 14]. The EPM attempts to minimize the
62 difference between observed and predicted methylation values amongst a
63 collection of sites through the implementation of a conditional expectation
64 maximization algorithm[15]. Under the EPM the observed methylation
65 status of a collection of sites is modeled linearly with respect to an input
66 factor of interest, such as age. A hidden epigenetic state, that is related
67 to the initial factor, but not necessarily linearly, is learned through the
68 course of model fitting. The EPM can capture the non-linear relationship
69 between methylation and age[16] and outputs an interpretable model for
70 each site. However, both the EPM and regression based methods suffer
71 from the same limitation, which is that they are limited to a single trait
72 predicted by, or used to model, observed methylation patterns. In reality,
73 the observed methylation landscape is likely impacted by a variety of
74 factors that act simultaneously to produce the observed methylome of an
75 individual.

76 To overcome this limitation, we have developed a multidimensional
77 extension of the EPM, the Multi-State Epigenetic Pacemaker (MSEPM).
78 We show that the (MSEPM) can accurately model site specific methyla-
79 tion variation driven by several factors, and given a trained model, ac-
80 curately predict the values of the factors associated with an individual's
81 observed methylation profile in both simulated methylation datasets and
82 a large aggregate blood tissue methylation dataset. Importantly, as fac-
83 tors that explain the observed methylation profile of an individual are

84 added to the model the ability to model the factors and methylation
 85 values improves. Additionally, we show that sites with similar associa-
 86 tions to modeled factors cluster together and are enriched for specific
 87 transcription factors. Therefore, unlike traditional epigenetic clocks, the
 88 MSEPM allows us to study mechanisms that may underlie age associated
 89 methylation changes. In our large dataset of blood samples, we find that
 90 sites that increase methylation with age are enriched for bivalent promoters,
 91 and are proximal to genes that are lowly expressed in blood. These
 92 results suggest that positively age associated sites may not have a sig-
 93 nificant functional impact on aging traits. The MSEPM is available as
 94 a Python package with scikit-learn style syntax under a MIT license at
 95 <https://github.com/NuttyLogic/MultistateEpigeneticPacemaker>.

96 2 Methods

97 2.1 Multi-State Epigenetic Pacemaker Model

98 The MSEPM model describes the observed methylation at site i and for
 99 individual j , $\hat{m}_{i,j}$, as a weighted linear combination of k individual epi-
 100 genetic factors $p_{j,k}$.
 101

$$\hat{m}_{i,j} = r_i^0 + \sum_{k=1}^n p_{j,k} r_{i,k}$$

102 Where k epigenetic factors are weighted by k site specific epigenetic rates
 103 of change, $r_{i,k}$, and offset by a sites specific intercept term, r_i^0 . Site pa-
 104 rameters, $r_{i,k}$ and r_i^0 , are characteristic of the site and shared amongst all
 105 individuals while epigenetic factors, $p_{j,k}$, are characteristic of an individ-
 106 ual and are the same across all sites for that individual. In practice, the
 107 observed methylation value is also dependent on a normally distributed
 108 error term $\epsilon_{i,j}$.
 109

$$\hat{m}_{i,j} = r_i^0 + \sum_{k=1}^n p_{j,k} r_{i,k} + \epsilon_{i,j}$$

110 Under this model epigenetic factors are related to observable individ-
 111 ual factors $p_{k,j}^0$, such as chronological age, sex and cell types, but may
 112 be transformed relative to observable factors. The epigenetic age factor,
 113 for example, often has a non-linear relationship with the observed age[16].
 114 The MSEPM learns the appropriate transformation during model fitting
 115 to describe the observed methylation status linearly in terms of the epi-
 116 genetic age factor, but not linearly with age.

117 Given an input matrix $\hat{M} = [m_{i,j}]$ of methylation values for i sites
 118 and matched observable epigenetic factors $\hat{P}^0 = [p_{j,k}^0]$ for j individuals
 119 the objective of the MSEPM is to find the optimal values of $r_{i,k}$ and $p_{j,k}$
 120 that minimize the residual sum of square (RSS) error,

$$\epsilon_{i,j}^2 = (m_{i,j} - r_i^0 - \sum_{k=1}^n p_{j,k} r_{i,k})^2$$

121 This is accomplished through the implementation of a conditional ex-
 122 pectation maximization algorithm. The maximum likelihood (ML) values

of $r_{i,k}$ and r_i^0 can be solved using ordinary least squares (OLS) regression. Provided the ML estimates for $r_{i,k}$, the site coefficients are fixed and epigenetic factors, $p_{j,k}$, are updated by minimizing the RSS across all i sites using gradient descent,

$$p_{j,k}^{n+1} = p_{j,k}^n - \lambda \nabla F(p_{j,k})$$

where λ is a specified learning rate. The optimization is accomplished by alternating between optimizing $r_{i,K}$ and $p_{j,k}$ until the reduction in sum of the site RSS is below a specified threshold or a set number of iterations is reached. Importantly, while the ML values of $p_{j,k}$ are by definition linear with the methylation status at any site, the original input factors for $p_{j,k}^0$ may not be.

Provided a trained MSEPM model and an unobserved methylation matrix, epigenetic factors are estimated by calculating each independent OLS for solution all i sites given the $r_{i,k}$ coefficients set for the respective input factor. These epigenetic factors can then be used to find the expected methylation value using the trained individual site models where

$$E[m_{i,j}] = r_{i,0} + P_j \dot{R}_i$$

where $P_j \dot{R}_i$ is a matrix of point values p and r.

2.2 MSEPM Simulation Framework

We implemented a simulation framework using the MSEPM formulation evaluate the performance of the MSEPM model under various conditions. To simulate the association between methylation status and an observable factor we modeled the epigenetic factor $p_{k,j}$ as function of time, or age, and magnitude, $p_{k,j}^0$ with a non-linear transformation γ_k , where $p_{k,j} = Age_j p_{k,j}^{0,\gamma_k}$. In practice the value of the $p_{k,j}$ is often unknown and the association between methylation status and $p_{k,j}$ is inferred through the observable factor $p_{k,j}^0$.

Methylation sites were simulated by first randomly setting the range of the methylation site, $-1 < \delta < 1$ a site intercept, r_i^0 , and the site error, $\sigma_i \sim \mathcal{U}(0.025, 0.05)$. The possible range of the methylation site is described by the initial methylation value, m_0 $\beta(.2, .2)$, and the target methylation value, m_t , where the range is $\delta_i = m_t - m_0$. $\beta(.2, .2)$ is the beta distribution with its parameters for randomly setting the sample. The value of m_t is set conditionally to ensure site variability is always larger than some specified threshold, θ , where $\theta \leq |\delta| \geq .r_i^0 \beta(.2, .2)$.

Simulated methylation sites are then randomly associated with a combination of zero, one, or multiple epigenetic factors. Rates for sites associated with multiple factors were set by sampling from a uniform distribution. The weighted factor rates are normalized so the input combination of traits describes the range of the simulated site, δ . If a site is associated with no factors the observed methylation status of a site is described by a random normal with a characteristic offset, $\hat{m}_i = r_{0,i} + N(\mu, \sigma)$.

