DNA methylation biomarker in blood predicts frailty in an HIV-positive veteran population

Chang Shu[1,2], Amy C. Justice[1,2], Xinyu Zhang[1,2], Vincent C. Marconi[3], Dana B. Hancock[4], Eric O. Johnson[4,5], Ke Xu[1,2]

1. Department of Psychiatry, Yale School of Medicine, New Haven, CT, 06516

2. Connecticut Veteran Healthcare System, West Haven, CT, 06515

3. Division of Infectious Disease, Emory University School of Medicine, Atlanta GA

4. Center for Omics Discovery and Epidemiology, Behavioral Health Research Division, RTI International, Research Triangle Park, NC, USA

5. Fellow Program, RTI International, Research Triangle Park, NC, USA.

All corresponding address to

Ke Xu, M.D., Ph.D.

Associate Professor of Psychiatry

Yale School of Medicine

Email: ke.xu@yale.edu

Tel: 203-932-5711x7430

Abstract

Background: With the improved life expectancy in people living with HIV, predicting individual frailty is important for clinical care. DNA methylation (DNAm) has emerged as a robust biomarker in precision medicine and has been previously linked to aging and mortality in non-HIV populations. Here, we aim to establish a panel of DNAm features selected from blood methylome to predict frailty in HIV-positive men from a veteran population, Veterans Aging Cohort Study (VACS).

Methods: We used a well-established score, VACS Index, as a measure of frailty. Samples (n_{total}=1,150) were divided into a training set (n=612) and a validation set (n=538). We first selected a panel of frailty-associated CpGs by conducting an epigenome-wide association analysis on the VACS index in the training set. We then applied four machine learning methods to build models to predict high and low frailty individuals in the training set. The prediction models were tested in the validation set. A methylation score constructed from the selected CpGs was tested an associated with mortality by performing survival analysis. To assess the biological relevance of the selected CpG sites, we performed a gene ontology enrichment analysis.

Results: A panel of 119 CpGs were identified from the training set (False Discovery Rate <0.2) that showed excellent performance on predicting high frailty individuals with Area Under Curve (AUC) of 0.835 (95%CI: 0.792, 0.879) and balanced accuracy of 0.693. The same panel showed good performance on predicting low frailty individuals with AUC of 0.735 (95%CI: 0.688-0.782) and a balanced accuracy of 0.528. The methylation score from the 119 CpGs was significantly associated with 5-year and 10-year mortality with hazard ratio of 1.40 (95% CI:1.033, 1.897 p=0.03) and 1.48 (95%CI:

1.10, 2.02; p=0.01) respectively. These 119 CpGs were located within or near 73 genes that were significantly enriched in 9 biological pathways relevant to immune and inflammation response.

Conclusions: We identified a panel of predictive DNAm features associated with frailty and mortality among HIV-positive population. These DNAm features may serve as biomarkers to discriminate high and low frailty people who live with HIV.

Keywords: DNA methylation, HIV-positive, frailty, mortality, DNA methylation score, machine learning prediction

Background

Combination antiretroviral therapy has significantly improved life expectancy among HIV-positive individuals. The prolonged life expectancy increases frailty risk in the HIV population. The prevalence of frailty in people living with HIV is significantly higher and onset of frailty occurs at an earlier age compared to the general population (1). High frailty is characterized by marked vulnerability and is associated with increased mortality. Thus, prediction of frailty is important to identify vulnerable group and deliver clinical care to highly vulnerable HIV-positive patients. At present, frailty is usually defined by clinical symptoms (e.g., Fried's frailty phenotype, Rockwood and Mitnitski's Frailty Index) or a combination of lab tests that indicate organ damage such as the Veterans Aging Cohort Study index (VACS index) (2). No biomarkers available to capture the early stage of the frailty to predict individual vulnerability in HIV-positive patients.

A large body of evidence has demonstrated that epigenetic modification adapts internal and external environmental changes and that is associated with early stage of pathophysiological processes (3-6). DNAm, one type of the most widely studied epigenetic marks, is strongly related with aging (7-9), substance use (e.g. cigarette smoking and alcohol consumption) (10-16), and a variety of diseases (3-6, 17, 18). Because DNAm is relatively stable and easy to detect from body fluids biospecimen through non-invasive procedure, DNAm marks have emerged as robust biomarkers for cancer diagnosis (19), disease subtype classification (20, 21) and treatment response monitoring (22, 23).

DNAm may play an important role in frailty among HIV-positive individuals. Frailty is associated with elevated inflammation markers such as IL-6 and hsCRP in HIV-positive individuals (24). Genes involved in immune and inflammation processes are also subject to epigenetic modification in immune cells. We previously reported association of two CpG sites in the promoter region of *NLRC5* with HIV infection (25), and *NLRC5* gene is a major transcriptional activator of MHC class I gene. DNAm was also linked to HIV comorbid medical diseases such as HIV-positive diabetes and kidney function (26, 27). Furthermore, aging, a critical contributor to frailty, is significantly associated with thousands of CpGs in the epigenome, and the epigenetic clock and DNAm age are becoming widely recognized (7-9). DNAm age is significantly correlated with physical frailty in HIV-negative individuals (28) and HIV-positive individuals (29). In HIV-positive individuals, the average DNAm age is accelerated 5 to 10 years faster than HIV-negative individuals (30-33) and 10 years faster in the HIV-positive frailty individuals in comparing to HIV-positive non-frailty individuals (29). These age-related DNAm signatures are predictive of mortality (34-38). Additional CpGs in blood have been identified to predict all-cause mortality (34, 39). Therefore, we hypothesize that DNAm is associated with frailty and that DNAm signatures in blood can serve as biomarkers to predict frailty and mortality among HIV-positive individuals.

In this study, our goal is to identify DNAm marks that can serve as biomarkers on frailty and to link frailty-associated DNAm to mortality among HIV-positive individuals. We also evaluated biological relevance of the identified CpG sites. We used a well-established frailty score, VACS index, as a measure of frailty in the HIV-positive population. The VACS index is a composite score constructed by a sum of pre-assigned

points on age, CD4 count, HIV-1 RNA, hemoglobin, platelets, aspartate and alanine transaminase (AST and ALT), creatine, estimated glomerular filtration rate (eGFR), and viral hepatitis C infection (40). Findings from this study provide a set of DNAm biomarkers to predict frailty for future clinic use and new insights into the epigenetic mechanism of frailty and mortality in HIV pathology.

Methods

As an overview, our analytical procedure is shown in the flowchart in **Figure 1**. Our sample was divided into the training set (n=612) and the validation set (n=538) which DNA methylation were processed at different time and using two different platforms. We first selected a panel of CpG sites relevant to frailty based on EWAS results in the training set. Then, we applied four commonly used machine learning classification methods to develop prediction models: random forest, Extreme Gradient Boosting Tree (XGBoost), Lasso and Elastic-Net Regularized Generalized Linear Models (GLMNET), and Support Vector Machines (SVM). The selected CpG sites were used as predictors to differentiate high and low frailty in the training set and each model was evaluated in the validation set. Additionally, the DNAm score was constructed based on the selected CpG sites and we conducted survival analysis to assess whether the DNAm score was associated with mortality in the HIV-positive samples. Lastly, we conducted a gene ontology enrichment analysis to reveal the underlying biological pathways of the selected CpG sites.

Study population

All samples in the training and validating sets were from the VACS cohort. The VACS is a prospective cohort study of veterans focusing on clinical outcomes in HIV infection (41). DNA samples were extracted from peripheral blood of 1,150 HIV-positive men from the VACS. Demographic and clinical information on age, race, smoking status, CD4 count, viral load, HIV medication adherence, VACS index, and mortality by training set and validation set are presented in **Table 1**. The training set included slightly older individuals, more African Americans, and greater VACS index than the validation set. There was no significant difference in HIV medication adherence, CD4 count, or $\log_{10}$ HIV-1 viral load between the training set and the validation set.

Table 1: Study sample characteristics

	Training set (N=612)	Validation set (N=538)	p value*
Age (year, mean +/-sd)	49.4 (7.6)	48.1 (7.8)	0.005
Male (%)	612 (100)	538 (100)	
Race (%)			
Caucasian	58 (9.5)	48 (8.9)	0.001
African Americans	526 (85.9)	435 (80.9)	
Other	28 (4.6)	55 (10.2)	
Smokers (%)	360 (59.4)	309 (58.4)	0.78
HIV positive (%)	614 (100)	538 (100.0)	
HIV treatment adherence (%)	469 (78.4)	407 (76.6)	0.519
CD4 count	428.97 (286.33)	447.46 (279.27)	0.287
log 10 HIV-1 viral load	2.61 (1.19)	2.67 (1.23)	0.423
VACS index (mean +/-)	35.64 (21.99)	30.78 (20.23)	<0.001
Mortality (%)	162 (26.5)	129 (24.0)	0.367

***t test is used for continous variables, chi-square test is used for categorical variables**

Genome-wide DNAm profiling and quality control

The DNA samples in the training set were profiled by Infinium Human Methylation 450K BeadChip (HM450K) and the DNA samples in the validation set were profiled by the

Infinium Human Methylation EPIC BeadChip. DNAm for the training and validation sets were evaluated using the same quality control (QC) protocol (42) in the R package minfi (43). In detail, CpG sites on sex chromosomes and within 10 base pairs of a single nucleotide polymorphism were removed. The detection p-value threshold was set at 10^{-12} for both the training and validation sets. After QC, 408,583 shared CpG sites between HM450K and EPIC arrays were used for analysis to ensure the same coverage between two sets. Proportions of 6 cell types (CD4+ T cells, CD8+ T cells, Natural Killer T cells, B cells, monocytes and granulocytes) were estimated using the established method (Houseman et al., 2012) for all samples in the training and validation sets.

Selection of CpG sites from EWAS on frailty in the training set

We performed an EWAS on VACS index in the training set using a two-step linear model approach as previously described (42). The analytical model was adjusted for systematic errors and clinical confounding factors. Here, log transformation was applied to VACS index to ensure normality distribution assumption for the linear model.

First, the following linear model was used to extract principal components (PCs) for potential confounding variables. Here, top 30 control PCs were derived from internal control probes in *minfi*:

Methylation β ~ age + race + smoking + self-reported HIV medication adherence + log viral load + WBC + CD4T + CD8T + Gran + NK + B cell + Mono + 30 Control PCs

Second, the top 20 residual PCs were extracted in the model above to fully adjust for unmeasured confounding. The final EWAS model on VACS index was:

log(VACS) ~ Methylation β + age + race + smoking + self-reported HIV medication adherence + log viral load + WBC + CD4T + CD8T + Gran + NK + B cell + Mono + 30 Control PCs + 20 Residual PCs

To ensure a sufficient number of CpGs as predictors for the prediction model development, we selected CpG sites with false discovery rate (FDR) <0.2 in the EWAS as predictors.

Additionally, to explore the underlying epigenetic mechanism of frailty, we also detected differentially methylated regions (DMRs) using *bumpHunter* (44) in the training set. DMRs with family-wise error rate (FWER)<0.2, which is equivalent to FDR<0.03, were considered significant.

