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

Colin Farrell[1,3], Kalsuda Lapborisuth[1], Sagi Snir[2],
and Matteo Pellegrini[1,3]

[1]Dept. of Molecular, Cell and Developmental Biology;
University of California, Los Angeles, CA 90095, USA;;
[2]Dept. of Evolutionary Biology, University of Haifa, Israel;
[3]Corresponding Authors; colinpatfarrell@g.ucla.edu, matteop@mcdb.ucla.edu

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 faction of the variability at thousands of DNA methylation sites. Moreover, we found modeled sites that are primarily affected by aging and no other factors. Among these, those that lose methylation over time are enriched for CTCF transcription factor chip peaks, while those that gain methylation over time are enriched for REST transcription factor chip peaks. Both transcription factors are associated with transcriptional maintenance and suggest a general dysregulation of transcription with age that is not impacted by sex or cell type heterogeneity. 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 predictably with the age of the individual from which the cells are drawn. This predictable nature of DNA methylation has led to the development of accurate DNA methylation based predictive models for age and health, termed epigenetic clocks. The difference between the predicted and the expected epigenetic age given an individual's chronological age has been interpreted as a measure of age acceleration[1], and has been associated with mortality[2, 3] and other adverse health outcomes[4–8].

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

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

To overcome this limitation, we have developed a multidimensional extension of the EPM, the Multi-State Epigenetic Pacemaker (MSEPM). We show that the (MSEPM) can accurately model site specific methylation variation driven by several factors, and given a trained model, accurately predict the values of the factors associated with an individual's observed methylation profile in both simulated methylation datasets and a large aggregate blood tissue methylation dataset. Importantly, as factors that explain the observed methylation profile of an individual are added to the model the ability to model the factors and methylation values improves. Additionally, we show that sites with similar associations to modeled factors cluster together and are enriched for specific transcription factors.

2 Results

2.1 Multi-State Epigenetic Pacemaker Model

The MSEPM attempts to describe the observed methylation status at any single methylation site as a linear combination of methylation associated factors specific to an individual, termed epigenetic factors. Under this model epigenetic factors are related to observable individual factors, such as chronological age, sex and cell types, but may be transformed relative to observable factors. The epigenetic age factor, for example, often has a non-linear relationship with the observed age [16]. The MSEPM learns the appropriate transformation during model fitting to describe the observed methylation status linearly in terms of the epigenetic age factor (and not linearly with age). Given a site i and individual j the observed methylation status can be modeled as $\hat{m_{i,j}} = r_i^0 + \sum_{k=1}^{n} f_{j,k} r_{i,k}$, where k epigenetic factors, $f_{j,k}$, are weighted by k site specific parameters, $r_{i,k}$, and offset by a sites specific intercept term, r_i^0. Site parameters, $r_{i,k}$ and r_i^0, are characteristic of the site and shared amongst all individuals while epigenetic factors, $f_{j,k}$, are characteristic of an individual and are the same across all sites for that individual. As a consequence, observed methylation differences between any two individuals are dependent on individual epigenetic factors $f_{j,k}$. In practice, the observed methylation value is also dependent on a normally distributed error term $\epsilon_{i,j}$, $\hat{m_{i,j}} = r_i^0 + \sum_{k=1}^{n} s_{j,n} r_{i,n} + \epsilon_{i,j}$.

Given an input matrix $\hat{M} = [m_{i,j}]$ the objective of the MSEPM is to find the optimal values of $r_{i,k}$ and $f_{k,j}$ that minimize the residual sum of square (RSS) error, $\epsilon_{i,j}^2 = (m_{i,j} - r_i^0 - \hat{\sum}_{n=1}^{n} f_{j,k} r_{i,k})^2$. This is accomplished through the implementation of a conditional expectation maximization algorithm. The maximum likelihood (ML) values of $r_{i,k}$ and r_i^0 for each i methylation site can be solved using ordinary least squares (OLS) regression. Given the ML $r_{i,k}$ site coefficients each k epigenetic factor is then updated by fixing the site coefficients and updating $s_{k,j}$ by minimizing the RSS across all i sites using gradient descent, $f_{k,j}^{n+1} = f_{k,j} - \lambda \nabla F(f_{k,j})$, where λ is a specified learning rate. The optimization is accomplished by alternating between optimizing $r_{k,i}$ and $f_{k,j}$ until the reduction in sum of the site RSS is below a specified threshold between iterations or a set number of interactions is reached. Importantly, while the ML values of $f_{j,k}$ are by definition linear with the methylation status at any site, the original input factors for $f_{j,k}$ 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 $E[m_{i,j}] = r_{i,0} + F_j \dot{R}_i$.