2.3 Blood MSEPM Model Training

MSEPM models were trained using a large aggregate dataset of blood derived methylation data from 17 publicly available datasets[7, 17–32]. Illumina methylation 450K Beadchip methylation array IDAT files were processed using minfi[33] (v1.34.0). Sample IDAT files were processed in batches according to GEO series and Beadchip identification. Methylation values within each batch were normal-exponential normalized using out-of-band probes[34]. Blood cell types counts were estimated using a regression calibration approach[35] and sex predictions were made using the median intensity measurements of the X and Y chromosomes as implemented in minfi[33]. Samples were filtered for quality control using the relative intensity of the methylated and unmethylated probes. Samples were used for downstream analysis if the sample median methylation probe intensity was greater than 10.5 and the difference between the observed and expected median unmethylation probe intensity is less than 0.4, where the expected median unmethylated intensity is described by $E[intensity_{unmethylated}] = 0.66intensity_{methylated} + 3.718$. This resulted in a total of 5687 samples.

We trained MSEPM models using data assembled from four GEO series[20, 22, 29, 36] ($n = 1605$). The samples were randomly split into training ($n = 1203$) and validation ($n = 402$) sets stratified by age. Methylation values for all samples were quantile normalized by probe type[37] using the median site methylation values across all training samples for each methylation site. Training set blood cell type abundance estimates were used to train a principal component analysis (PCA) model which was then used to calculate cell type PCA estimates for the validation and testing sets. Methylation sites were selected for modeling with MSEPM if the site methylation values were correlated with age ($n = 276$), sex ($n = 49$), CT-PC1 ($n = 120$), CT-PC2 ($n = 116$) or a combination of factors ($n = 238$) by absolute pearson correlation coefficient. Where a absolute pearson correlation coefficient greater than 0.7, 0.995, 0.92 and 0.64 for age, sex, CT-PC1 and CT-PC2 respectively. Sites with a sum of absolute pearson coefficients across the four factors greater than 1.8 were also included ($n = 238$) for a total of 778 methylation sites. Min-max, (0-1), scalers were fit using the training input features. Validation and testing sample features were transformed with the trained scalers. Age was min-max scaled on a range from 0-100 years. MSEPM models were trained with a learning rate of 0.01 with an iteration limit of 200.

2.4 Blood MSEPM Model Cluster Transcription Factor Overlap Analysis

We evaluated the relationship between modeled sites, input factors and regulatory transcription factors using overlap enrichment analysis. We built a custom transcription factor reference set using ENCODE V4 transcription factor chromatin immunoprecipitation[38, 39] (release 1.4.0 - 2.1.2) irreproducible discovery rate narrow bed peaks, which contains peaks with high rank consistency between replicates, that were not audited for non-compliance or errors. GRCh38 region coordinates were lifted

194 to GRCh37 coordinates using liftOver[40]. The overlap reference contains
195 714 transcription factor targets from 1621 accession IDs.

196 We then performed hierarchical clustering of the four factor MSEPM
197 model sites based on the similarity of their regression coefficients. Indi-
198 vidual methylation site coefficients were first normalized by the standard
199 deviation of methylation values of the site among the training samples,
200 $r_{i,k}/\sigma_i$. A distance matrix was then created by taking the Euclidean
201 distance between the normalized site model coefficients. Sites were then
202 clustered using Ward's method which seeks to minimize within cluster
203 variance by minimizing the increase in the error sum of squares (ESS)
204 through successive cluster fusions. Clusters label by tree cutting at a
205 height of 18. All clustering analysis was carried out using SciPy v1.6.3[41].

206 Transcription factor enrichment analysis was performed with LOLA[42]
207 which assesses the genomic region set overlap between a set of query re-
208 gions and a set of reference regions, within a specified shared background
209 set, using Fisher's exact test. Overlap analysis was performed for sites
210 within a cluster against the ENCODE V4 reference region (1BP minimum
211 overlap) using all sites assayed with Infinium HumanMethylation450K
212 BeadChip as background.

213 2.5 Clustering sites with age-associated increases 214 in methylation

215 To better understand age associated methylation in whole blood, we ex-
216 amined each site within MSEPM four factor blood model cluster 7 indi-
217 vidually, as this cluster contains sites that have methylation that increases
218 with age but is not strongly affected by other factors. Using the EWAS
219 Data Hub (Xiong, et al. 2016), we validated our results by obtaining ad-
220 dditional methylation by age data in whole blood for each site in the cluster
221 (McCartney, et al. 2019). We created a matrix with every sample and its
222 associated methylation and age from cluster 7, then used age associated
223 methylation levels to create a clustered heatmap using the Matlab func-
224 tion Clustergram. We then clustered the tree into four groups which were
225 analyzed separately.

226 We also identified the genes that were proximal to each site using
227 Cistrome-GO (Li et al. 2019). We then examined the expression of the
228 genes across tissues in the Genotype-Tissue Expression (GTEx) database
229 database. We used the GTEx Multi Gene Query to find which tissues
230 those genes belonged to.

231 We utilized the Toolkit for Cistrome Data Browser [43, 44] for the
232 analysis of significant factors in each cluster. This allowed us to input
233 .bed files of each sub-cluster and generate a GIGGLE score for specific
234 transcription factors, histone marks, and chromatin regions to assess sig-
235 nificance of these elements. A GIGGLE score tailored ranking of loci
236 based on overlap of genomic features provided by the user[45].

237 2.6 H3K4me3 enrichment analysis

238 Enrichment of analysis for H3K4me3 (figure 7A) was carried out by down-
239 loading rpm normalized bigwig files of H3K4me3 ChIP-seq data from

240 epigenomesportal[46] for CD38+ B Cells and CD56+ NTK Cells (for both
 241 0-5 years old and 60-65 years old individuals). Heatmaps of H3K4me3
 242 were generated using deepTools2[47] using the computeMatrix and plotHe-
 243 atmap function to plot the bigwig signal over genomic regions of cluster
 244 7 as the BED input. The IGV genome browser[48] was used to generate
 245 an image of the KCTD1 and IRS2 promoter regions shown in figure 7B
 246 using downloaded bigwig tracks.

247 2.7 Analysis Environment

248 Analysis was carried out in a Jupyter[49] analysis environment. Joblib[50],
 249 SciPy[51], Matplotlib[52], Seaborn[53], Pandas[54] and TQDM[55] pack-
 250 ages were utilized during analysis.

251 3 Results

252 3.1 Simulated Methylation Associated Traits

We simulated individuals whose methylation is determined by four factors and their associated epigenetic factors: a uniformly distributed factor approximating age with a non-linear association with methylation status

$$q \sim \mathcal{U}(0, 100), s_{Age} = q^{0.5}, \text{Figure 1A-B}$$

a binary distributed trait resembling sex, linearly associated with methylation status

$$q \sim B(1, .5), s_{Sex} = q, \text{Figure 1C-D}$$

a continuous normal (CN) phenotype a linear association with methylation status

$$q \sim \mathcal{N}(1, 0.1), s_{CN} = q, \text{Figure 1E-F}$$

and a continuous exponentially (CE) distributed trait with a linear association with methylation status

$$q \sim \frac{1}{20}e^{-x/20}, s_{CE} = q, \text{Figure 1G-H}$$

253 We simulated 90 methylation sites (Figure 1I). We then evaluated
 254 the MSEPM model as follows. We simulated 1000 samples with the four
 255 epigenetic factors described above. We then simulated methylation values
 256 using the simulated site rates. Simulated samples were then split for train-
 257 ing ($n = 500$) and testing ($n = 500$). MSEPM models were then fitted
 258 using the values of the input factors, $p_{k,j}^0$. We generated 1000 simulated
 259 datasets and fit MSEPM models using four combinations of input factors
 260 (Age, Age-Sex, Age-Sex-CN, Age-Sex-CN-CE). Within each simulation,
 261 epigenetic state predictions and methylation site predictions were made
 262 for all testing samples. All models captured the nonlinear association be-
 263 tween simulated age and methylation (Supp. Figure 1). As the number of
 264 factors in the model is increased the mean absolute error (MAE) between
 265 the predicted epigenetic states and the simulated epigenetic factors de-
 266 creases (Figure 2A). Importantly, to accurately assess simulated age it is

267 necessary to account for the influence of the other simulated factors (Sex,
268 CN, CE). The MAE between the predicted and simulated methylation
269 values decreases as simulated factors are added to the model, and accu-
270 rately assessing the methylation status of a simulated site requires that
271 the factor associated with the methylation status at the site is included
272 in the model (Figure 2A).