Developing machine learning prediction models for frailty

Based on the distribution of the VACS index in the entire sample (**Figure S1**) and clinical significance (45), high HIV frailty was defined as VACS index > 50, and low HIV frailty was defined as VACS index < 16. Models were developed to predict high frailty (VACS index >50 versus ⩽50) and low frailty (VACS index <16 versus ≥16). CpGs relevant to frailty were selected from EWAS as described above. The following steps were taken to develop separate prediction models on high and low frailty:

1) Model development in the training set: Four machine learning models, random forest, Extreme Gradient Boosting Tree (XGBoost), support vector machines(SVM), Lasso and Elastic-Net Regularized Generalized Linear Models (GLMNET), were separately applied to predict VACS index by using the R package *caret* (46). These four models are commonly used in the supervised learning on classification (47-50). Each

model was developed separately and 10-fold cross-validation was used in model training process to minimize data overfitting.

2) Model evaluation in the validation set: All four models built in the training set were evaluated in the validation set. Area Under Curve (AUC) and balanced accuracy were used to assess the prediction performance. We used balanced accuracy for model evaluation because in the presence of imbalance samples in each class since there are 228 subjects with VACS index greater than or equal to 50 and 916 subjects with VACS index less than 916. Balanced accuracy is defined as the average accuracy obtained on each class as shown in the following formula (51). Balanced accuracy was used in this study to avoid biased accuracy due to imbalanced samples (51).

$$Balanced\ accuracy = \frac{1}{2}\left(\frac{True\ positive}{True\ positive + False\ negative} + \frac{True\ negative}{True\ negative + False\ positive}\right)$$

Association of a DNAm score for VACS index with mortality

We constructed a DNAm score by the selected frailty-associated CpG sites. A survival analysis was conducted to assess whether the DNAm score was associated with 5-year and 10-year mortality respectively.

DNAm score was constructed based on a previously reported method (12). We computed the DNAm score by normalizing the score with mean β value (μ_c) and standard deviation, σ_c, from subjects with VACS index <50 (control group). For selected n methylation sites, W_c is 1 for a hypermethylated methylation site c, and -1 for a

198 hypomethylated site c, β_{ci} is the β value for subject i and methylation site c. The
199 methylation score for subject i was constructed by the following formula:

200 $$Methylation\ Score_i = \frac{1}{n}\sum_{c=1}^{n} W_c \frac{\beta_{ci} - \mu_c}{\sigma_c}$$

201 The mean methylation score was 0, and samples were divided into high
202 methylation score (>0) and low methylation score (⩽0) groups.
203 We preformed Kaplan-Meier survival curves during 10-year follow-up to visualize
204 the survival differences between high and low methylation score groups. Survival
205 analysis was conducted by cox proportional hazards model on 5-year and 10-year
206 mortality comparing high and low methylation groups. We used age as time scale *t*, and
207 our model was adjusted for race, smoking, self-reported HIV medication adherence, log
208 10 of HIV viral load and CD4 count.
209
210 $$h(t) = h_0(t)\exp\left(\beta_1\, methylation\ score\ +\ \beta_2\, race + \beta_3\, smoking\right.$$
211 $$\left. +\ \beta_4\, HIV\ medication\ adherence\ +\ \beta_5 \log_{10}(viral\ load) + \beta_6\, CD4\ count\right)$$
212
213 ***Biological interpretation of the predictive panel of CpG sites on frailty***
214 We performed gene ontology (GO) enrichment analysis using Database for
215 Annotation, Visualization and Integrated Discovery (DAVID) (52). To avoid redundancy
216 in pathway names, we only used level 4 GO terms defined in DAVID in the enrichment
217 analysis. Genes with at least one frailty-associated CpG site was used for GO analysis.
218 We considered biological pathways with FDR <0.05 as statically significant pathways.
219

Results

Selection a panel of 119 frailty-associated CpG sites by EWAS in the training set

In the training set, we conducted an EWAS on VACS index (**Figure 2**; λ=1.038) and selected 119 CpGs with FDR<0.2 to ensure sufficient number of predictors (**Table 2)**. Eighty out of 119 CpGs were negatively associated with VACS index, while 39 out of 119 CpGs were positively associated with VACS index. The majority of CpGs were located within or near known genes, except for 19 CpGs that were intergenic. Only 40 CpGs were located in gene bodies, while 79 CpGs were located in promoter regions, first exons, or 3' UTR. Thirty-two CpGs were located in CpG islands.

Notably, 22 out of 119 CpGs reached epigenome-wide significance (FDR<0.05), including 16 CpG sites negatively associated with the VACS index and 6 CpG sites with positive associations. These 22 CpGs were located on 17 genes, including 13 of 22 CpGs located in promoter regions, 6 CpGs in gene body, and 3 CpGs in 3'UTR. Among these genes, some harbored more than one significant CpG. For example, three CpGs were annotated within or near *PSMB8* (cg01309328, $p=4.44\times10^{-9}$; cg08099136, $p=4.63\times10^{-9}$; cg00533183, $p=3.18\times10^{-7}$), two CpGs were located near *PARP9* (cg08122652, $p=2.18\times10^{-7}$; cg22930808, $p=6.78\times10^{-7}$), two CpG sites were located near *IFITM1* (cg03038262, $p=5.83\times10^{-7}$; cg23570810, $p=1.58\times10^{-6}$), and the rest of the significant CpG sites were annotated to 13 genes including *MX1*, *TAP1*, *ZNF32*, *NLRC5*, *IFI44L*.

We found 9 significant DMRs in the training set (**Table S1**). Methylation β values between high and low frailty ranged from -0.15 to 0.05 in these regions. As examples, 3

region plots showing DMRs in the *MX1, PARP9, and IFI44L* genes are shown in **Figure 3**.

Machine learning prediction on frailty by the selected 119 CpGs

Table 3: Performance of machine learning model predicting frailty in an HIV-infected population

	High frailty (VACS index >50 vs. ≤50)				Low frailty (VACS index >16 vs. ≤16)			
	Training		Validation		Training		Validation	
Method	AUC	Balanced Accuracy	AUC	Balanced Accuracy	AUC	Balanced Accuracy	AUC	Balanced Accuracy
Random Forest	0.775	0.622	0.835 (0.792,0.879)	0.693	0.774	0.509	0.735 (0.688,0.782)	0.528
XGBoost	0.760	0.624	0.830 (0.786,0.874)	0.684	0.820	0.637	0.716 (0.669,0.762)	0.500
GLMNET	0.783	0.642	0.757 (0.701,0.813)	0.668	0.795	0.589	0.715 (0.667,0.763)	0.503
SVM	0.759	0.618	0.744 (0.685,0.802)	0.500	0.756	0.594	0.501 (0.442,0.560)	0.492

AUC: Area Under Curve
VACS: Veteran Aging Cohort Study
XGBoost: Extreme Gradient Boosting Tree
GLMNET: Lasso and Elastic-Net Regularized Generalized Linear Models
SVM: Support Vector Machines

In the training set, we developed our prediction models based on the selected 119 CpGs in the training set using four commonly used machine learning classification models: XGBoost, random forest, SVM and GLMNET (47-50). Performances of four prediction models are presented in **Table 3**. We found that AUCs ranged from 0.756 to 0.820 and balanced accuracies ranged from 0.509 to 0.642, suggesting that the 119 CpGs had good to excellent predictive performance on frailty. The performances of four models were mostly comparable. For predicting high frailty, GLMNET outperformed the three other models (AUC=0.783, balanced accuracy=0.642) while SVM performed slightly worse (AUC=0.759, balanced accuracy = 0.618). For the prediction of low frailty,

XGBoost performed the best among the four models (AUC=0.820, balanced accuracy=0.637) while SVM performed worst (AUC=0.756, balanced accuracy =0.594).

In the validation set, for the prediction of high frailty, random forest and XGBoost showed the best performance. AUC of random forest model was 0.835 (95%CI: 0.792, 0.879) and XGBoost was 0.830 (95%CI: 0.786, 0.874) for predicting high frailty. Balanced accuracy was 0.69 and 0.68, respectively, for random forest and XGBoost (**Figure 4**). For the prediction of low frailty, XGBoost and random forest models also outperformed GLMNET and SVM. AUC was 0.735 (95%CI: 0.688, 0.782) for the random forest and 0.716 (95%CI: 0.669, 0762) for the XGBoost. The balanced accuracy was 0.528 and 0.500 for the random forest and XGBoost, respectively.

These results suggest that the selected panel of 119 CpG sites were able to predict frailty, and the prediction models performed better at predicting high frailty than predicting low frailty. Among four machine learning methods, SVM showed poor performance of predicting frailty in our samples.

DNAm score by 119 CpGs was significant associated with mortality

We constructed DNAm score based on the selected 119 CpGs, and we further assessed whether VACS index-related DNAm score were associated with mortality. In **Figure 5**, the Kaplan Meier curves during 10-year follow-up showed that the individuals with high DNAm score were at higher risk of mortality than those with low DNAm scores. After adjusting for confounding factors, our cox proportional hazards regression model showed that the hazard ratio of 5-year and 10-year mortality comparing high and low DNAm score groups were 1.40 (95%CI: 1.03-1.90, p=0.03) and 1.48 (95%CI: 1.10-

278 2.02, p=0.01) respectively. In conclusion, we found that the DNAm score constructed
279 from the 119 selected CpGs was significantly associated with mortality.

280 ***Biological interpretations of 119 CpGs by gene ontology enrichment analysis***

281 The selected panel of 119 CpGs were located within or near 73 genes. Gene
282 ontology enrichment analysis using these 73 genes resulted in 9 significant pathways
283 with FDR<0.05 (**Figure 6, Table 4**). These pathways included response to type I
284 interferon response (FDR=1.05×10^{-7}), innate immune response (FDR=3.45×10^{-5}),
285 cytokine response (FDR=3.69×10^{-3}) and defense response to virus (**Figure 6**). Our
286 findings suggested that the selected 119 CpG sites are biologically relevant to HIV
287 pathogenesis and progression.