2.2 Simulated Methylation Associated Factors

We implemented a simulation framework that extends the MSEPM model where the methylation status at site i for an individual j is described by a weighted sum of epigenetic factors, $\hat{m_{i,j}} = r_i^0 + \sum_{l=1}^{n} f_{k,j} r_{i,k} + \epsilon_{i,j}$. As shown in our previous work [16] the association between a methylation associated input factor and methylation status is not necessarily linear. To account for non-linear associations, $f_{k,j}$, have the form $f_{j,k} = q_{j,k}^{\gamma_k}$, where γ_k is the transformation between a factor of magnitude $q_{n,j}$ and the epigenetic factor. In practice the value of the $f_{k,j}$ is often unknown and the association between methylation status and $f_{k,j}$ is inferred through $q_{n,j}$. We simulated individuals whose methylation is determined by four factors and their associated epigenetic factors: a uniform factor approximating age with a non-linear association with methylation status ($q \sim \mathcal{U}(0, 100), s_{Age} = q^{0.5}$, Figure 1A-B), a binary trait resembling sex, linearly associated with methylation status ($q \sim B(1, .5), s_{Sex} = q$, Figure 1C-D), a continuous normal (CN) phenotype resembling a cell type with a linear association with methylation status ($q \sim \mathcal{N}(1, 0.1), s_{CN} = q$, Figure 1E-F), and a continuous exponentially (CE) distributed trait resembling obesity with a linear association with methylation status ($q \sim \frac{1}{20}e^{-x/20}, s_{CE} = q, Figure G - H$). Methylation sites were simulated by first randomly setting the dynamic range of the methylation site, ($-1 < \delta < 1$), a site intercept, r_i^0, and the site error, ($\sigma_i \sim \mathcal{U}(0.025, 0.05)$). Simulated methylation sites were 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. Site rates were normalized to describe the dynamic range of the simulated methylation site unless the rates for all phenotypes were zero.

We simulated 90 methylation sites (Figure 1I). We then evaluated the MSEPM model as follows. We simulated 1000 samples with the four epigenetic factors described above. We then simulated methylation values using the simulated site rates. Simulated samples were then split for training ($n = 500$) and testing ($n = 500$). MSEPM models were then fitted using the values of the input factors, $q_{k,j}$, as the initial guess for $f_{k,j}$. We generated 1000 simulated datasets and fit MSEPM models using four combinations of input factors (Age, Age-Sex, Age-Sex-CN, Age-Sex-CN-CE). Within each simulation, epigenetic state predictions and methylation site predictions were made for all testing samples. All models captured the nonlinear association between simulated age and methylation (S. Figure 1). As the number of factors in the model is increased the mean absolute error (MAE) between the predicted epigenetic states and the simulated epigenetic factors decreases (Figure 2A). Importantly, to accurately assess simulated age it is necessary to account for the influence of the other simulated factors (Sex, CN, CE). The MAE between the predicted and simulated methylation values decreases as simulated factors are added to the model, and accurately assessing the methylation status of a simulated site requires that the factor associated with the methylation status at the site is included in the model (Figure 2A).

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

2.3 Blood MSEPM Model

We utilized a large aggregated dataset composed of Illumina 450k array data from 17 publicly available datasets [7, 17–32] deposited in the Gene Expression Omnibus [33] (GEO) generated from blood derived samples (whole blood, peripheral blood lymphocytes, and peripheral blood mononuclear cells). All datasets were processed through a unified pipeline from raw array intensity data (IDAT) files using minfi (Aryee et al., 2014). Sex and blood cell type abundance predictions were made for each as previously described[34, 35]. The aggregate data spanned a wide age range (0.0 - 99.0 years, Figure 3A), contained more predicted females ($n = 3392$) than males ($n = 2295$, Figure 3B) and produced reasonable cell type abundance estimates (Figure 3C).

We trained MSEPM models using data assembled from four GEO series[20, 22, 29, 36] ($n = 1605$) with samples spanning a wide age range (0.01 - 94.0 years). The samples were randomly split into training ($n = 1203$) and validation ($n = 402$) sets stratified by age. 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. The first cell type principal component (CT-PC1) is largely driven by the relative abundance of granulocytes (Figure 3D), while the second PC (CT-PC2) captures relative differences in the abundance of differentiated lymphocytes (Figure 3D). 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.