273 The MSEPM model generated using all four simulated factors can
274 capture the relative magnitude of the simulated site-specific rates (Figure
275 2C-F). However, the model has difficulty capturing the exact relationship
276 between the simulated factors (age, CN and CE) and the inferred factors
277 (Figure 2C, E-F). This is likely due to limitations of the model at cap-
278 turing nonlinear methylation association and a limited training range for
279 normally and exponentially distributed traits. Regardless, the four-factor
280 model can accurately predict the simulated methylation value (Figure 2
281 D) and site intercept (Supp. Figure 1A). We also assessed the model ro-
282 bustness to variation in the number of samples and sites used for model
283 training by randomly selecting a reduced subset of samples or sites for
284 model training. MSEPM models trained with age, sex, CN, and CE can
285 accurately assess all simulated phenotypes with few samples and sites
286 (Supp. Figure 2 B-E).

287 3.2 Blood MSEPM Model

288 We next applied the MSEPM to real data. We utilized a large aggregated
289 dataset composed of Illumina 450k array data from 17 publicly available
290 datasets[7, 17–32] deposited in the Gene Expression Omnibus[56] (GEO)
291 generated from blood derived samples (whole blood, peripheral blood lym-
292 phocytes, and peripheral blood mononuclear cells). The aggregate data
293 spanned a wide age range (0.0 - 99.0 years, Figure 3A), contained more
294 predicted females ($n = 3392$) than males ($n = 2295$, Figure 3B) and rea-
295 sonable predicted cell type abundance estimates (Figure 3C). The first
296 principal component of a PCA modele trained cell type abundance es-
297 timates (CT-PC1) is largely driven by the relative abundance of gran-
298 uocytes (Figure 3D), while the second PC (CT-PC2) captures relative
299 differences in the abundance of differentiated lymphocytes (Figure 3D).

300 We trained MSEPM models using methylation sites ($n = 778$) that
301 were correlated with the observable input factors. MSEPM models were
302 fit using four combinations of input factors (Age, Age Sex, Age Sex CT-
303 PC1, and Age Sex CT-PC1 CT-PC2). The association between the fit epi-
304 genetic factor predictions against the input modeled factors was assessed
305 by fitting a trendline between epigenetic state predictions and scaled con-
306 tinuous input factors using the state prediction made for the MSEPM
307 model trained with all four input factors. Performance of the MSEPM
308 model was then evaluated using the testing samples ($n = 4,082$). The per-
309 formance of the MSEPM largely closely resembles the simulation results.
310 All four MSEPM models capture the nonlinear relationship between age
311 and methylation status (Supp. Figure 6). The epigenetic state prediction
312 associated with age improves as the underlying methylation data are more
313 fully explained through the addition of epigenetic factors (Supp. Figure
314 6). The MSEPM model fit with Age, Sex, CT-PC1 and CT-PC2 can

315 accurately model the associated epigenetic state for each factor (Figure
316 4 A-D) and accurately predicts the methylation levels at individual sites
317 ($R^2 = 0.935$, $MAE = 0.035$, Figure 4 E). The trained MSEPM produces a
318 collection of methylation site models that can help explain the association
319 between modeled factors and methylation status.

320 **3.3 Analysis of chromatin regulators of site clus-
321 ters**

322 We evaluated the relationship between sites that are influenced by age,
323 sex, CT-PC1 or CT-PC2 and potential regulatory factors by performing
324 overlap enrichment analysis of these sites with transcription factor chro-
325 matin immunoprecipitation peaks present in the ENCODE V4[38, 39] re-
326 lease. We first identified sites with similar coefficients of epigenetic factors
327 through hierarchical clustering. The resulting tree was cut at a height of
328 18 to produce 10 distinct clusters with clear associations to the modeled
329 factors (Figure 5A).

330 The site clusters largely conform to underlying biological expectations.
331 Cluster one contains sites that are wholly associated with sex status and
332 localized to the X chromosome (Supp. Table 1) and is enriched for peaks
333 of transcription factors associated with sex specific regulation such as
334 MAZ[57]. Clusters nine and ten contain sites whose methylation sta-
335 tus is largely driven by CT-PC1, and are enriched for transcription fac-
336 tors associated with granulocyte development (CEPB, CEBPA, EP300,
337 ETV6)[58, 59]. Similarly, clusters two, five and eight are associated with
338 CT-PC2 and are enriched for transcription factor peaks associated with
339 immune development (ZBED1, ETV6, FOSL2, FOS, TBX21). Clusters
340 four and six are associated with loss of methylation with age. Cluster six
341 is highly enriched for CTCF binding sites; CTCF is known to increase
342 at sites where methylation is lost during aging[60]. Cluster four is en-
343 riched for STAT3 whose activation during exercise is age dependent[61,
344 62]. Cluster seven is associated with the accumulation of methylation
345 with age and is enriched for immunoprecipitation peaks for aging as-
346 sociated transcription factors SMAD4 and RE1-Silencing Transcription
347 Factor (REST). SMAD4 encodes a protein involved in the transforming
348 growth factor beta (TGF- β) signaling pathway. Age related dysregulation
349 of TGF- β has been linked to reduced skeletal muscle regeneration[63, 64]
350 and SMAD4 polymorphisms are associated with longevity[65]). REST
351 is a transcriptional repressor of neuron specific genes in non-neuronal
352 cells[66, 67]. REST expression is upregulated in aged prefrontal cortex
353 tissue and the absence of REST expression is associated with cognitive
354 impairment[68] and cellular senescence in neurons[69]).