Table 4: Gene ontology term enrichment analysis of the selected CpG sites that predict frailty in an HIV infected population

Term	N	%	Genes	Fold Enrichment	P value	FDR
GO:0034340~response to type I interferon	10	11.8	*OAS2, IFITM1, IFIT3, MX1, NLRC5, HLA-F, XAF1, HLA-E, IRF7, PSMB8*	28.1	6.62E-11	1.05E-07
GO:0045087~innate immune response	19	22.4	*TRIM22, MAP4K2, IFITM1, TRIM25, MX1, GFI1, IFIT5, XAF1, LY86, PSMB8, TRIM14, PARP9, PLSCR1, OAS2, IFIT3, NLRC5, HLA-F, HLA-E, IRF7*	5.0	2.17E-08	3.45E-05
GO:0051607~defense response to virus	11	12.9	*TRIM22, PLSCR1, OAS2, IFITM1, TRIM25, IFIT3, MX1, IFI44L, NLRC5, IFIT5, IRF7*	10.5	7.55E-08	1.20E-04
GO:0051707~response to other organism	18	21.2	*TRIM22, PDE4B, EPO, IFITM1, NOTCH1, TRIM25, MX1, GFI1, IFIT5, IFI44L, LY86, PLSCR1, OAS2, IFIT3,*	4.8	1.17E-07	1.85E-04

			CCDC88B, NLRC5, HLA-E, IRF7			
GO:0016032~viral process	19	22.4	*TRIM22, IFITM1, TFRC, NOTCH1, TRIM25, MX1, PPIB, GFI1, IFIT5, IFI44L, TAP1, PSMB8, TRIM14, PLSCR1, OAS2, SLC1A5, IFIT3, NLRC5, IRF7*	4.3	1.93E-07	3.07E-04
GO:0009615~response to virus	11	12.9	*TRIM22, PLSCR1, OAS2, IFITM1, TRIM25, IFIT3, MX1, IFI44L, NLRC5, IFIT5, IRF7*	7.7	1.27E-06	2.02E-03
GO:0098542~defense response to other organism	13	15.3	*TRIM22, PLSCR1, OAS2, IFITM1, TRIM25, IFIT3, MX1, CCDC88B, IFIT5, NLRC5, IFI44L, HLA-E, IRF7*	5.8	1.94E-06	3.09E-03
GO:0034097~response to cytokine	16	18.8	*TRIM22, EPO, IFITM1, TRIM25, MX1, GFI1, XAF1, PSMB8, PARP9, PLSCR1, OAS2, IFIT3, NLRC5, HLA-F, HLA-E, IRF7*	4.4	2.32E-06	3.69E-03
GO:0034341~response to interferon-gamma	8	9.4	*TRIM22, OAS2, IFITM1, TRIM25, NLRC5, HLA-F, HLA-E, IRF7*	10.9	7.95E-06	1.26E-02

Discussion

In this study, we present evidences that DNAm marks in blood are predictive of frailty and are associated mortality in an HIV-positive population. We identified a panel of 119 CpG sites that were highly predictive for high frailty and moderately predictive for low frailty. We also found that the DNAm score constructed by these 119 CpGs was strongly associated with mortality in the HIV-positive population. More importantly, these

clinically informative 119 DNAm features lied in genes involved in HIV pathogenesis and progression. Thus, we discovered a panel of 119 DNAm biomarkers that add knowledge to the epigenetic mechanisms underlying frailty and mortality among people living with HIV.

We demonstrated the utility of using DNAm marks to predict frailty in HIV-positive individuals. We took the several steps to avoid overfitting in developing the prediction models: 1) model development and evaluation were conducted separately in training set and validation set. DNAm in two sets were profiled in different time with different platforms; 2) 10-fold cross validation were performed during training in each model. The performances of four machine learning methods were in general consistent except that the SVM did not perform well for low frailty. Our results suggested that this panel of CpG sites was relatively stable and robust although performances of predictive models differed using different methods. Of note, SVM showed the worst performance in predicting high and low frailty that highlights the importance of choosing appropriate machine learning method for model development. Compared to our previously reported a panel of 698 smoking-associated CpGs that are moderately predictive of frailty (15), the panel of 119 CpG sites in the present study includes fewer CpGs (119 CpGs) and shows greater prediction performance on frailty (AUC=0.73 for smoking-associated CpGs and AUC 0.83 for the VACS index-associated CpGs), suggesting that the panel of 119 CpG sites has greater clinical utility for frailty among people living with HIV. The improvement of prediction performance on frailty in the present study may be due to different strategies of CpG selections. In this study, the 119 CpGs are selected from the

317 entire blood methylome and may be more informative than the smoking-associated
318 CpGs in our previous report.
319 We also found that the VACS index-related DNAm score derived from the 119
320 CpGs is strongly associated with mortality. This result provides further knowledge of
321 epigenetic profile into previous report that VACS index is predictive of mortality in
322 people who live with HIV (53). The result is also consistent with previous literature
323 showing that DNAm marks in blood predict mortality in the general population (39).
324 Interestingly, we found no overlapped CpG between our predictive panel and the
325 previously identified CpGs for all-cause mortality that included CpGs associated with a
326 variety of diseases such as diabetes and cancer(39). The discrepancy suggest that
327 methylation-based prediction of mortality may be relevant to causal disease. A recent
328 study reports a significant overlap of mortality-associated CpGs and aging-associated
329 CpGs (Lund et al, 2019). We found no overlapped CpG site between our mortality-
330 related 119 CpG panel and aging-related 353 CpGs in epigenetic clock (7), suggesting
331 that epigenetic mechanisms in HIV-related mortality differ from aging-related mortality.
332 Most importantly, this panel of 119 CpG sites are biologically meaningful and
333 may shed light on our understanding of the biological mechanisms of frailty for HIV-
334 positive individuals. The majority of the 119 CpG sites were located within or near
335 genes that are involved in known HIV pathology and progression. The results from DMR
336 analysis are corresponding to some of the significant genes for frailty. For example,
337 cg26312951, cg22862003 and cg21549285 from the 119 CpG sites are located in the
338 *MX1* gene (Interferon-Induced GTP-Binding Protein Mx1) and were negatively
339 associated with frailty. A DMR on *MX1* also showed a significant with frailty. *MX1*

encodes a GTP-metabolizing protein induced by interferon I and II and is involved in interferon gamma signaling and Toll-like signaling pathway. Multiple CpGs near the *HLA-F* and *HLA-E* genes are part of the predictive panel, and these two genes are actively involved in immune response (54). Our previously reported HIV-associated CpG site, cg07839457 in the *NLRC5* gene is also a member of the predictive panel on frailty (25). Other interesting gene regions reveled by DMR analysis include *PARP9 and IFI44L,* which both genes are involved in *HIV pathogenesis.* The biological relevance of these 119 CpG sites was further supported by the gene ontology enrichment analysis. The top enriched pathways such as type I interferon response and cytokine response may point out important biological pathways that leads to frailty and increased risk of mortality among people living with HIV.

We acknowledge several limitations in this study. The generalizability of the selected CpG sites may be limited since our samples were predominantly middle-aged men, which may not generalize to frailty in a different age group. All samples in our study are HIV-positive and the identified CpGs have a limited application to HIV-negative population as we discussed above. Future studies in diverse populations is warranted to validate the selected methylation features. Lastly, the prediction of low frailty is moderate due to imbalanced sample distribution in our present study. We expect that including more samples with low frailty will improve the predictive performance.

Conclusions

We identified a panel of 119 predictive DNAm features in blood that are associated with frailty and mortality among people living with HIV. These DNAm features may serve as biomarkers to discriminate high and low frailty groups and may help to prioritize genes to better understand the mechanisms of frailty in HIV-positive population. These DNAm features have potential to serve as biomarkers to monitor HIV progression in future clinical care.

Legends

Figure 1: Flowchart of analytical procedures on selection of CpG sites in peripheral blood methylome, machine learning prediction models to predict frailty, DNA methylation score to assess mortality, and gene ontology enrichment analysis

Figure 2: Manhattan and quantile-quantile plots on Veteran Aging Cohort Study (VACS) index (λ=1.038) in the training set; blue line indicates False Discovery Rate (FDR)=0.05

Figure 3: Differentially methylation regions (DMRs) are associated with frailty, measured by Veteran Aging Cohort Study (VACS) index. a: *MX1* regional plot showing methylation beta value among subjects with VACS index > 50 and VACS index ≤ 50 in the training set. b: *IFI11L* regional plot showing methylation beta value among subjects with VACS index > 50 and VACS index ≤ 50 in the training set. c: *PARP9* regional plot showing methylation beta value among subjects with VACS index > 50 and VACS index ≤ 50 in the training set

Figure 4: Receiver operating characteristic curve by XGBoost prediction model on HIV frailty; a) Predicting high HIV frailty (VACS index >50 vs. ≤50), Area Under Curve (AUC): 0.830 (95% CI: 0.786,0.874); b) Predicting low HIV frailty (VACS index >16 vs. ≤16), AUC: 0.735 (95% CI: 0.688,0.782)

Figure 5: Kaplan-Meier curves comparing high and low HIV frailty methylation score groups. Methylation score is calculated by 119 selected probes. High methylation score group has methylation score >=0, while low methylation score group <0.

Figure 6: Gene ontology enrichment analysis of 119 CpG probes (FDR<0.05) for prediction

391 **Figure S1**: Distribution of Veteran Aging Cohort Study index (VACS index) between
392 training set and testing set
393
394 **Table 1**: Study sample characteristics
395 **Table 2**: A panel of 119 CpG sites in blood that predicts high and low frailty in a HIV-
396 positive veteran population
397 **Table 3**: Performance of machine learning model predicting high and low frailty in an
398 HIV-positive population
399 **Table 4**: Gene ontology term enrichment analysis of the selected 119 CpG sites that
400 predict frailty in an HIV-positive population
401 **Table S1**: Differentially methylated regions between high and low frailty in training set
402

Table 2: A panel of 119 CpG sites in blood that predicts high and low frailty in a HIV-infected veteran population

CpG	Chr	Position	Gene	Gene group	Relation to CpG Island	Effect size	SE	P value	FDR*
cg01309328	6	32811253	*PSMB8*	Body	N_Shore	-5.30	0.888	4.44E-09	9.32E-04
cg08099136	6	32811251	*PSMB8*	Body	N_Shore	-4.77	0.800	4.63E-09	9.32E-04
cg26312951	21	42797847	*MX1*	TSS200;5UTR	N_Shore	-2.46	0.417	6.40E-09	9.32E-04
cg08818207	6	32820355	*TAP1*	Body	N_Shore	-3.40	0.627	9.17E-08	1.00E-02
cg08122652	3	122281939	*PARP9*	5UTR;TSS1500	N_Shore	-1.79	0.341	2.18E-07	1.45E-02
cg07352001	10	44144445	*ZNF32*	TSS200;TSS1500	Island	-14.30	2.730	2.31E-07	1.45E-02
cg07325529	6	6613762	*LY86*	Body		20.70	3.980	2.77E-07	1.45E-02
cg16242615	19	4059988	*ZBTB7A*	5UTR	Island	-6.27	1.200	2.85E-07	1.45E-02
cg00533183	6	32810742	*PSMB8*	Body	N_Shore	-5.81	1.120	3.18E-07	1.45E-02
cg12649038	10	116282534	*ABLIM1*	5UTR		5.03	0.973	3.32E-07	1.45E-02
cg08926253	11	614761	*IRF7*	Body	Island	-2.91	0.571	4.94E-07	1.96E-02
cg03038262	11	315262	*IFITM1*	3UTR	N_Shore	-2.86	0.565	5.83E-07	2.12E-02
cg07839457	16	57023022	*NLRC5*	TSS1500	N_Shore	-1.99	0.396	6.55E-07	2.12E-02
cg22930808	3	122281881	*PARP9*	5UTR	N_Shore	-1.54	0.306	6.78E-07	2.12E-02
cg03607951	1	79085586	*IFI44L*	TSS1500		-2.00	0.398	7.41E-07	2.16E-02
cg08260450	6	34993987	*ANKS1A*	Body		6.35	1.270	8.43E-07	2.30E-02
cg03917473	17	38764244				9.10	1.840	1.04E-06	2.67E-02
cg18234224	1	172917851				6.59	1.350	1.34E-06	3.22E-02
cg06188083	10	91093005	*IFIT3*	Body		-2.02	0.414	1.40E-06	3.22E-02
cg23570810	11	315102	*IFITM1*	Body	N_Shore	-2.31	0.475	1.58E-06	3.41E-02
cg03816851	22	18324769	*MICAL3*	Body	Island	10.30	2.120	1.64E-06	3.41E-02
cg20676542	17	54991456	*TRIM25*	TSS200	Island	-3.02	0.629	2.10E-06	4.17E-02
cg05019807	9	139410393	*NOTCH1*	Body	N_Shore	5.97	1.260	2.70E-06	5.07E-02
cg03753191	13	43566902	*EPSTI1*	TSS1500	S_Shore	-5.39	1.140	2.88E-06	5.07E-02
cg01971407	11	313624	*IFITM1*	TSS1500	N_Shelf	-3.53	0.745	2.90E-06	5.07E-02
cg03359362	19	47289611	*SLC1A5*	TSS1500;Body;5UTR	N_Shore	-12.60	2.670	3.13E-06	5.10E-02
cg22116398	5	196162			S_Shore	5.15	1.090	3.15E-06	5.10E-02
cg26922780	16	88769443	*RNF166*	Body	N_Shelf	4.27	0.907	3.28E-06	5.12E-02
cg15748006	2	9772375	*YWHAQ*	TSS1500	S_Shore	4.45	0.949	3.45E-06	5.20E-02