Methylation sites were selected for modeling with MSEPM if the sites were highly correlated with age ($n = 276$) , sex ($n = 49$), CT-PC1 ($n = 120$), CT-PC2 ($n = 116$) or a combination of factors ($n = 238$), a total of 778 unique sites. Four MSEPM models were fit using the selected sites with four combinations of input factors (Age, Age Sex, Age Sex CT-PC1, and Age Sex CT-PC1 CT-PC2). Training sample factors were min-max scaled between 0 and 1 before model training. The association between the fit epigenetic factor predictions against the input modeled factors was assessed by fitting a trendline between epigenetic state predictions and scaled continuous input factors using the state prediction made for the MSEPM model trained with all four input factors. Performance of the MSEPM model was then evaluated using the testing samples ($n = 4,082$). The performance of the MSEPM largely closely resembles the simulation results. All four MSEPM models capture the nonlinear relationship between age and methylation status (Supp. Figure 2). The epigenetic state prediction associated with age improves as the underlying methylation data are more fully explained through the addition of epigenetic factors (Supp. Figure 2). The MSEPM model fit with Age, Sex, CT-PC1 and CT-PC2 can accurately model the associated epigenetic state for each factor (Figure 4 A-D) and accurately predicts the methylation levels at individual sites ($R^2 = 0.935$, $MAE = 0.035$, Figure 4 E). The trained MSEPM produces a collection of methylation site models that can help explain the association between modeled factors and methylation status.

2.4 Analysis of chromatin regulators of site clusters

We evaluated the relationship between sites that are influenced by age, sex, CT-PC1 or CT-PC2 and potential regulatory factors by performing overlap enrichment analysis of these sites with transcription factor chromatin immunoprecipitation peaks present in the ENCODE V4 [38, 39] release. We first identified sites with similar coefficients of epigenetic factors through hierarchical clustering. Briefly, sites were clustered using Ward's method by the euclidean distance between site regression coefficients that were normalized by the standard deviation of the methylation values amongst the training samples. The resulting tree was cut at a height of 18 to produce 10 distinct clusters with clear associations to the modeled factors (Figure 5A).

The site clusters largely conform to underlying biological expectations. Cluster one contains sites that are wholly associated with sex status and localized to the X chromosome (Supp. Table 1) and is enriched for peaks of transcription factors associated with sex specific regulation such as MAZ [40]. Clusters nine and ten contain sites whose methylation status is largely driven by CT-PC1, and are enriched for transcription factors associated with granulocyte development (CEBPB, CEBPA, EP300, ETV6)[41] [42]. Similarly, clusters two, five and eight are associated with CT-PC2 and are enriched for transcription factor peaks associated with immune development (ZBED1, ETV6, FOSL2, FOS, TBX21). Clusters four and six are associated with loss of methylation with age. Cluster six is highly enriched for CTCF binding sites; CTCF is known to increase at sites where methylation is lost during aging [43]. Cluster four is enriched for STAT3 whose activation during exercise is age dependent [44, 45]. Cluster seven is associated with the accumulation of methylation with age and is enriched for immunoprecipitation peaks for aging associated transcription factors SMAD4 and RE1-Silencing Transcription Factor (REST). SMAD4 encodes a protein involved in the transforming growth factor beta (TGF-β) signaling pathway. Age related dysregulation of TGF-β has been linked to reduced skeletal muscle regeneration[46, 47] and SMAD4 polymorphisms are associated with longevity [48]. REST is a transcriptional repressor of neuron specific genes in non-neuronal cells [49, 50]. REST expression is upregulated in aged prefrontal cortex tissue and the absence of REST expression is associated with cognitive impairment [51] and cellular senescence in neurons [52].

3 Discussion

Epigenetic clocks are widely used tools to study human aging and health. Despite their widespread use, the biological interpretation of model parameters is limited. A methylome is influenced by many different biological processes occurring simultaneously over time that may differ among individuals. Epigenetic clocks, while producing accurate predictions of age, attempt to capture this complexity through a single dependent variable. Additionally, the penalized regression based methods used to fit most epigenetic clocks select sites that minimize, or regress out, the influence of other factors and omit groups of sites that are correlated. To overcome these limitations, here we propose a multidimensional extension of the EPM model, the MSEPM.

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

Additionally, the individual methylation site linear models fit as part of the MSEPM optimization can provide information about the relationship between modeled factors and site specific biology. To this end, we find that the blood MSEPM model conforms to expected biology. Sites with a strong sex association localize to the X chromosome and sites associated with cell types are enriched for transcription factors associated with the development of immune cells. Sites that are primarily affected only by age in the blood MSEPM model are of particular interest. As others have previously described, sites that progressively lose methylation over time are strongly enriched for CTCF [53, 54], while in the blood MSEPM model sites that gain methylation over time are enriched for REST. The association between age and REST is well established, but the association between gain of methylation and age has previously not been reported. Both transcription factors are associated with the maintenance of transcriptional profiles and suggest that there is a general dysregulation of maintenance of cell specific chromosomal organization over time.