355 **3.4 Analysis of sites with age-associated increases
356 in methylation**

357 Because of our interest in the mechanisms that underlie ages associated
358 increase in methylation, we focused on cluster seven, as these sites have
359 methylation increases that depend primarily on age rather than sex and

360 cell types. Cluster 7 consisted of 93 CpG sites. To obtain an independent
361 measure of how these sites change with age, we obtained age associated
362 methylation

363 data from the EWAS Data Hub[70], with a focus on whole blood
364 methylation. The dataset consisted of about 1600 individuals with ages
365 ranging from 0 to 113 years old[71]. We clustered the sites based on age
366 associated methylation levels, meaning the rate of methylation based on
367 age for each marker. Each site was organized into an ordered matrix with
368 methylation levels at each age, then grouped into four sub-clusters: A,
369 B, C, and D. As seen in Figure 6A, Cluster A had the highest average
370 methylation across ages, and each consecutive cluster had a decrease in
371 average methylation. We next examined chromatin accessibility, tran-
372 scription factors, histone marks, and genes associated with each cluster.
373 As shown in Supp. Figure 7, genes proximal to Cluster 7 sites were lowly
374 expressed in blood compared to other tissues. We analyzed chromatin ac-
375 cessibility, transcription factors, and histone marks associated with these
376 four groups. We computed levels of H3K27ac, H3K27me3, H3K4me3, and
377 H3K9me3 across the four subclusters. As seen in Figure 6C, H3K4me3
378 increased from clusters A through D. Figure 6E shows that H3K27ac in-
379 creased from clusters A through C, but then decreased in D. These results
380 suggest that subcluster D is enriched for bivalent domains, characterized
381 by H3K4me3 and H3K27me3.

382 Based on these results we hypothesize that the mechanisms that un-
383 derlies the gain of methylation with age at these bivalent promoters is the
384 age-associated loss of H3K4me3. It is well established that the presence
385 of trimethylation on H3K4 inhibits de novo methylation, and this effect
386 explains the hypomethylation that is typical of promoters, including biva-
387 lent promoters. We therefore hypothesize that the gain of methylation at
388 these sites may be caused by an age associated loss of H3K4me3. In or-
389 der to demonstrate that H3K4me3 decreases with age for genomic regions
390 where DNA methylation increases, we used published H3K4me3 ChIP-seq
391 data from epigenomesportal[46]. We selected two different blood cell types
392 CD38+ B Cells and CD56+ NTK Cells and plotted the H3K4me3 signal
393 of young (0 to 5 years old) versus old individuals (60 to 65 years old) over
394 genomic regions of cluster 7 (Figure 7A). Our analysis shows that younger
395 individuals have higher levels of H3K4me3 compared to older ones (Figure
396 7A) as also shown for two selected genomic loci of cluster 7 (the promoters
397 of KCTD1 and IRS2 genes) where we can observe a marked decrease in the
398 levels of H3K4me3 as age increases (Figure 7B). All together these data
399 suggest that genomic regions whose DNA methylation is increased with
400 age exhibit an age dependent loss of H3K4me3, thus showing an inverse
401 correlation between DNA methylation and H3K4me3 at these genomic
402 loci.

403 4 Discussion

404 Epigenetic clocks are widely used tools to study human aging and health.
405 Despite their widespread use, the biological interpretability of the mod-
406 els is limited. A methylome is influenced by many different biological

407 processes occurring simultaneously over time that may differ among individuals.
408 Epigenetic clocks, while producing accurate predictions of age,
409 attempt to capture this complexity through a single dependent variable.
410 Additionally, the penalized regression based methods used to fit most epigenetic
411 clocks select sites that minimize, or regress out, the influence of other factors and
412 omit groups of sites that are correlated. To overcome these limitations, here we propose a
413 multidimensional extension of the EPM model, the MSEPM.
414

415 In contrast to previous methods, the MSEPM aims to simultaneously
416 model the effect of multiple factors on the methylome. The simulation
417 and blood MSEPM models show that concurrently modeling age, cell
418 type composition and sex can minimize model residuals when compared
419 with the MSEPM model fit with age only. The residual of the age only
420 model is often interpreted as a measure of age acceleration. When multiple
421 methylome associated traits are modeled simultaneously this residual can
422 be explained directly by other factors and the association between the
423 methylome and a trait of interest can be inferred.

424 Additionally, the individual methylation site linear models fit as part of
425 the MSEPM optimization can provide information about the relationship
426 between modeled factors and site specific biology. To this end, we find
427 that the blood MSEPM model conforms to expected biology. Sites with a
428 strong sex association localize to the X chromosome and sites associated
429 with cell types are enriched for transcription factors associated with the
430 development of immune cells.

431 CpG sites that are primarily affected only by age in the blood MSEPM
432 model are of particular interest. As others have previously described,
433 sites that progressively lose methylation over time are strongly enriched
434 for CTCF[72, 73]. As CTCF plays a key role in long range chromatin
435 interactions, this may suggest that there are age-associated changes in
436 three dimensional chromatin structure, and that the structure may be-
437 come more disordered with age. In fact, alterations in CTCF binding and
438 function with age have been implicated in the pathogenesis of various age-
439 related diseases, including cancer. For example, changes in the chromatin
440 structure and gene expression due to altered CTCF binding can contribute
441 to the genomic instability and altered cell proliferation characteristic of
442 cancerous cells (Hnisz et al., 2016; Phillips et al., 2009).

443 We identified a cluster of sites that showed increasing methylation
444 with age and that were not significantly affected by other factors. We
445 found that these sites are enriched for the transcription factor REST.
446 The RE1-Silencing Transcription Factor (REST), also known as Neuron-
447 Restrictive Silencer Factor (NRSF), is a key regulatory protein involved
448 in the development and differentiation of neurons. It plays a crucial role
449 in neurogenesis, neuronal differentiation, and in the maintenance of the
450 neuronal phenotype by regulating gene expression[74]. REST achieves
451 this by binding to the neuron-restrictive silencer element (NRSE) or RE1
452 sites in the DNA, leading to the repression of gene transcription in non-
453 neuronal cells or in neuronal progenitor cells, ensuring that neuronal genes
454 are expressed only in neurons[66, 75, 76]. The fact that this factor is
455 enriched at the positively age-associated sites suggests that these sites are
456 likely expressed in neuronal cells but not in blood. In fact this is what we

457 find when we examine the tissue specific expression of the genes proximal
458 o these sites.

459 We also examined the histone modifications associated with the posi-
460 tively age-associated sites and found that they were enriched for H3K4me3
461 and H3K27me3. These sites are characteristic of bivalent promoters. Bi-
462 valent promoters play a crucial role in the regulation of gene expression
463 during development and differentiation. Characterized by the simultane-
464 ous presence of both activating (H3K4me3) and repressive (H3K27me3)
465 histone modifications, bivalent promoters mark genes that are poised for
466 transcription but are not actively transcribed. This dual modification
467 serves as a regulatory mechanism, ensuring that genes essential for differ-
468 entiation and development are ready to be activated at the appropriate
469 time. Bivalent domains are predominantly found in embryonic stem cells
470 and are crucial for maintaining the cells in a pluripotent state, allowing
471 for the rapid activation or repression of gene expression in response to de-
472 velopmental cues. The significance of bivalent promoters extends to their
473 role in cell fate decisions, where they contribute to the tight control of de-
474 velopmental pathways and the maintenance of stem cell identity[77, 78].
475 Our results suggest that the bivalent promoters we identified in blood are
476 inactive (as seen by the fact that the proximal genes are not expressed).
477 However, the fact that DNA methylation at these sites increases with age
478 suggests that they may be losing H3K4me3 with age. H3K4me3 is a crit-
479 ical regulator of DNA methylation as it inhibits the binding of DNMT3
480 to histones, as the DNMT3 ADD domain preferentially binds to the un-
481 methylated H3K4 residue[79]. This explains why promoters, which are
482 enriched for H3K4me3, are generally hypomethylated. Our results sug-
483 gests that there must therefore be an age associated loss of H3K4me3 at
484 these bivalent promoters. That is in fact what we saw when we examined
485 these marks in B cells and Nk cells of both young and old individuals.
486 These mechanisms further suggest that the age associated DNA methyla-
487 tion increases may not have a functional consequence in blood and that
488 their proximal genes remain repressed throughout life.