cg22862003	21	42797588	*MX1*	TSS1500;5UTR	N_Shore	-1.66	0.356	3.88E-06	5.65E-02
cg21366673	6	30459512	*HLA-E*	Body	S_Shore	-4.19	0.900	4.04E-06	5.69E-02
cg05588757	2	95825608	*ZNF514*	TSS1500	Island	-19.40	4.220	5.25E-06	6.64E-02
cg06872964	1	79085250	*IFI44L*	TSS1500		-1.92	0.416	5.26E-06	6.64E-02
cg06376949	10	91173811	*IFIT5*	TSS1500	N_Shore	-3.21	0.698	5.29E-06	6.64E-02
cg01154505	2	112940409	*FBLN7*	Body	S_Shore	6.16	1.340	5.32E-06	6.64E-02
cg01871595	13	39336304	*FREM2*	Body		6.90	1.500	5.59E-06	6.78E-02
cg05696877	1	79088769	*IFI44L*	5UTR		-1.21	0.265	6.01E-06	6.94E-02
cg24497541	1	55352819	*DHCR24*	1stExon;5UTR	Island	-9.81	2.150	6.12E-06	6.94E-02
cg11829870	22	50988451	*KLHDC7B*	3UTR;1stExon	S_Shore	-3.63	0.795	6.28E-06	6.94E-02
cg13720750	5	112309167			N_Shelf	7.28	1.600	6.37E-06	6.94E-02
cg18255813	6	7195966	*RREB1*	Body		9.38	2.060	6.76E-06	6.94E-02
cg21549285	21	42799141	*MX1*	5UTR	S_Shore	-1.01	0.224	7.08E-06	6.94E-02
cg14945867	14	54908007	*CNIH*	1stExon	Island	-29.70	6.550	7.16E-06	6.94E-02
cg09026253	11	313267	*IFITM1*	TSS1500	S_Shore	-3.48	0.768	7.25E-06	6.94E-02
cg14651616	11	64563992	*MAP4K2*	Body		-6.98	1.540	7.27E-06	6.94E-02
cg24082730	3	126076366	*KLF15*	TSS200	Island	-10.90	2.410	7.31E-06	6.94E-02
cg16302816	11	16834800	*PLEKHA7*	Body		6.50	1.440	7.49E-06	6.96E-02
cg01765174	9	100880960	*TRIM14*	Body	N_Shore	-4.47	0.993	8.35E-06	7.60E-02
cg10552523	11	313478	*IFITM1*	TSS1500	N_Shelf	-3.14	0.702	9.42E-06	8.40E-02
cg19025187	3	195808255	*TFRC*	5UTR	N_Shore	-6.73	1.510	1.07E-05	9.16E-02
cg21081878	21	38334730	*HLCS*	5UTR	N_Shelf	3.49	0.785	1.08E-05	9.16E-02
cg14299044	10	19972179				5.32	1.200	1.09E-05	9.16E-02
cg24871132	3	149688846	*PFN2*	TSS200	Island	-21.70	4.900	1.17E-05	9.55E-02
cg27294701	16	88107336	*BANP*	Body	Island	7.04	1.590	1.18E-05	9.55E-02
cg09296453	6	29692035	*HLA-F*	Body	Island	-3.24	0.735	1.31E-05	1.04E-01
cg16400434	11	73882363	*PPME1;C2CD3*	TSS200;TSS1500	Island	-25.60	5.850	1.46E-05	1.13E-01
cg16656286	17	4981603	*ZFP3*	TSS200	Island	22.40	5.130	1.52E-05	1.13E-01
cg26480543	19	55629279	*PPP1R12C*	TSS1500	S_Shore	-7.33	1.680	1.52E-05	1.13E-01
cg19371652	12	113415883	*OAS2*	TSS1500		-3.59	0.822	1.52E-05	1.13E-01
cg09597638	17	3907349			N_Shore	-5.48	1.260	1.56E-05	1.14E-01
cg00994629	14	22694547				-7.44	1.710	1.62E-05	1.16E-01

cg02314339	10	91020653				-4.35	1.010	1.86E-05	1.29E-01
cg21468416	9	127019914	*NEK6*	1stExon;TSS1500;5UTR	N_Shore	6.61	1.530	1.89E-05	1.29E-01
cg12149905	3	61547208	*PTPRG*	TSS200	Island	-12.20	2.820	1.89E-05	1.29E-01
cg20927242	6	29692011	*HLA-F*	Body	Island	-7.14	1.660	1.97E-05	1.32E-01
cg03332034	16	1823786	*EME2; MRPS34*	TSS1500	Island	-12.30	2.870	2.12E-05	1.36E-01
cg05489271	12	123871695	*SETD8*	Body	N_Shelf	7.83	1.820	2.13E-05	1.36E-01
cg25843003	6	31431312	*HCP5*	3UTR		-3.29	0.768	2.19E-05	1.36E-01
cg14090510	7	64839023	*ZNF92*	5UTR;Body	Island	-14.70	3.440	2.21E-05	1.36E-01
cg25467833	1	212335061				4.86	1.140	2.28E-05	1.36E-01
cg03725115	6	30458102	*HLA-E*	Body	Island	-5.09	1.190	2.29E-05	1.36E-01
cg15620384	6	34164405			Island	-6.69	1.570	2.34E-05	1.36E-01
cg08450404	19	58326441	*ZNF552*	TSS200	S_Shore	-8.37	1.960	2.35E-05	1.36E-01
cg06927297	12	117175889	*RNFT2*	TSS200;TSS1500	Island	-27.60	6.460	2.36E-05	1.36E-01
cg21684411	6	31431573	*HCP5*	3UTR		-6.71	1.570	2.40E-05	1.36E-01
cg20998539	17	62208374	*ERN1*	TSS1500	S_Shore	-7.32	1.720	2.42E-05	1.36E-01
cg25138854	1	9555557			Island	-7.03	1.650	2.45E-05	1.36E-01
cg13250752	4	138281992				-6.09	1.430	2.46E-05	1.36E-01
cg26562691	16	23850404	*PRKCB*	Body	S_Shelf	6.87	1.620	2.47E-05	1.36E-01
cg23677352	4	149715675				-7.92	1.860	2.49E-05	1.36E-01
cg22107533	15	45028083	*TRIM69*	TSS1500		-3.14	0.739	2.53E-05	1.36E-01
cg06412917	12	124858844	*NCOR2*	Body	S_Shelf	6.47	1.520	2.57E-05	1.37E-01
cg01190666	20	62204908	*PRIC285*	5UTR	N_Shore	-3.47	0.818	2.67E-05	1.40E-01
cg11977562	13	114845297	*RASA3*	Body	N_Shelf	7.05	1.660	2.70E-05	1.40E-01
cg13304609	1	79085162	*IFI44L*	TSS1500		-1.54	0.364	2.72E-05	1.40E-01
cg01537765	19	42914828	*LIPE*	Body	Island	22.00	5.210	2.75E-05	1.40E-01
cg08275025	11	314493	*IFITM1*	Body	N_Shore	-4.98	1.180	2.78E-05	1.40E-01
cg05362517	13	37393368	*RFXAP*	5UTR;1stExon	Island	-13.80	3.260	2.92E-05	1.42E-01
cg21145248	5	176816679	*SLC34A1*	Body		6.04	1.430	2.94E-05	1.42E-01
cg07537655	6	161415612	*MAP3K4*	Body	S_Shelf	1.77	0.419	2.95E-05	1.42E-01
cg26912671	1	66458803	*PDE4B*	1stExon		10.70	2.540	3.01E-05	1.42E-01

cg11905821	6	31770932	*LSM2*	Body	N_Shelf	11.80	2.800	3.05E-05	1.42E-01
cg06202053	11	64109021	*CCDC88B*	Body	N_Shore	-5.84	1.390	3.09E-05	1.42E-01
cg13794687	7	123302399	*LMOD2*	Body		6.73	1.600	3.11E-05	1.42E-01
cg23892836	6	29692085	*HLA-F*	Body	Island	-3.19	0.759	3.12E-05	1.42E-01
cg15331332	6	29692111	*HLA-F*	Body	S_Shore	-3.53	0.840	3.12E-05	1.42E-01
cg15804432	15	64454965	*PPIB*	Body	Island	-4.27	1.020	3.15E-05	1.42E-01
cg09251764	17	6659070	*XAF1*	TSS200		-5.73	1.370	3.23E-05	1.44E-01
cg20893717	7	100318190	*EPO*	TSS1500	Island	13.80	3.290	3.26E-05	1.44E-01
cg08159663	16	57022486	*NLRC5*	TSS1500	N_Shore	-3.44	0.821	3.29E-05	1.44E-01
cg10019429	19	32836659	*ZNF507*	1stExon;5UTR	Island	-17.20	4.120	3.55E-05	1.54E-01
cg11791770	11	611791	*PHRF1*	3UTR	Island	-6.52	1.560	3.64E-05	1.56E-01
cg11653134	2	66805547			S_Shore	-4.24	1.020	3.81E-05	1.61E-01
cg06981309	3	146260954	*PLSCR1*	5UTR	N_Shore	-1.96	0.471	3.84E-05	1.61E-01
cg07168939	8	143763412	*PSCA*	Body		-5.88	1.420	3.88E-05	1.61E-01
cg04611649	2	152681240	*ARL5A*	5UTR	N_Shelf	7.42	1.790	4.04E-05	1.66E-01
cg16297569	1	92952517	*GFI1*	TSS1500;TSS200	Island	-8.34	2.020	4.16E-05	1.70E-01
cg01518846	6	26246970	*HIST1H4G*	1stExon	Island	-17.50	4.240	4.41E-05	1.74E-01
cg26852894	4	11157341				6.31	1.530	4.42E-05	1.74E-01
cg06811183	3	48510438	*SHISA5*	3UTR		-8.23	2.000	4.42E-05	1.74E-01
ch.10.1305203R	10	63407435				-8.63	2.090	4.43E-05	1.74E-01
cg12461141	11	5710654	*TRIM22*	TSS1500		-3.40	0.829	4.79E-05	1.87E-01
cg13149600	21	43374220	*C2CD2*	TSS1500	S_Shore	-9.35	2.280	4.83E-05	1.87E-01
cg12660813	1	3192343	*PRDM16*	Body	N_Shelf	7.28	1.780	4.96E-05	1.89E-01
cg24447788	19	795310			N_Shore	11.30	2.760	5.02E-05	1.89E-01
cg13861758	9	138148099			N_Shelf	5.08	1.240	5.02E-05	1.89E-01
cg09950208	10	130841235				-4.99	1.220	5.24E-05	1.95E-01
cg01482620	19	48835971	*TMEM143*	3UTR	N_Shore	18.40	4.500	5.28E-05	1.95E-01
cg03634735	7	1992524	*MAD1L1*	Body	S_Shore	6.19	1.520	5.37E-05	1.97E-01