In conclusion, we introduced a multi-dimensional extension of the Epigenetic Pacemaker, the MSEPM. The MSEPM is capable of accurately modeling multiple methylation associated factors simultaneously. This paradigm can elucidate the site specific regulation underpinning methylome dynamics. The MSEPM is available under the MIT license at https://github.com/NuttyLogic/MultistateEpigeneticPacemaker.

4 Methods

4.1 MSEPM

MSEPM model training begins with an input matrix of i methylation sites for j individuals and k input traits for j individuals. The goal of the MSEPM is to minimize the difference between the observed and predicted methylation values at any of the i input methylation sites as explained by the k input traits. The MSEPM model is fit as follows:

- Fit i ordinary least squares regression models with the form $\hat{m_{i,j}} = r_i^0 + \sum_{k=1}^{n} s_{j,k} r_{i,k} + \epsilon_{i,j}$ to estimate site specific parameters r_i shared amongst all individuals.
  - The k untransformed input traits are used as the initial guess of $S_{j,k}$ for the first model iteration.
- Update $S_{j,k}$ to minimize the RSS cost function, $J(k) = \frac{1}{n}\sum_{i=1}^{n}(\hat{m_{ij}} - (r_i^0 + R_i S_j))^2)$, using gradient descent with fixed r_i site parameters.
  - $\nabla J(k)$
    - $\frac{d}{dk} = \frac{2}{n}\sum_{i=1}^{n}(\hat{m_{ij}} - (r_i^0 + R_i S_j))r_{i,k}$
    - $s_{j,k}^1 = s_{j,k}^0 - \eta * \frac{d}{dk}$, $\eta =$ learning rate
- Repeat steps 1-2 until the reduction in sum of site RSS between iterations is $\leq$ specified threshold or a set number of iterations has been reached.

4.2 Simulation Framework

Under the MSEPM formulation a single site can be described linearly where the observed methylation value is dependent on a weighted sum of n epigenetic states, $\hat{m_{i,j}} = r_i^0 + \sum_{n=1}^{n} p_{j,n} r_{i,n} + \epsilon_{i,j}$, where $p_{j,n}$ is an individual trait associated with the methylation status at site i. Using this formulation we implemented a simulation framework in python with only numpy [55] as a dependency. The details of the approach are as follows:

- Individuals are simulated by setting individual traits with the following form $p_{n,j} = Age_j^{\gamma_n} q_{n,j}$ Where the simulated individual trait $p_{n,j}$ is dependent time and exposure, $q_{n,j}$, over time. Under this formulation traits may have a nonlinear association with the simulated methylation status as mediated by time. The form of $Q[q]$ varies between simulated traits as shown above. Simulated MSEPM models are fit using $q_{n,j}$ as input for and the association between $p_{n,j}$ is learned through model fitting.$p_{n,j} = Age_j^{\gamma_n} q_{n,j}$ Where the simulated individual trait $p_{n,j}$ is dependent time and exposure, $q_{n,j}$, over time. Under this formulation traits may have a nonlinear association with the simulated methylation status as mediated by time. The form of $Q[q]$ varies between simulated traits as shown above. Simulated MSEPM models are fit using $q_{n,j}$ as input for and the association between $p_{n,j}$ is learned through model fitting.

- Methylation site profiles are then simulated by setting the dynamic range of possible methylation values and the error profile of the methylation site. The simulated methylation values are bounded by 0 and 1, and represent the proportion of methylated bases to the total number of observed cytosines at any site. The dynamic range is set by first initial methylation value $m_0 \lfloor\rceil\sqcup\dashv(.2, .2)$ and then setting a target methylation value m_t where the dynamic range is $\delta_i = m_t - m_o$. The value of m_t is set conditionally to ensure dynamic range is always larger than some specified threshold ,θ, where $\theta \leq |\delta| \geq$. $m_0 \lfloor\rceil\sqcup\dashv(.2, .2)$ and then setting a target methylation value m_t where the dynamic range is $\delta_i = m_t - m_o$. The value of m_t is set conditionally to ensure dynamic range is always larger than some specified threshold , θ, where $\theta \leq |\delta| \geq$.
- Methylation sites profiles are then randomly selected and associated with a weighted set of simulated traits. The weighted traits are then normalized so the input combination of traits describes the dynamic range of the simulation site, δ. If no weights are provided the individual methylation sites are simulated by returning $m_0 + N(\mu, \sigma)$. δ. If no weights are provided the individual methylation sites are simulated by returning $m_0 + N(\mu, \sigma)$.