489 In conclusion, we introduced a multi-dimensional extension of the Epi-
490 genetic Pacemaker, the MSEPM. The MSEPM is capable of accurately
491 modeling multiple methylation associated factors simultaneously. This
492 paradigm can elucidate the site specific regulation underpinning methy-
493 lome dynamics. It allows us to characterize the mechanisms underlying
494 age associated increases in methylation sites, suggesting that these were
495 caused by the loss of H3K4Me3 at bivalent promoters of genes that are
496 silenced in blood. The MSEPM is available under the MIT license at
497 <https://github.com/NuttyLogic/MultistateEpigeneticPacemaker>.

498 4.1 Supplementary Information

499 All analysis code, data processing code, and supplementary material asso-
500 ciated with this manuscript can be found at <https://github.com/NuttyLogic/MSEPMManuscript>.
501 The methylation simulation utility can be found at <https://github.com/NuttyLogic/MethSim>.
502 The data supporting these findings are openly available at GEO un-
503 der the series GSE87640, GSE87648, GSE51057, GSE51032, GSE87571,
504 GSE125105, GSE42861, GSE69138, GSE111629, GSE128235, GSE121633,

505 GSE73103, GSE61496, GSE59065, GSE97362, GSE156994, GSE128064
506 and GSE43976.

507 5 Acknowledgments

508 This work has benefited from the equipment and framework of the COMP-
509 HUB and COMP-R Initiatives, funded by the ‘Departments of Excellence’
510 program of the Italian Ministry for University and Research (MIUR, 2018-
511 2022 and MUR, 2023-2027).

512 6 Ethical Statement/Conflict of Interest

513 We have no conflicts of interest to disclose.

References

1. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. en. *Nat. Rev. Genet.* **19**, 371–384 (June 2018).
2. Perna, L. *et al.* Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort 2016.
3. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. en. *Genome Biol.* **16**, 25 (Jan. 2015).
4. Dugué, P.-A. *et al.* DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. en. *Int. J. Cancer* **142**, 1611–1619 (Apr. 2018).
5. Huang, R.-C. *et al.* Epigenetic Age Acceleration in Adolescence Associates With BMI, Inflammation, and Risk Score for Middle Age Cardiovascular Disease. en. *J. Clin. Endocrinol. Metab.* **104**, 3012–3024 (July 2019).
6. Armstrong, N. J. *et al.* Aging, exceptional longevity and comparisons of the Hannum and Horvath epigenetic clocks. en. *Epigenomics* **9**, 689–700 (May 2017).
7. Chuang, Y.-H. *et al.* Parkinson’s disease is associated with DNA methylation levels in human blood and saliva. en. *Genome Med.* **9**, 76 (Aug. 2017).
8. Horvath, S. & Ritz, B. R. Increased epigenetic age and granulocyte counts in the blood of Parkinson’s disease patients. en. *Aging* **7**, 1130–1142 (Dec. 2015).
9. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing 2019.
10. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. en. *Aging* **11**, 303–327 (Jan. 2019).

11. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. en. *Aging* **10**, 573–591 (Apr. 2018).
12. Belsky, D. W. *et al.* Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. en. *Elife* **9** (May 2020).
13. Farrell, C., Snir, S. & Pellegrini, M. The Epigenetic Pacemaker: modeling epigenetic states under an evolutionary framework. en. *Bioinformatics* **36**, 4662–4663 (Nov. 2020).
14. Snir, S., vonHoldt, B. M. & Pellegrini, M. A Statistical Framework to Identify Deviation from Time Linearity in Epigenetic Aging. en. *PLoS Comput. Biol.* **12**, e1005183 (Nov. 2016).
15. Snir, S. Epigenetic pacemaker: closed form algebraic solutions. en. *BMC Genomics* **21**, 257 (Apr. 2020).
16. Snir, S., Farrell, C. & Pellegrini, M. Human epigenetic ageing is logarithmic with time across the entire lifespan. en. *Epigenetics* **14**, 912–926 (Sept. 2019).
17. Venthram, N. T. *et al.* Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. en. *Nat. Commun.* **7**, 13507 (Nov. 2016).
18. Demetriou, C. A. *et al.* Methylome analysis and epigenetic changes associated with menarcheal age. en. *PLoS One* **8**, e79391 (Nov. 2013).
19. Polidoro, S. *et al.* EPIC-Italy at HuGeF. GSE51032. *Gene Expression Omnibus* (2013).
20. Johansson, A., Enroth, S. & Gyllensten, U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. en. *PLoS One* **8**, e67378 (June 2013).
21. Arloth, J. *et al.* DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. en. *PLoS Comput. Biol.* **16**, e1007616 (Feb. 2020).
22. Liu, Y. *et al.* *Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis* 2013.
23. Soriano-Tárraga, C. *et al.* Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia. en. *Hum. Mol. Genet.* **25**, 609–619 (Feb. 2016).
24. Zannas, A. S. *et al.* *Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB-driven inflammation and cardiovascular risk* 2019.
25. Kurushima, Y. *et al.* Epigenetic findings in periodontitis in UK twins: a cross-sectional study. en. *Clin. Epigenetics* **11**, 27 (Feb. 2019).
26. Voisin, S. *et al.* Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. en. *Genome Med.* **7**, 103 (Oct. 2015).

27. Tan, Q. *et al.* Epigenetic signature of birth weight discordance in adult twins. en. *BMC Genomics* **15**, 1062 (Dec. 2014).
28. Tserel, L. *et al.* Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. en. *Sci. Rep.* **5**, 13107 (Aug. 2015).
29. Butcher, D. T. *et al.* CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions. en. *Am. J. Hum. Genet.* **100**, 773–788 (May 2017).
30. Dabin, L. C. *et al.* Altered DNA methylation profiles in blood from patients with sporadic Creutzfeldt-Jakob disease. en. *Acta Neuropathol.* **140**, 863–879 (Dec. 2020).
31. Marabita, F. *et al.* An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. en. *Epigenetics* **8**, 333–346 (Mar. 2013).
32. Del Valle J *et al.* GSE128064 Title of the publication associated with this dataset: Comprehensive constitutional genetic and epigenetic characterization of Lynch-like individuals. Mar. 2019.
33. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. en. *Bioinformatics* **30**, 1363–1369 (May 2014).
34. Triche Jr, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. en. *Nucleic Acids Res.* **41**, e90 (Apr. 2013).
35. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. en. *BMC Bioinformatics* **13**, 86 (May 2012).
36. Dámaso, E. *et al.* Comprehensive Constitutional Genetic and Epigenetic Characterization of Lynch-Like Individuals. en. *Cancers* **12** (July 2020).
37. Horvath, S. *DNA methylation age of human tissues and cell types* 2013.
38. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. en. *Nature* **489**, 57–74 (Sept. 2012).
39. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. en. *Nucleic Acids Res.* **46**, D794–D801 (Jan. 2018).
40. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. en. *Nucleic Acids Res.* **34**, D590–8 (Jan. 2006).
41. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* (Feb. 2020).
42. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. en. *Bioinformatics* **32**, 587–589 (Feb. 2016).