*FDR: False discovery rate

403

404

Table S1: Differentially methylated regions between high and low fraity in training set

chr	start	end	value	area	Number of CpG	p.value	fwer	Gene.Symbol	FDR	CpG
3	122281881	122281975	-0.068	0.203	3	2.30E-07	0.001	*DTX3L,PARP9*	1.03E-04	cg00959259; cg08122652; cg22930808
12	11700321	11700489	-0.054	0.162	3	1.61E-06	0.006	*LINC01252*	3.60E-04	cg06202470; cg18232235; cg19651115
1	79088769	79118191	-0.039	0.154	4	1.17E-05	0.049	*IFI44,IFI44L*	1.75E-03	cg00458211; cg01079652; cg05696877; cg07107453
21	42797588	42797847	-0.046	0.092	2	2.41E-05	0.081	*MX1*	2.68E-03	cg22862003; cg26312951
1	79085162	79085765	-0.034	0.168	5	2.99E-05	0.122	*IFI44L*	2.68E-03	cg00855901; cg03607951; cg06872964; cg13304609; cg17980508
11	319555	319718	-0.036	0.109	3	4.73E-05	0.166	*IFITM3*	2.80E-03	cg20045320; cg09122035; cg17990365
11	312518	313624	-0.023	0.230	10	4.89E-05	0.194	*IFITM1*	2.80E-03	cg01886988; cg01971407; cg04582010; cg05432003; cg09026253; cg10552523; cg11694510; cg20566897; cg22963452; cg27032101
11	314493	317767	-0.021	0.228	11	5.01E-05	0.195	*IFITM1*	2.80E-03	cg12047941; cg15013527; cg25050332; cg27331665; cg03038262; cg08000731; cg08275025; cg16379091; cg18434560; cg21686213; cg23570810
21	42798747	42799141	-0.041	0.083	2	5.95E-05	0.197	*MX1*	2.96E-03	cg08924203; cg21549285

405

406 **List of abbreviations**

407 AUC: Area Under Curve

408 CI: Confidence interval

409 DMR: differentially methylated region

410 DNA: Deoxyribonucleic acid

411 DNAm: DNA methylation

412 DAVID: Database for Annotation, Visualization and Integrated Discovery

413 EWAS: epigenome-wide association study

414 FDR: False discovery rate

415 FWER: Family-wise error rate

416 GLMNET: Lasso and Elastic-Net Regularized Generalized Linear Models

417 GO: Gene ontology

418 HIV: Human immunodeficiency virus

419 HM450K: Human Methylation 450K BeadChip

420 NK: Natural killer

421 PC: Principal component

422 QC: Quality control

423 SVM: Support Vector Machines

424 VACS: Veterans Aging Cohort Study

425 VACS index: Veterans Aging Cohort Study index

426 XGBoost: Extreme Gradient Boosting Tree

427

428

Declarations

Ethics approval and consent to participate

The study was approved by the committee of the Human Research Subject Protection at Yale University and the Institutional Research Board Committee of the Connecticut Veteran Healthcare System. All subjects provided written consents.

Availability of data and materials

Demographic and clinical variables and DNAm data for the VACS samples were submitted to GEO dataset (GSE117861) and are available to the public. All codes for analysis are also available upon a request to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Funding

The project was supported by the National Institute on Drug Abuse (R03DA039745, R01DA038632, R01DA047063, R01DA047820) and the National Center for Post-Traumatic Stress Disorder, USA.

Authors' contributions

CS was responsible for data analysis and manuscript preparation. ACJ provided DNA samples, clinical data, and contributed to manuscript preparation. XZ was responsible for the bioinformatics data processing. VM involved clinical data collection and manuscript preparation. DH and EJ contributed to analytical approach and the manuscript preparation. KX was responsible for the study design, study protocol, sample preparation, data analysis, interpretation of findings, and manuscript preparation.

Acknowledgements

The authors appreciate the support of the Veteran Aging Study Cohort Biomarker Core and Yale Center of Genomic Analysis.

References

1. Kooij KW, Wit FW, Schouten J, van der Valk M, Godfried MH, Stolte IG, et al. HIV infection is independently associated with frailty in middle-aged HIV type 1-infected individuals compared with similar but uninfected controls. AIDS. 2016;30(2):241-50.
2. Dent E, Kowal P, Hoogendijk EO. Frailty measurement in research and clinical practice: A review. Eur J Intern Med. 2016;31:3-10.
3. Lam K, Pan K, Linnekamp JF, Medema JP, Kandimalla R. DNA methylation based biomarkers in colorectal cancer: a systematic review. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2016;1866(1):106-20.
4. Teroganova N, Girshkin L, Suter CM, Green MJ. DNA methylation in peripheral tissue of schizophrenia and bipolar disorder: a systematic review. BMC genetics. 2016;17(1):27.
5. Bakusic J, Schaufeli W, Claes S, Godderis L. Stress, burnout and depression: A systematic review on DNA methylation mechanisms. Journal of Psychosomatic Research. 2017;92:34-44.
6. Li M, D'Arcy C, Li X, Zhang T, Joober R, Meng X. What do DNA methylation studies tell us about depression? A systematic review. Translational psychiatry. 2019;9(1):68.
7. Horvath S. DNA methylation age of human tissues and cell types. Genome biology. 2013;14(10):3156.
8. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature Reviews Genetics. 2018:1.
9. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell. 2013;49(2):359-67.
10. Liu C, Marioni RE, Hedman ÅK, Pfeiffer L, Tsai P-C, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Molecular psychiatry. 2018;23(2):422.
11. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. The American Journal of Human Genetics. 2011;88(4):450-7.
12. Gao X, Zhang Y, Breitling LP, Brenner H. Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration. Oncotarget. 2016;7(30):46878-89.
13. Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environmental health perspectives. 2012;120(10):1425-31.
14. Lee KW, Pausova Z. Cigarette smoking and DNA methylation. Frontiers in genetics. 2013;4:132.
15. Zhang X, Hu Y, Aouizerat BE, Peng G, Marconi VC, Corley MJ, et al. Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality. 2018;10(1):155.
16. Zhang R, Miao Q, Wang C, Zhao R, Li W, Haile CN, et al. Genome-wide DNA methylation analysis in alcohol dependence. Addiction biology. 2013;18(2):392-403.
17. Kraiczy J, Nayak KM, Howell KJ, Ross A, Forbester J, Salvestrini C, et al. DNA methylation defines regional identity of human intestinal epithelial organoids and undergoes dynamic changes during development. Gut. 2019;68(1):49-61.

18. Nano J, Ghanbari M, Wang W, de Vries PS, Dhana K, Muka T, et al. Epigenome-Wide Association Study Identifies Methylation Sites Associated With Liver Enzymes and Hepatic Steatosis. Gastroenterology. 2017;153(4):1096-106.e2.
19. Delpu Y, Cordelier P, Cho W, Torrisani J. DNA methylation and cancer diagnosis. International journal of molecular sciences. 2013;14(7):15029-58.
20. Figueroa ME, Lugthart S, Li Y, Erpelinck-Verschueren C, Deng X, Christos PJ, et al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. Cancer cell. 2010;17(1):13-27.
21. Holm K, Hegardt C, Staaf J, Vallon-Christersson J, Jönsson G, Olsson H, et al. Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. Breast cancer research. 2010;12(3):R36.
22. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. Nature Reviews Genetics. 2019;20(2):109-27.
23. Mohammad HP, Barbash O, Creasy CL. Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer. Nat Med. 2019;25(3):403-18.
24. Erlandson KM, Ng DK, Jacobson LP, Margolick JB, Dobs AS, Palella Jr FJ, et al. Inflammation, immune activation, immunosenescence, and hormonal biomarkers in the frailty-related phenotype of men with or at risk for HIV infection. The Journal of infectious diseases. 2016;215(2):228-37.
25. Zhang X, Justice AC, Hu Y, Wang Z, Zhao H, Wang G, et al. Epigenome-wide differential DNA methylation between HIV-infected and uninfected individuals. Epigenetics. 2016;11(10):750-60.
26. Chen J, Huang Y, Hui Q, Mathur R, Gwinn M, So-Armah K, et al. Epigenetic Associations with Estimated Glomerular Filtration Rate (eGFR) among Men with HIV Infection. Clin Infect Dis. 2019.
27. Mathur R, Hui Q, Huang Y, Gwinn M, So-Armah K, Freiberg MS, et al. DNA Methylation Markers of Type 2 Diabetes Mellitus Among Male Veterans With or Without Human Immunodeficiency Virus Infection. J Infect Dis. 2019;219(12):1959-62.
28. Breitling LP, Saum K-U, Perna L, Schöttker B, Holleczek B, Brenner H. Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. Clinical Epigenetics. 2016;8(1):21.
29. Sanchez-Conde M, Rodriguez-Centeno J, Dronda F, Lopez JC, Jimenez Z, Berenguer J, et al. Frailty phenotype: a clinical marker of age acceleration in the older HIV-infected population. Epigenomics. 2019;11(5):501-9.
30. Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock. The Journal of infectious diseases. 2015;212(10):1563-73.
31. Rickabaugh TM, Baxter RM, Sehl M, Sinsheimer JS, Hultin PM, Hultin LE, et al. Acceleration of age-associated methylation patterns in HIV-1-infected adults. PloS one. 2015;10(3):e0119201.
32. Gross AM, Jaeger PA, Kreisberg JF, Licon K, Jepsen KL, Khosroheidari M, et al. Methylome-wide analysis of chronic HIV infection reveals five-year increase in biological age and epigenetic targeting of HLA. Molecular cell. 2016;62(2):157-68.