4.3 Blood MSEPM Model Training

Methylation sites with an absolute pearson correlation coefficient greater than 0.7, 0.995, 0.92 and 0.64 between age ($n = 276$), sex ($n = 49$), cell type PC1 ($n = 120$) and cell type PC2 ($n = 238$) and site methylation values respectively among training samples samples($n = 1203$) were included in the model. Additionally, sites with a sum of all absolute pearson coefficients 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, note 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.

4.4 Blood MSEPM Model Cluster Transcription Factor Overlap Analysis

A custom transcription factor reference set was built from all 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 to GRCh37 coordinates using liftOver [56]. The overlap reference contains 714 transcription factor targets 1621 accession IDs (table . . .).

Blood MSEPM model site hierarchical clustering was performed as follows. Individual methylation site coefficients were first normalized by the standard deviation of methylation values of the site among the training samples, $r_{i,n}/\sigma_i$. A distance matrix was then created by taking the Euclidean distance between the normalized site model coefficients. Sites were then clustered using Ward's method which seeks to minimize within cluster variance by minimizing the increase in the error sum of squares (ESS) through successive cluster fusions. Clusters label by tree cutting at a height of 18. All clustering analysis was carried out using SciPy v1.6.3 [57]. $r_{i,n}/\sigma_i$. A distance matrix was then created by taking the Euclidean distance between the normalized site model coefficients. Sites were then clustered using Ward's method which seeks to minimize within cluster variance by minimizing the increase in the error sum of squares (ESS) through successive cluster fusions. Clusters label by tree cutting at a height of 18. All clustering analysis was carried out using SciPy v1.6.3 (Virtanen et al. 2020).

Transcription factor enrichment analysis was performed with LOLA [58] which assesses the genomic region set overlap between a set of query regions and a set of reference regions, within a specified shared background set, using Fisher's exact test. Overlap analysis was performed for sites within a cluster against the ENCODE V4 reference region (1BP minimum overlap) using all sites assayed with Infinium HumanMethylation450K BeadChip as background.

4.5 Analysis Environment

Analysis was carried out in a Jupyter[59] analysis environment. Joblib[60], SciPy[61], Matplotlib[62], Seaborn[63], Pandas[64] and TQDM[65] packages were utilized during analysis.

4.6 Supplementary Information

All analysis code, data processing code, and supplementary material associated with this manuscript can be found at https://github.com/NuttyLogic/MSEPMManuscript. The methylation simulation utility can be found at https://github.com/NuttyLogic/MethSim. The data supporting these findings are openly available at GEO under the series GSE87640, GSE87648, GSE51057, GSE51032, GSE87571, GSE125105, GSE42861, GSE69138, GSE111629, GSE128235, GSE121633, GSE73103, GSE61496, GSE59065, GSE97362, GSE156994, GSE128064 and GSE43976.

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. Ventham, 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. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. en. *Nucleic Acids Res.* **41,** D991–D995 (Nov. 2012).

34. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. en. *BMC Bioinformatics* **13,** 86 (May 2012).

35. 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).

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. 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).

41. 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).

42. 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).

43. 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).

44. Trenerry, 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).

45. 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).

46. 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).

47. Carlson, M. E. *et al.* Relative roles of TGF-Î²1 and Wnt in the systemic regulation and aging of satellite cell responses. en. *Aging Cell* **8,** 676–689 (Dec. 2009).

48. 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).

49. 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).

50. Coulson, J. M. Transcriptional regulation: cancer, neurons and the REST. en. *Curr. Biol.* **15,** R665–8 (Sept. 2005).

51. Lu, T. *et al. REST and stress resistance in ageing and Alzheimer's disease* 2014.

52. Rocchi, A. *et al.* REST/NRSF deficiency impairs autophagy and leads to cellular senescence in neurons. en. *Aging Cell* **20,** e13471 (Oct. 2021).

53. Han, Y. *et al.* New targeted approaches for epigenetic age predictions. en. *BMC Biol.* **18,** 71 (June 2020).

54. 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).

55. Harris, C. R. *et al.* Array programming with NumPy. en. *Nature* **585,** 357–362 (Sept. 2020).

56. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. en. *Nucleic Acids Res.* **34,** D590–8 (Jan. 2006).

57. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* (Feb. 2020).

58. 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).

59. Basu, A. *Reproducible research with jupyter notebooks*

60. Varoquaux, G. & Grisel, O. Joblib: running python function as pipeline jobs. *packages. python. org/joblib* (2009).

61. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* (Feb. 2020).

62. Hunter, J. D. *Matplotlib: A 2D Graphics Environment* 2007.

63. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6,** 3021 (Apr. 2021).

64. McKinney, W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* en ("O'Reilly Media, Inc.", Oct. 2012).

65. Da Costa-Luis, C. O. tqdm: A Fast, Extensible Progress Meter for Python and CLI. *JOSS* **4,** 1277 (May 2019).

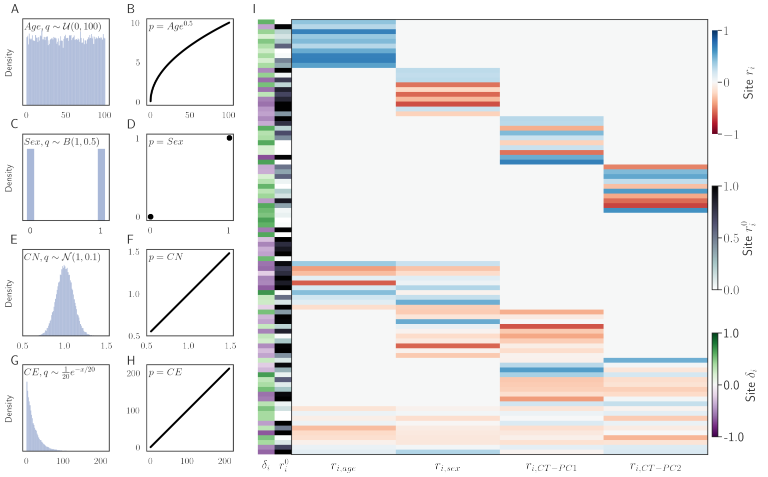

Figure 1: Simulated factors and the association with simulated methylation values. (A) Age with a non-linear association with methylation (B). Sex (C) with a binary association with methylation (D). Normal factor (E) with a linear relationship with methylation (F). Continuous exponential trait (G) with 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.