43. Zheng, R. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. en. *Nucleic Acids Res.* **47**, D729–D735 (Jan. 2019).
44. Mei, S. *et al.* Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. en. *Nucleic Acids Res.* **45**, D658–D662 (Jan. 2017).
45. Layer, R. M. *et al.* GIGGLE: a search engine for large-scale integrated genome analysis. en. *Nat. Methods* **15**, 123–126 (Feb. 2018).
46. Bujold, D. *et al.* The International Human Epigenome Consortium Data Portal. en. *Cell Syst* **3**, 496–499.e2 (Nov. 2016).
47. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. en. *Nucleic Acids Res.* **44**, W160–5 (July 2016).
48. Robinson, J. T. *et al.* Integrative genomics viewer. en. *Nat. Biotechnol.* **29**, 24–26 (Jan. 2011).
49. Basu, A. *Reproducible research with jupyter notebooks*
50. Varoquaux, G. & Grisel, O. Joblib: running python function as pipeline jobs. *packages. python. org/joblib* (2009).
51. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* (Feb. 2020).
52. Hunter, J. D. *Matplotlib: A 2D Graphics Environment* 2007.
53. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (Apr. 2021).
54. McKinney, W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* en (“O’Reilly Media, Inc.”, Oct. 2012).
55. Da Costa-Luis, C. O. tqdm: A Fast, Extensible Progress Meter for Python and CLI. *JOSS* **4**, 1277 (May 2019).
56. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. en. *Nucleic Acids Res.* **41**, D991–D995 (Nov. 2012).
57. Lopes-Ramos, C. M. *et al.* Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. en. *Cell Rep.* **31**, 107795 (June 2020).
58. Theilgaard-Mönch, K. *et al.* Transcription factor-driven coordination of cell cycle exit and lineage-specification in vivo during granulocytic differentiation : In memoriam Professor Niels Borregaard. en. *Nat. Commun.* **13**, 3595 (June 2022).
59. Guerzoni, C. *et al.* Inducible activation of CEBPB, a gene negatively regulated by BCR/ABL, inhibits proliferation and promotes differentiation of BCR/ABL-expressing cells. en. *Blood* **107**, 4080–4089 (May 2006).
60. Tharakan, R. *et al.* Blood DNA Methylation and Aging: A Cross-Sectional Analysis and Longitudinal Validation in the InCHIANTI Study. en. *J. Gerontol. A Biol. Sci. Med. Sci.* **75**, 2051–2055 (Oct. 2020).

61. Mohamed, E. A. & Sayed, W. M. Implication of JAK1/STAT3/SOCS3 Pathway in Aging of Cerebellum of Male Rat: Histological and Molecular study. en. *Sci. Rep.* **10**, 8840 (June 2020).
62. Trencerry, M. K., Carey, K. A., Ward, A. C., Farnfield, M. M. & Cameron-Smith, D. Exercise-induced activation of STAT3 signaling is increased with age. en. *Rejuvenation Res.* **11**, 717–724 (Aug. 2008).
63. Carlson, M. E. *et al.* Relative roles of TGF- β 1 and Wnt in the systemic regulation and aging of satellite cell responses. en. *Aging Cell* **8**, 676–689 (Dec. 2009).
64. Paris, N. D., Soroka, A., Klose, A., Liu, W. & Chakkalakal, J. V. Smad4 restricts differentiation to promote expansion of satellite cell derived progenitors during skeletal muscle regeneration. en. *Elife* **5** (Nov. 2016).
65. Carrieri, G. *et al.* The G/C915 polymorphism of transforming growth factor beta1 is associated with human longevity: a study in Italian centenarians. en. *Aging Cell* **3**, 443–448 (Dec. 2004).
66. Chong, J. A. *et al.* REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. en. *Cell* **80**, 949–957 (Mar. 1995).
67. Coulson, J. M. Transcriptional regulation: cancer, neurons and the REST. en. *Curr. Biol.* **15**, R665–8 (Sept. 2005).
68. Lu, T. *et al.* REST and stress resistance in ageing and Alzheimer's disease 2014.
69. Rocchi, A. *et al.* REST/NRSF deficiency impairs autophagy and leads to cellular senescence in neurons. en. *Aging Cell* **20**, e13471 (Oct. 2021).
70. Xiong, Z. *et al.* EWAS Data Hub: a resource of DNA methylation array data and metadata. en. *Nucleic Acids Res.* **48**, D890–D895 (Jan. 2020).
71. McCartney, D. L. *et al.* An epigenome-wide association study of sex-specific chronological ageing. en. *Genome Med.* **12**, 1 (Dec. 2019).
72. De Lima Camillo, L. P., Lapierre, L. R. & Singh, R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. en. *npj Aging* **8**, 1–15 (Apr. 2022).
73. Han, Y. *et al.* New targeted approaches for epigenetic age predictions. en. *BMC Biol.* **18**, 71 (June 2020).
74. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. en. *Science* **267**, 1360–1363 (Mar. 1995).
75. Ooi, L. & Wood, I. C. Chromatin crosstalk in development and disease: lessons from REST. en. *Nat. Rev. Genet.* **8**, 544–554 (July 2007).
76. Bruce, A. W. *et al.* Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. en. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 10458–10463 (July 2004).

77. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. en. *Cell* **125**, 315–326 (Apr. 2006).
78. Voigt, P., Tee, W.-W. & Reinberg, D. A double take on bivalent promoters. en. *Genes Dev.* **27**, 1318–1338 (June 2013).
79. Ooi, S. K. T. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. en. *Nature* **448**, 714–717 (Aug. 2007).

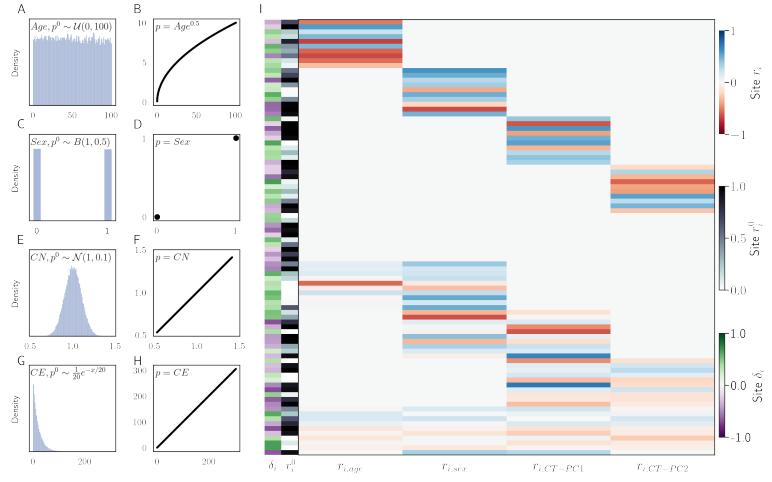


Figure 1: Simulated factors and the association with simulated methylation values. (A) Age with a non-linear association with methylation (B). Sex (C) with a binary association with methylation (D). Normal factor (E) with a linear relationship with methylation (F). Continuous exponential trait (G) with a linear relationship with methylation. (I) Simulated methylation sites. Each simulation site has a starting methylation value r_i^0 , rate of change associated with each simulated factor $r_{i,factor}$ and range of variation δ_i .