542 33. Nelson KN, Hui Q, Rimland D, Xu K, Freiberg MS, Justice AC, et al. Identification of HIV
543 infection-related DNA methylation sites and advanced epigenetic aging in HIV-positive,
544 treatment-naive U.S. veterans. AIDS. 2017;31(4):571-5.
545 34. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA methylation age
546 of blood predicts all-cause mortality in later life. Genome biology. 2015;16(1):25.
547 35. Marioni RE, Harris SE, Shah S, McRae AF, von Zglinicki T, Martin-Ruiz C, et al. The
548 epigenetic clock and telomere length are independently associated with chronological age and
549 mortality. International journal of epidemiology. 2016;45(2):424-32.
550 36. Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, Brenner H. Epigenetic age
551 acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort.
552 Clinical epigenetics. 2016;8(1):64.
553 37. Christiansen L, Lenart A, Tan Q, Vaupel JW, Aviv A, McGue M, et al. DNA methylation
554 age is associated with mortality in a longitudinal Danish twin study. Aging Cell. 2016;15(1):149-
555 54.
556 38. Fransquet PD, Wrigglesworth J, Woods RL, Ernst ME, Ryan J. The epigenetic clock as a
557 predictor of disease and mortality risk: a systematic review and meta-analysis. Clin Epigenetics.
558 2019;11(1):62.
559 39. Zhang Y, Wilson R, Heiss J, Breitling LP, Saum K-U, Schöttker B, et al. DNA methylation
560 signatures in peripheral blood strongly predict all-cause mortality. Nature communications.
561 2017;8:14617.
562 40. Tate JP, Justice AC, Hughes MD, Bonnet F, Reiss P, Mocroft A, et al. An internationally
563 generalizable risk index for mortality after one year of antiretroviral therapy. Aids.
564 2013;27(4):563-72.
565 41. Justice AC, Dombrowski E, Conigliaro J, Fultz SL, Gibson D, Madenwald T, et al. Veterans
566 aging cohort study (VACS): overview and description. 2006;44(8 Suppl 2):S13.
567 42. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan S-T, et al. A coherent approach for
568 analysis of the Illumina HumanMethylation450 BeadChip improves data quality and
569 performance in epigenome-wide association studies. 2015;16(1):37.
570 43. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al.
571 Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA
572 methylation microarrays. 2014;30(10):1363-9.
573 44. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to
574 identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol.
575 2012;41(1):200-9.
576 45. Bebu I, Tate J, Rimland D, Mesner O, Macalino GE, Ganesan A, et al. The VACS index
577 predicts mortality in a young, healthy HIV population starting highly active antiretroviral
578 therapy. Journal of acquired immune deficiency syndromes (1999). 2014;65(2):226.
579 46. Kuhn MJJoss. Building predictive models in R using the caret package. 2008;28(5):1-26.
580 47. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data
581 mining, inference and prediction. The Mathematical Intelligencer. 2005;27(2):83-5.
582 48. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of
583 classification techniques. Emerging artificial intelligence applications in computer engineering.
584 2007;160:3-24.

585 49. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the
586 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016:
587 ACM.
588 50. Ogutu JO, Piepho H-P, Schulz-Streeck T. A comparison of random forests, boosting and
589 support vector machines for genomic selection. BMC Proc. 2011;5 Suppl 3(Suppl 3):S11-S.
590 51. Brodersen KH, Ong CS, Stephan KE, Buhmann JM, editors. The balanced accuracy and its
591 posterior distribution. 2010 20th International Conference on Pattern Recognition; 2010: IEEE.
592 52. Fresno C, Fernández EAJB. RDAVIDWebService: a versatile R interface to DAVID.
593 2013;29(21):2810-1.
594 53. Justice AC, Modur SP, Tate JP, Althoff KN, Jacobson LP, Gebo KA, et al. Predictive
595 accuracy of the Veterans Aging Cohort Study index for mortality with HIV infection: a North
596 American cross cohort analysis. Journal of acquired immune deficiency syndromes (1999).
597 2013;62(2):149-63.
598 54. O'Callaghan CA, Bell JIJIr. Structure and function of the human MHC class Ib molecules
599 HLA-E, HLA-F and HLA-G. 1998;163(1):129-38.
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

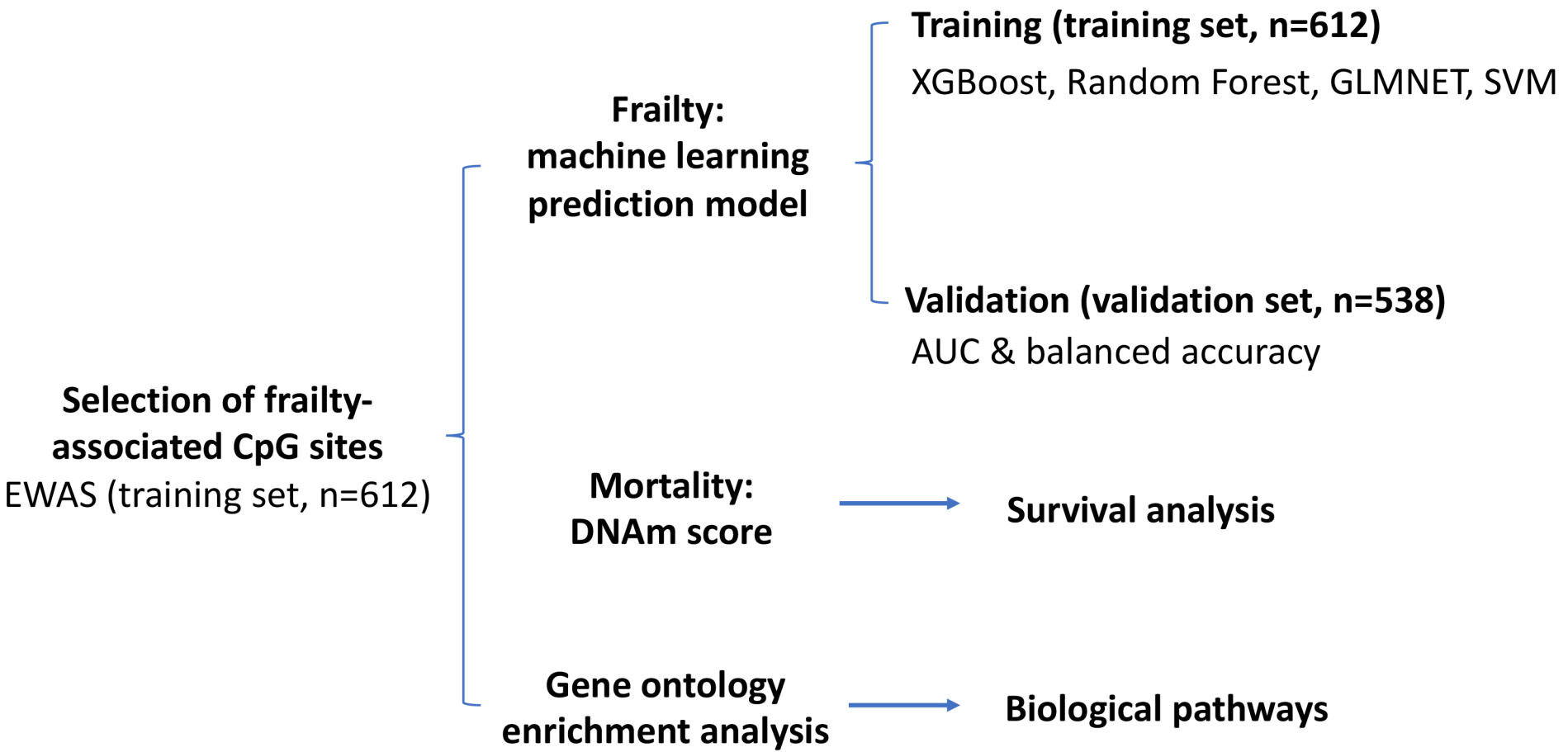
Training (training set, n=612)
XGBoost, Random Forest, GLMNET, SVM
Frailty:
machine learning
prediction model
Validation (validation set, n=538)
AUC & balanced accuracy
Selection of frailty-associated CpG sites
EWAS (training set, n=612)
Mortality:
DNAm score
Survival analysis
Gene ontology
enrichment analysis
Biological pathways

Figure 1

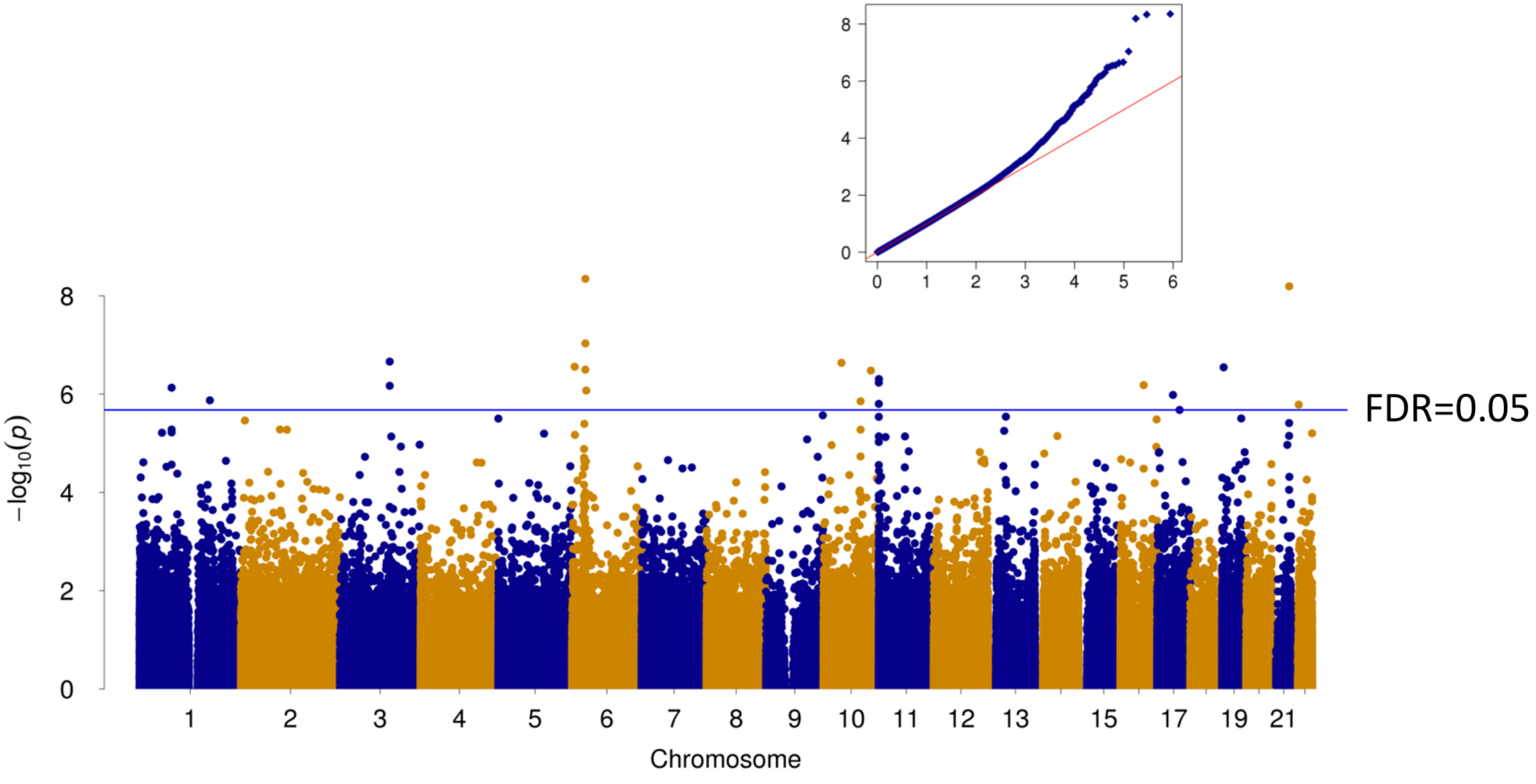
8
6
4
2
0
0 1 2 3 4 5 6
8
6
4
2
0
$-\log_{10}(p)$
FDR=0.05
1 2 3 4 5 6 7 8 9 10 11 12 13 15 17 19 21
Chromosome