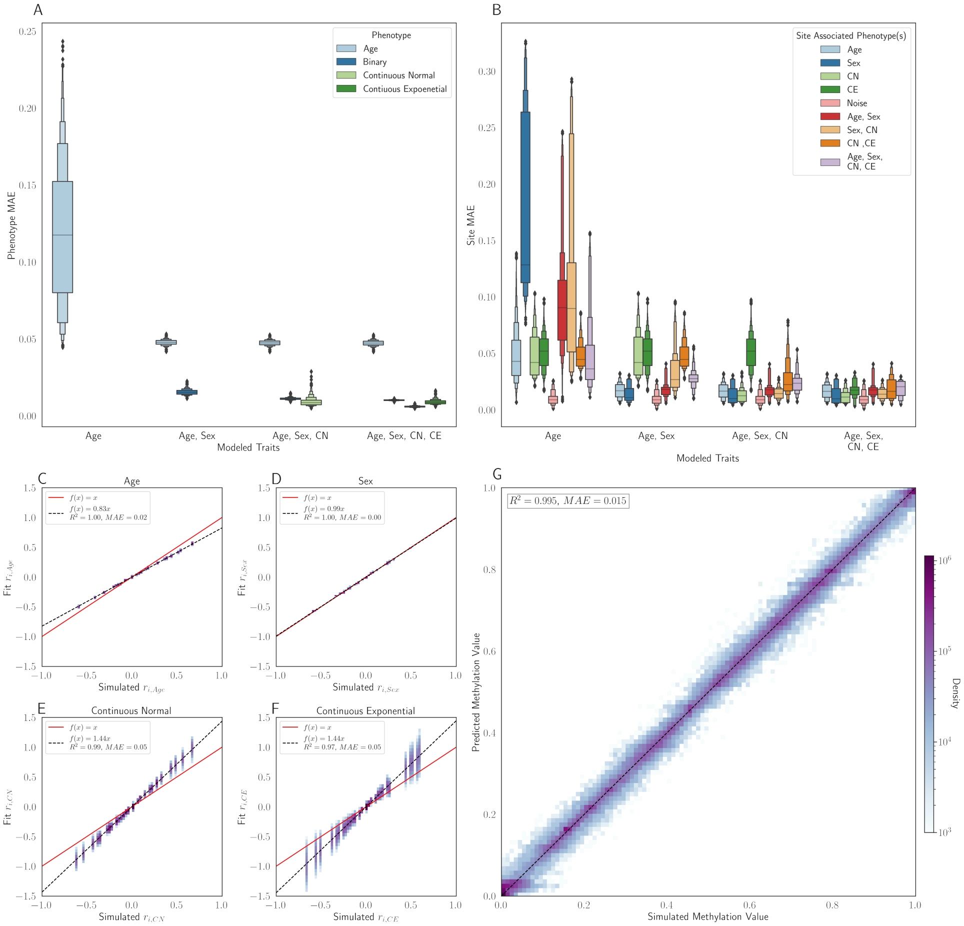

Figure 2: (A) The MAE of the factor predictions on the testing set as multiple factors are modeled simultaneously. (B) Prediction methylation MAE as factors are included in the MSEPM model. (C) Model coefficients for Age, Sex, Continuous Normal and Continuous Exponential factors for models trained ($n = 500$) with all four simulated factors. (D). Simulated and predicted methylation values for all simulated testing sites across all training folds.

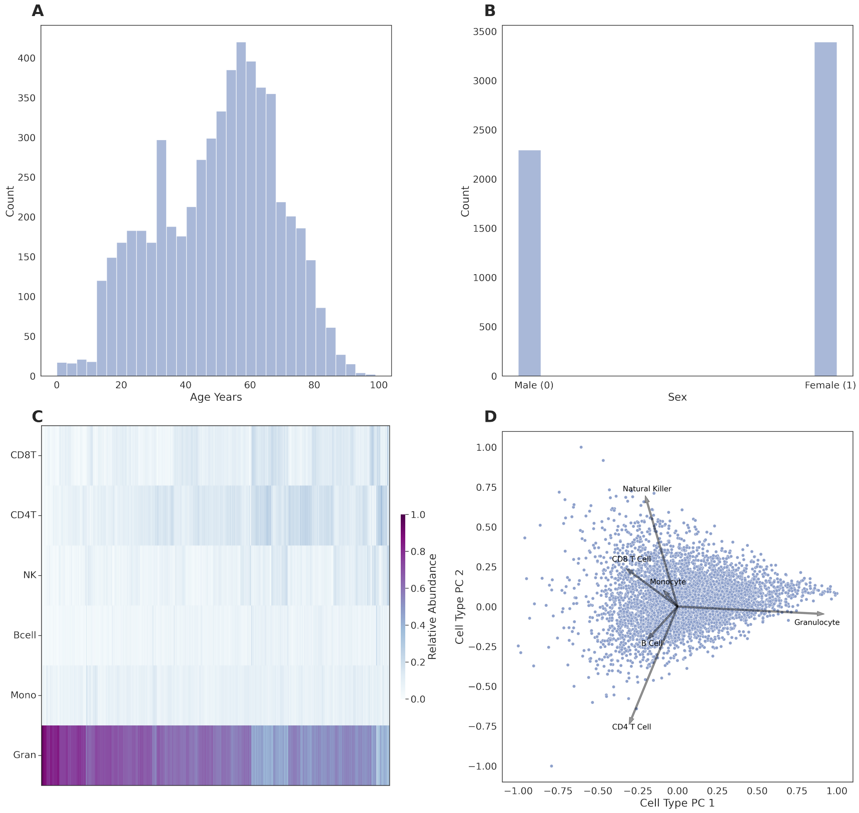

Figure 3: Distribution of age (A) and (B) sex in aggregate blood dataset. (C) Calculated cell type composition and (D) loading plot of principal components of cell type composition in aggregate blood data set. .