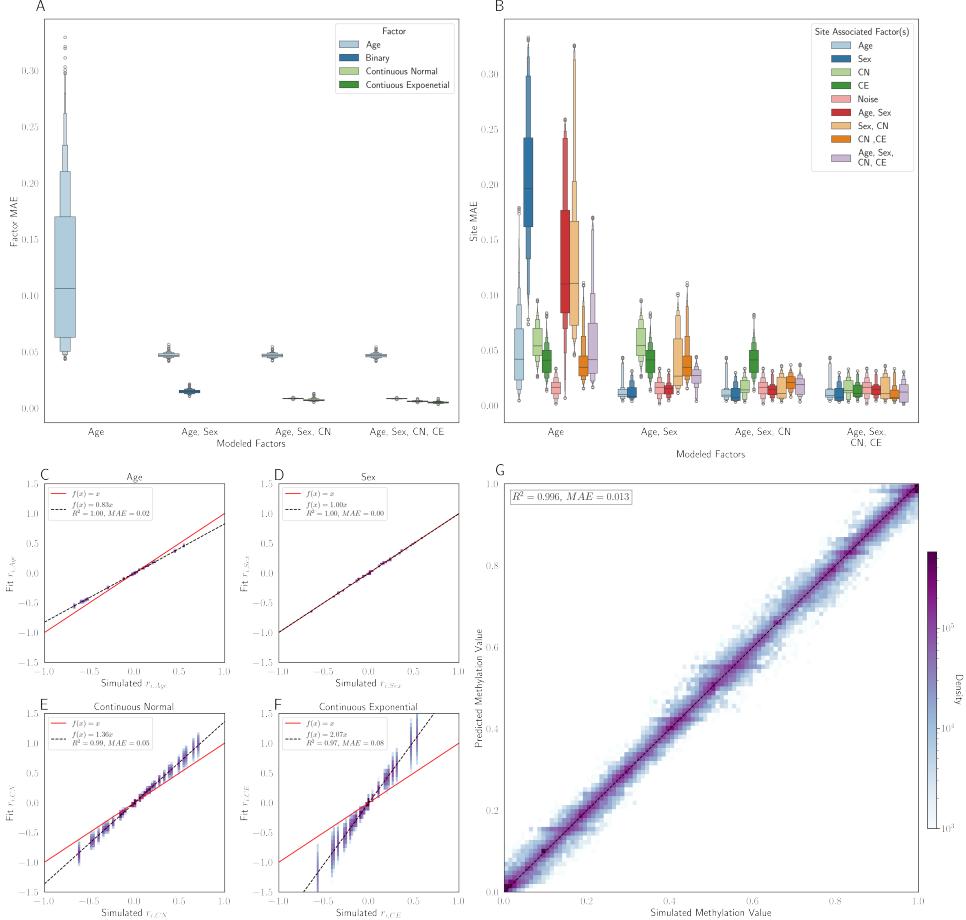


Figure 2: (A) The MAE of the factor predictions on the testing set as multiple factors are modeled simultaneously and (B) predicted methylation MAE as factors are included in the MSEPM model where the centerline is the 50th quantile and the box with greatest width contains 50% of the underlying data with each smaller box containing 50% of the remaining data with 6 levels of box width. (C) Model coefficients for Age, Sex, Continuous Normal and Continuous Exponential factors for models trained ($n = 500$) with all four simulated factors. (D). Simulated and predicted methylation values for all simulated testing sites across all training fold

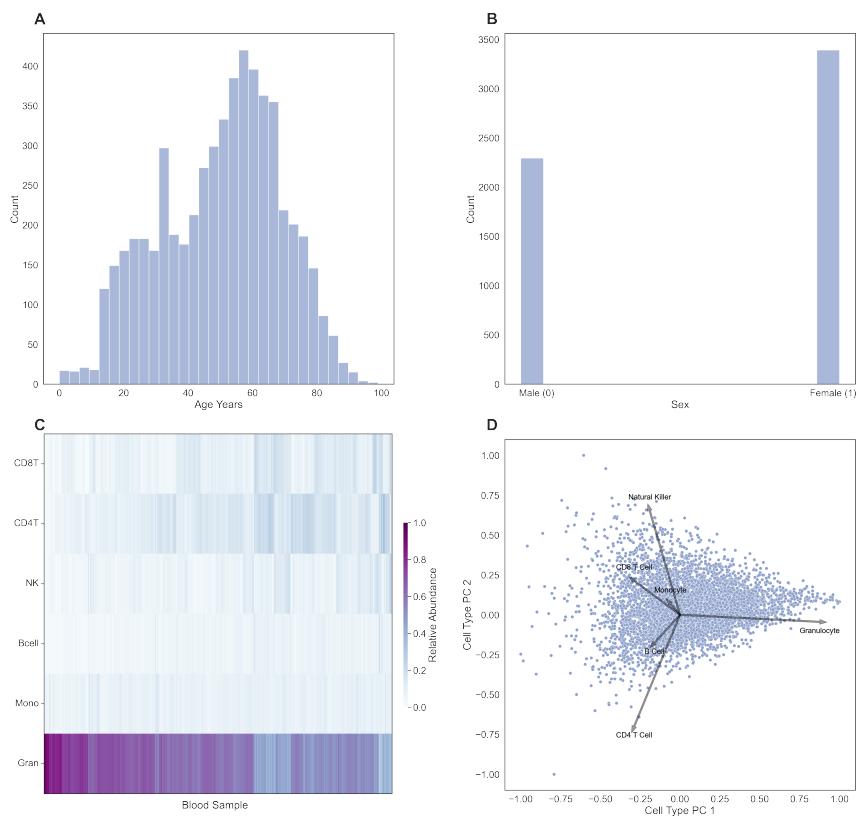


Figure 3: Distribution of age (A) and (B) sex in aggregate blood dataset. (C) Calculated cell type composition and (D) loading plot of principal components of cell type composition in aggregate blood data set.

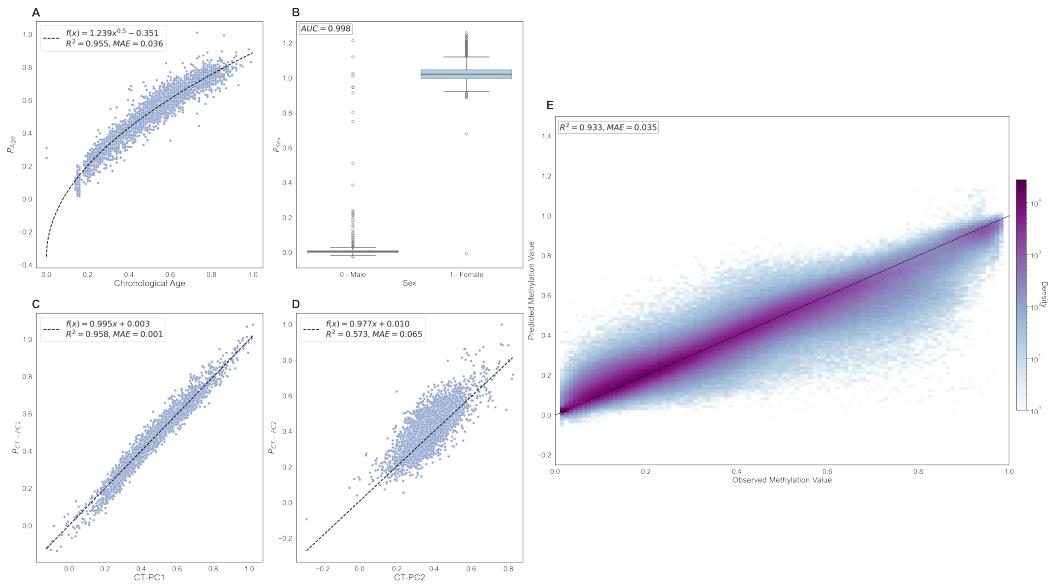


Figure 4: MSEPM model trained with age, sex, CT-PC1 and CT-PC2 predictions within testing set for epigenetic factors (A) age, (B) sex, (C) CT-PC1 and (D) CT-PC2. (E) Observed and predicted methylation values for training set has high concordance ($R^2 = 0.933$, $MAE = 0.035$)