Figure 2

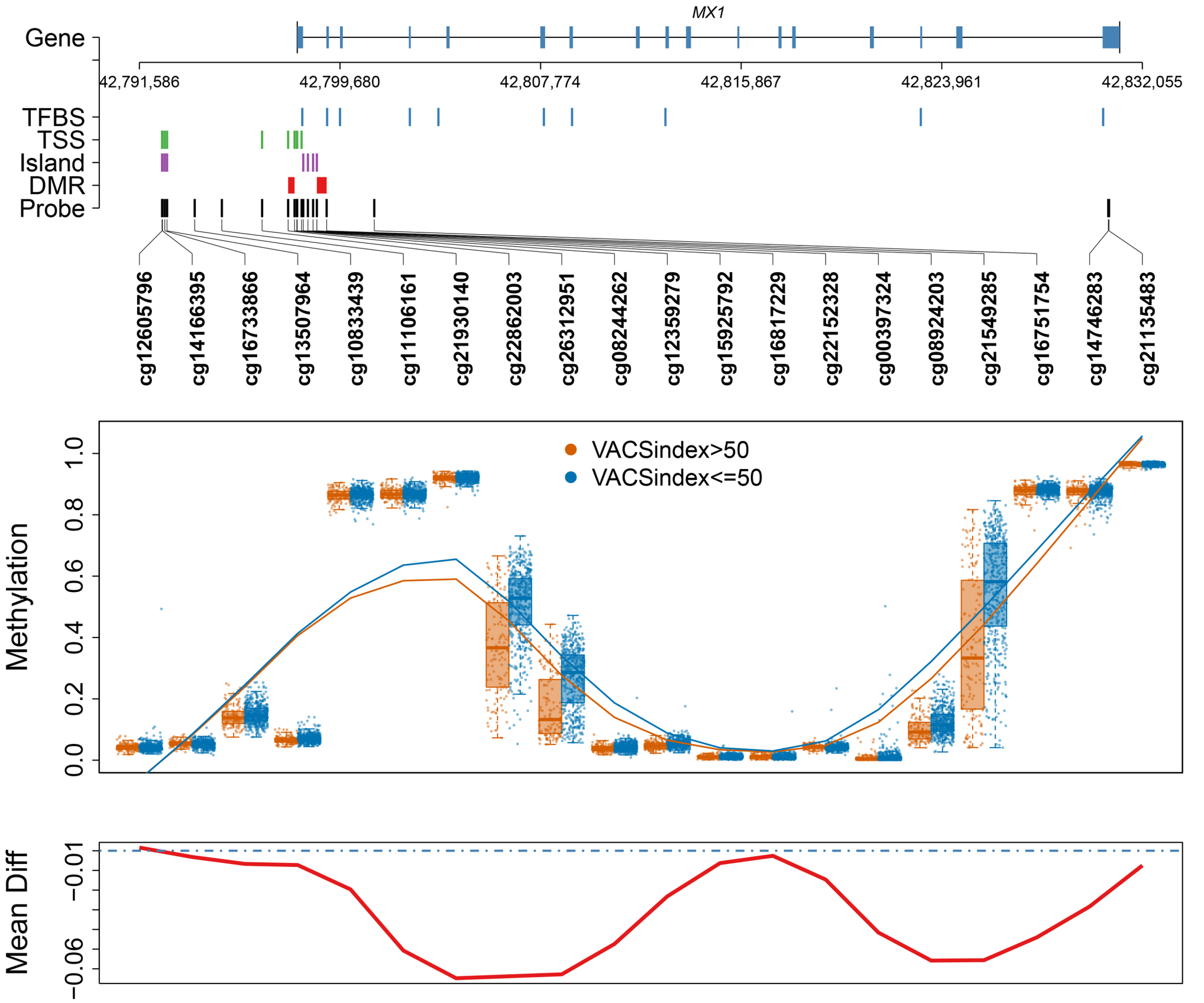
MX1
Gene
42,791,586
42,799,680
42,807,774
42,815,867
42,823,961
42,832,055
TFBS
TSS
Island
DMR
Probe
cg12605796
cg14166395
cg16733866
cg13507964
cg10833439
cg11106161
cg21930140
cg22862003
cg26312951
cg08244262
cg12359279
cg15925792
cg16817229
cg22152328
cg00397324
cg08924203
cg21549285
cg16751754
cg14746283
cg21135483
VACSindex>50
VACSindex<=50
Methylation
0.0
0.2
0.4
0.6
0.8
1.0
Mean Diff
−0.01
−0.06

Figure 3a

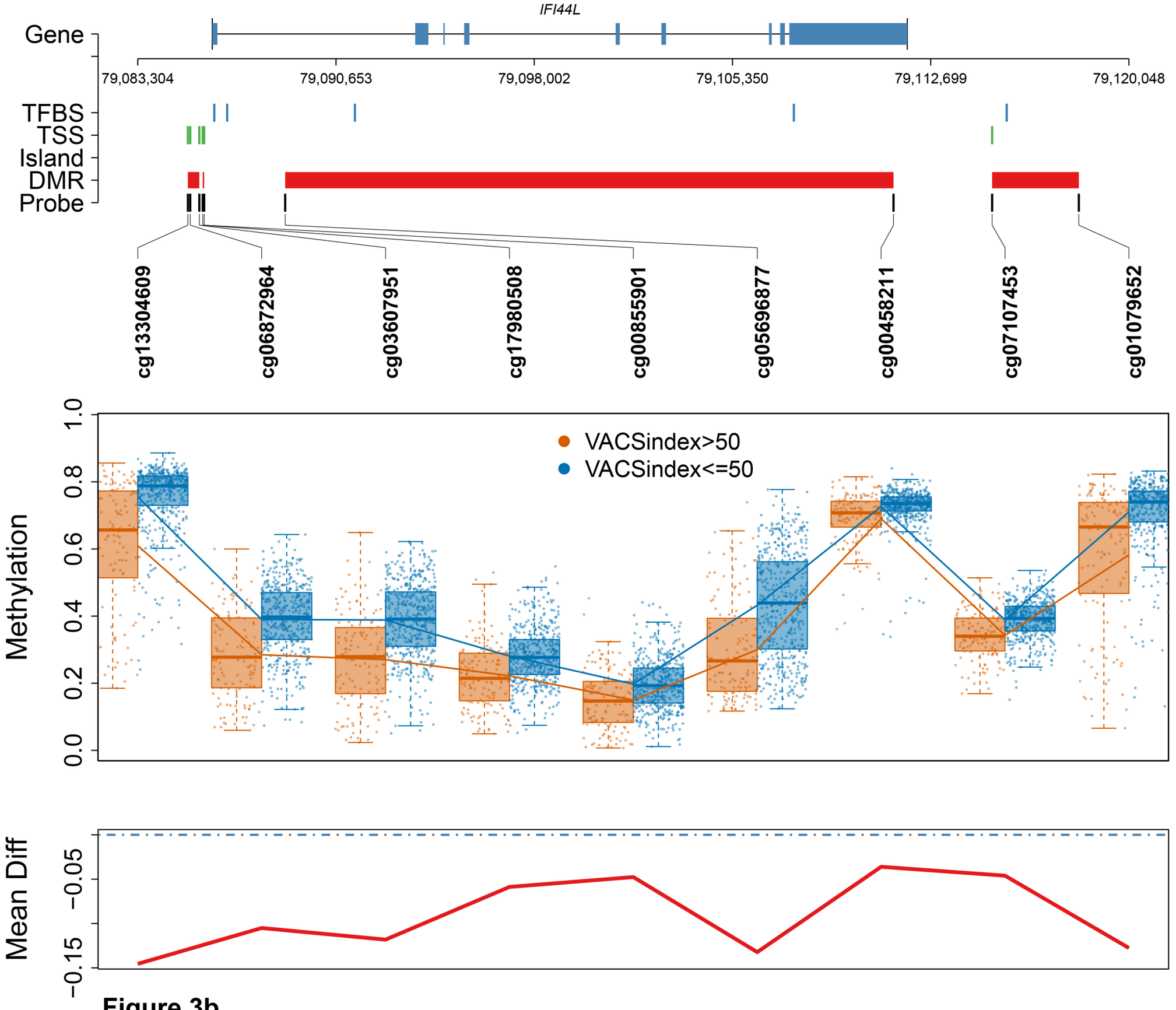
IFI44L
Gene
79,083,304
79,090,653
79,098,002
79,105,350
79,112,699
79,120,048
TFBS
TSS
Island
DMR
Probe
cg13304609
cg06872964
cg03607951
cg17980508
cg00855901
cg05696877
cg00458211
cg07107453
cg01079652
VACSindex>50
VACSindex<=50
Methylation
0.0
0.2
0.4
0.6
0.8
1.0
Mean Diff
−0.15
−0.05