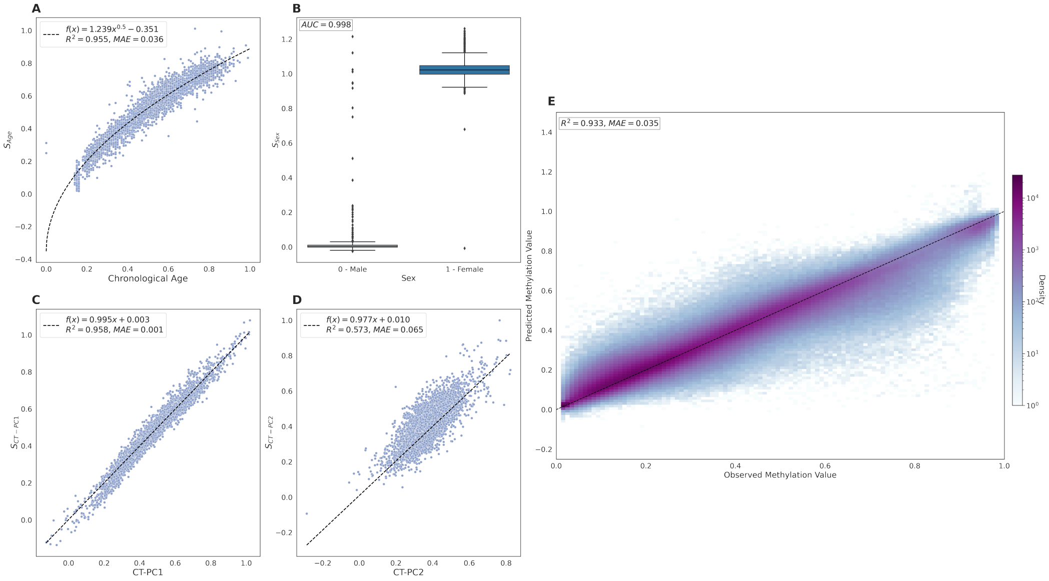

Figure 4: MSEPM model trained with age, sex, CT-PC1 and CT-PC2 predictions within testing set for methylation associated factors (A) age, (B) sex, (C) CT-PC1 and (D) CT-PC2. (E) Observed and predicted methylation values for training set has high concordance ($R^2 = 0.833, MAE = 0.035$)

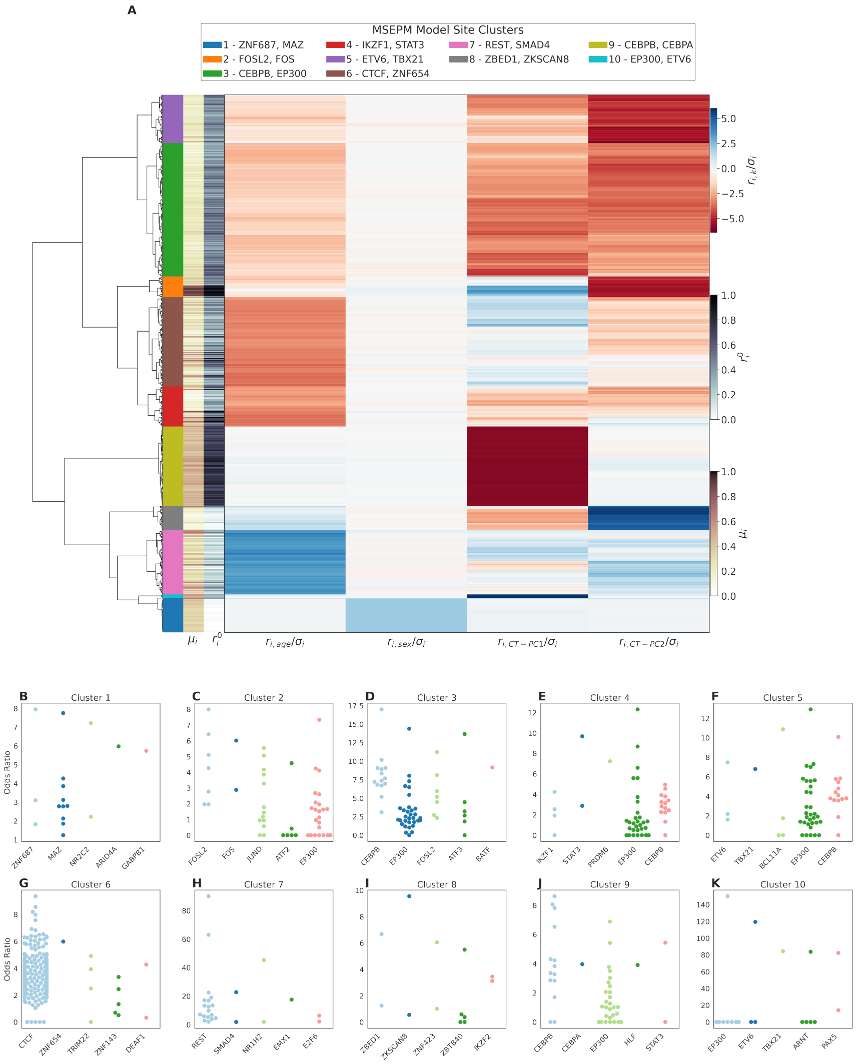

Figure5: (A) Site clustering by standardized model coefficients. Sites clusters show distinct relationships with modeled traits. (B-K) Top five enriched transcription factors for clusters 1 - 10.