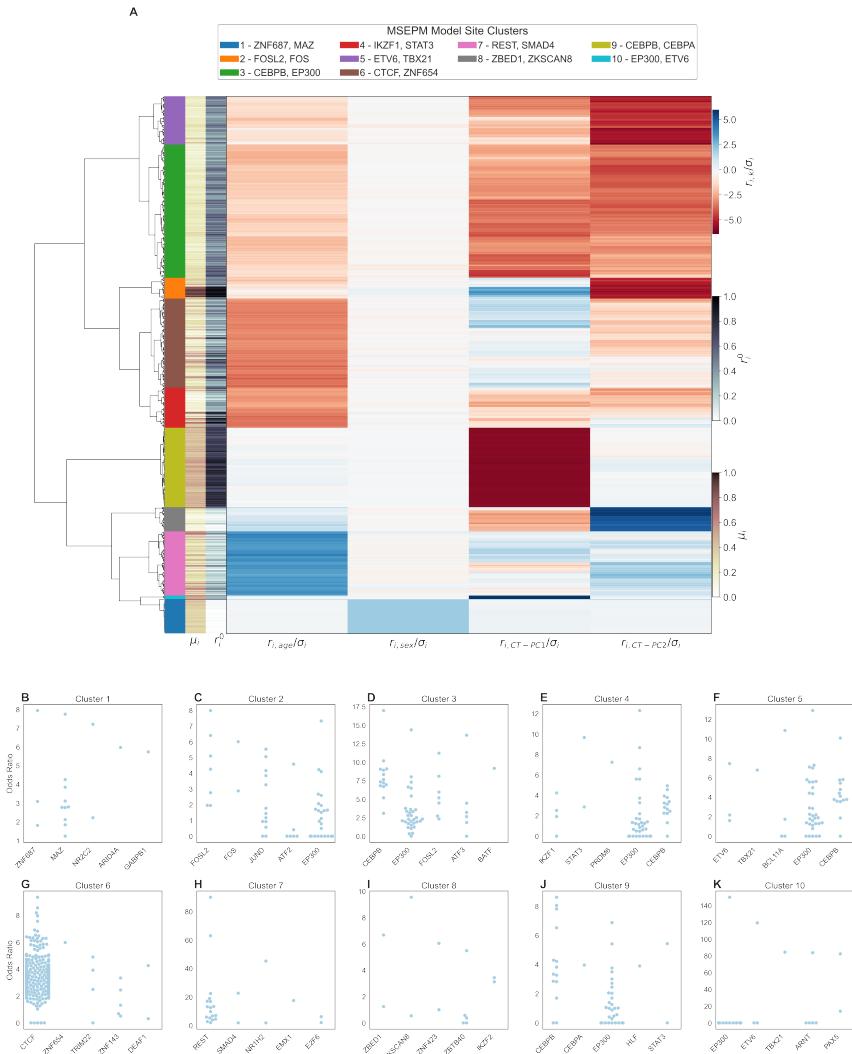


Figure5: (A) Site clustering by standardized model coefficients. Sites clusters show distinct relationships with modeled traits. (B-K) Top five enriched transcription factors for clusters 1 - 10.

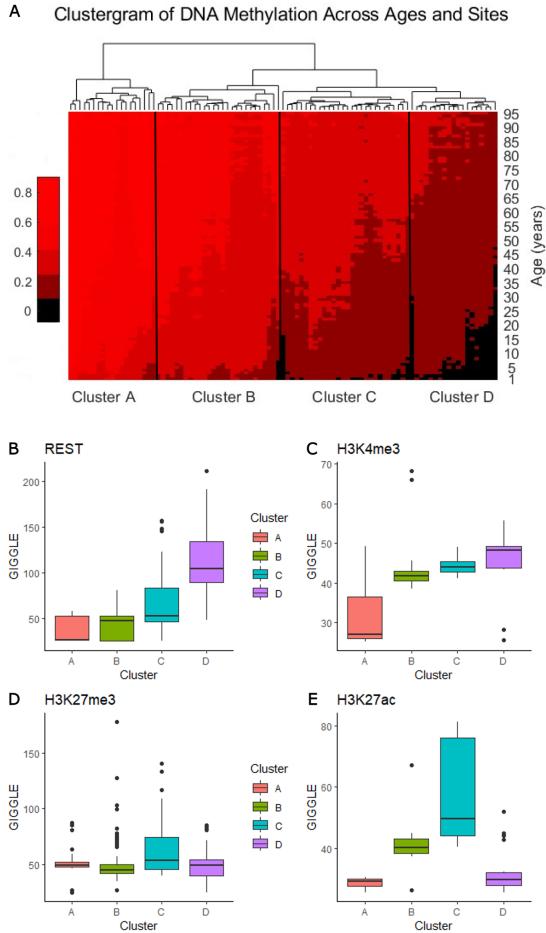


Figure6: (A) Heatmap of H3K4me3 ChIP-seq enrichment for two different blood cell types ($CD38^+$ B Cells and $CD56^+$ NTK Cells) in two cohorts of individual within 0 to 5 years old and 60 to 65 years old. The average level within 2kb up and downstream for centered genomic regions of cluster 7 is represented above the heatmap. (B) Genome browser view of H3K4me3 levels in each cohort at the promoter regions of *KCTD1* and *IRS2* genes.

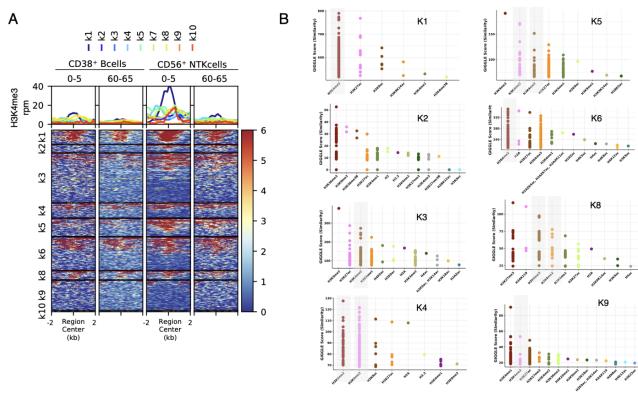


Figure 7: (A) Heatmap of H3K4me3 ChIP-seq enrichment for two different blood cell types (CD38⁺ B Cells and CD56⁺ NTK Cells) in two cohorts of individual within 0 to 5 years old and 60 to 65 years old. The average level within 2kb up and downstream for centered genomic regions of cluster 7 is represented above the heatmap. (B) Genome browser view of H3K4me3 levels in each cohort at the promoter regions of *KCTD1* and *IRS2* genes.