Figure 3b

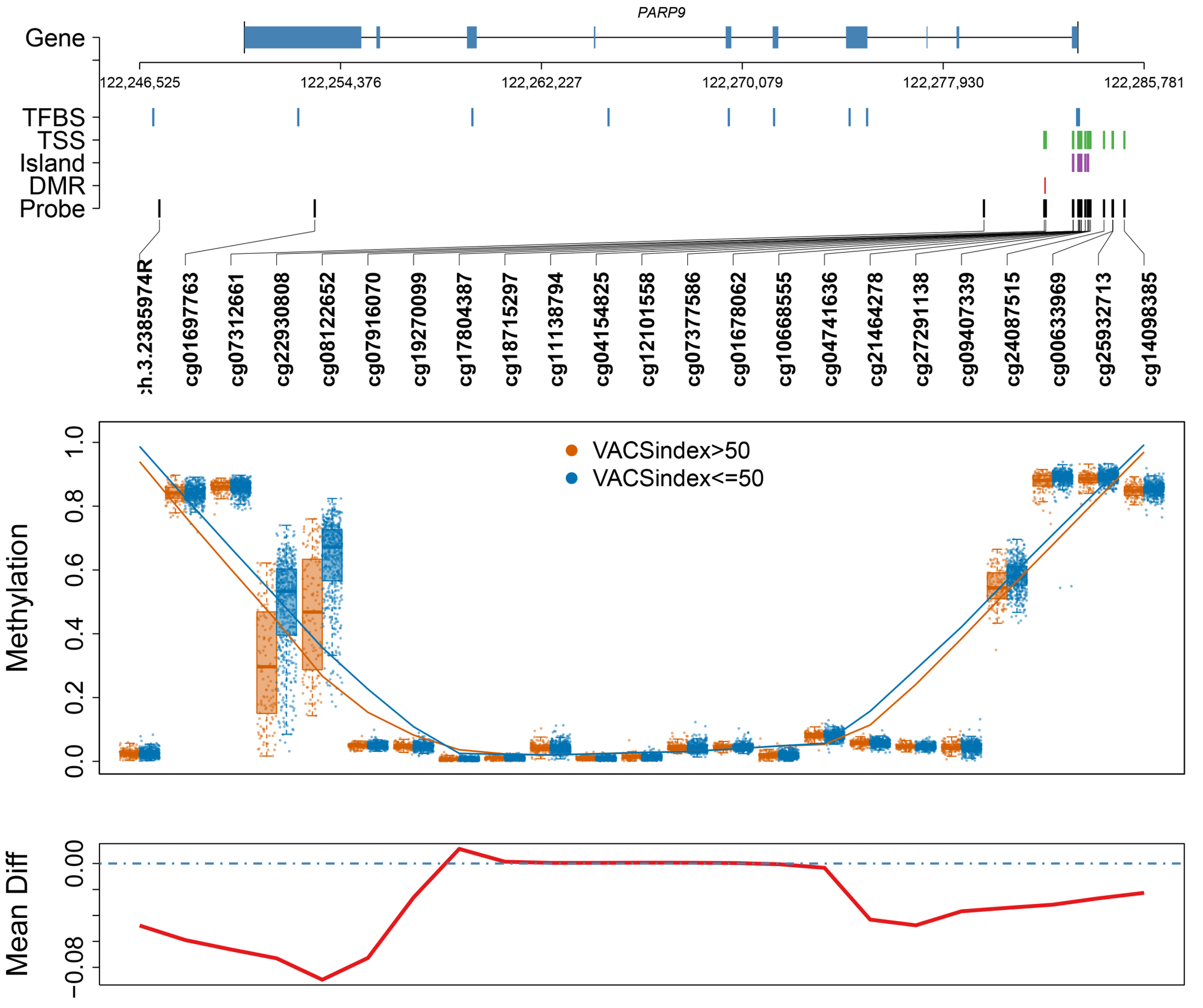
PARP9
Gene
122,246,525
122,254,376
122,262,227
122,270,079
122,277,930
122,285,781
TFBS
TSS
Island
DMR
Probe
ch.3.23859748R
cg01697763
cg07312661
cg22930808
cg08122652
cg07916070
cg19270099
cg17804387
cg18715297
cg11138794
cg04154825
cg12101558
cg07377586
cg01678062
cg10668555
cg04741636
cg21464278
cg27291138
cg09407339
cg24087515
cg00633969
cg25932713
cg14098385
VACSindex>50
VACSindex<=50
Methylation
0.0
0.2
0.4
0.6
0.8
1.0
Mean Diff
0.00
−0.08

Figure 3c

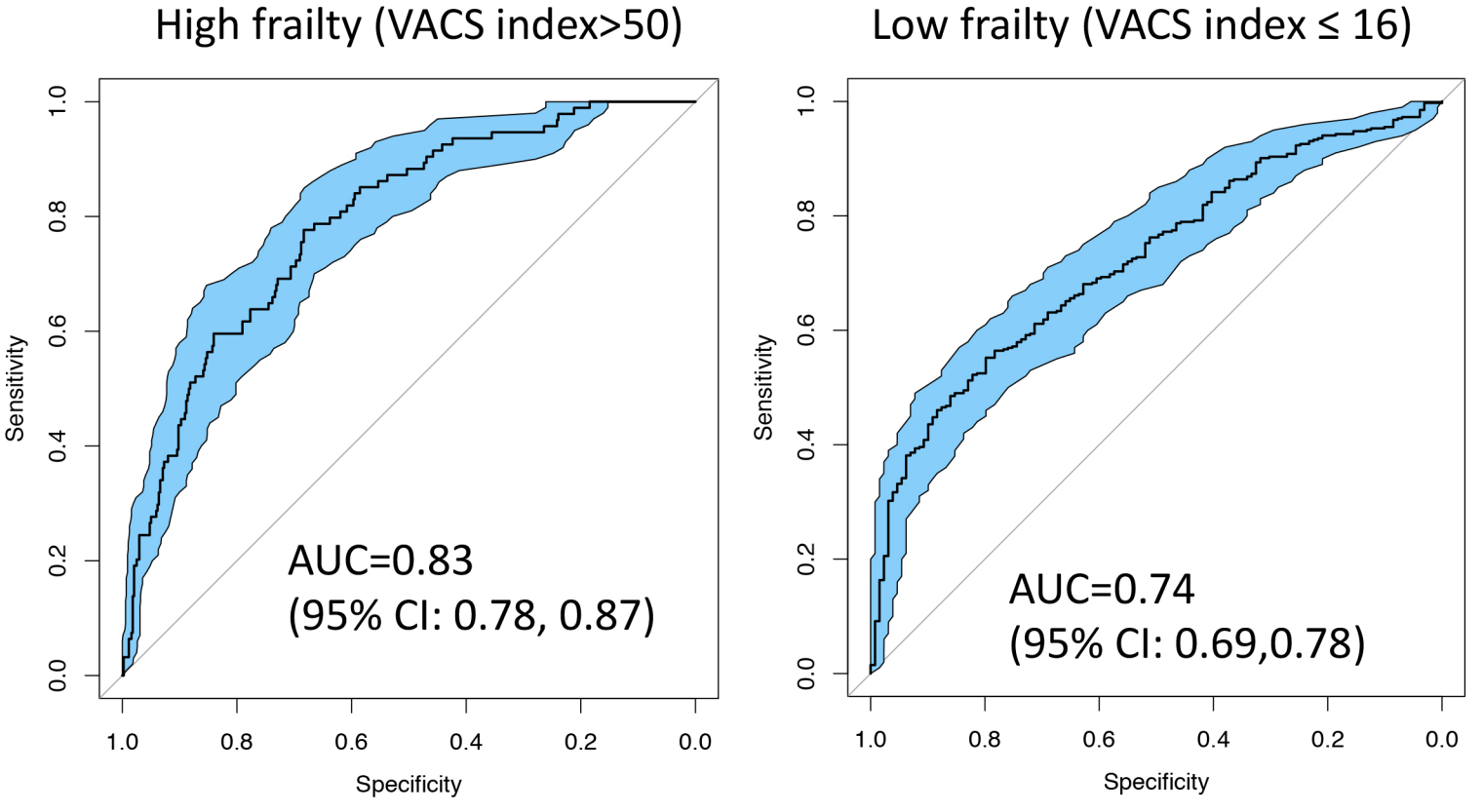
High frailty (VACS index>50)
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC=0.83
(95% CI: 0.78, 0.87)
Low frailty (VACS index ≤ 16)
AUC=0.74
(95% CI: 0.69,0.78)

Figure 4

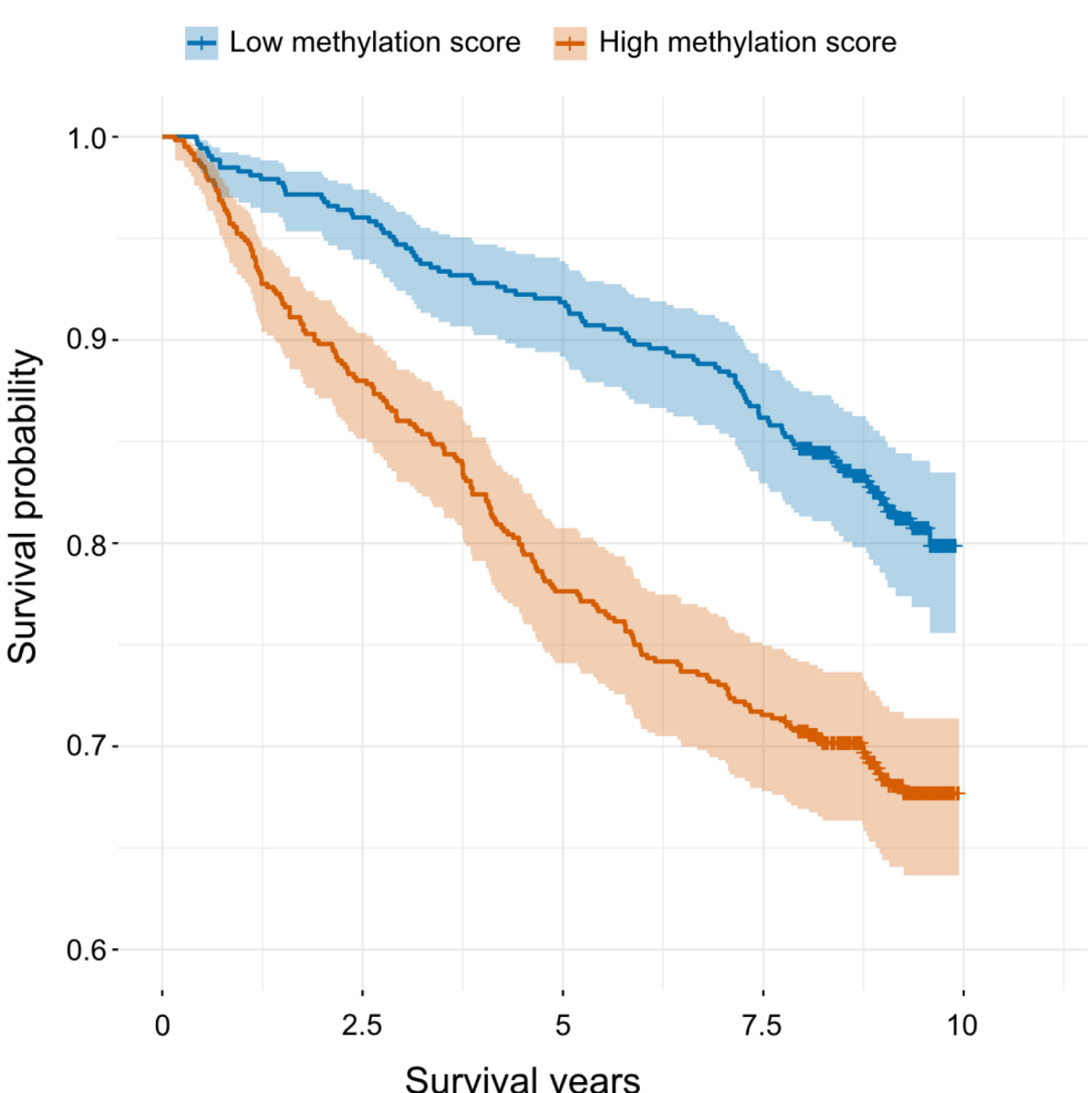
Low methylation score
High methylation score
1.0
0.9
0.8
0.7
0.6
Survival probability
0
2.5
5
7.5
10
Survival years

Figure 5

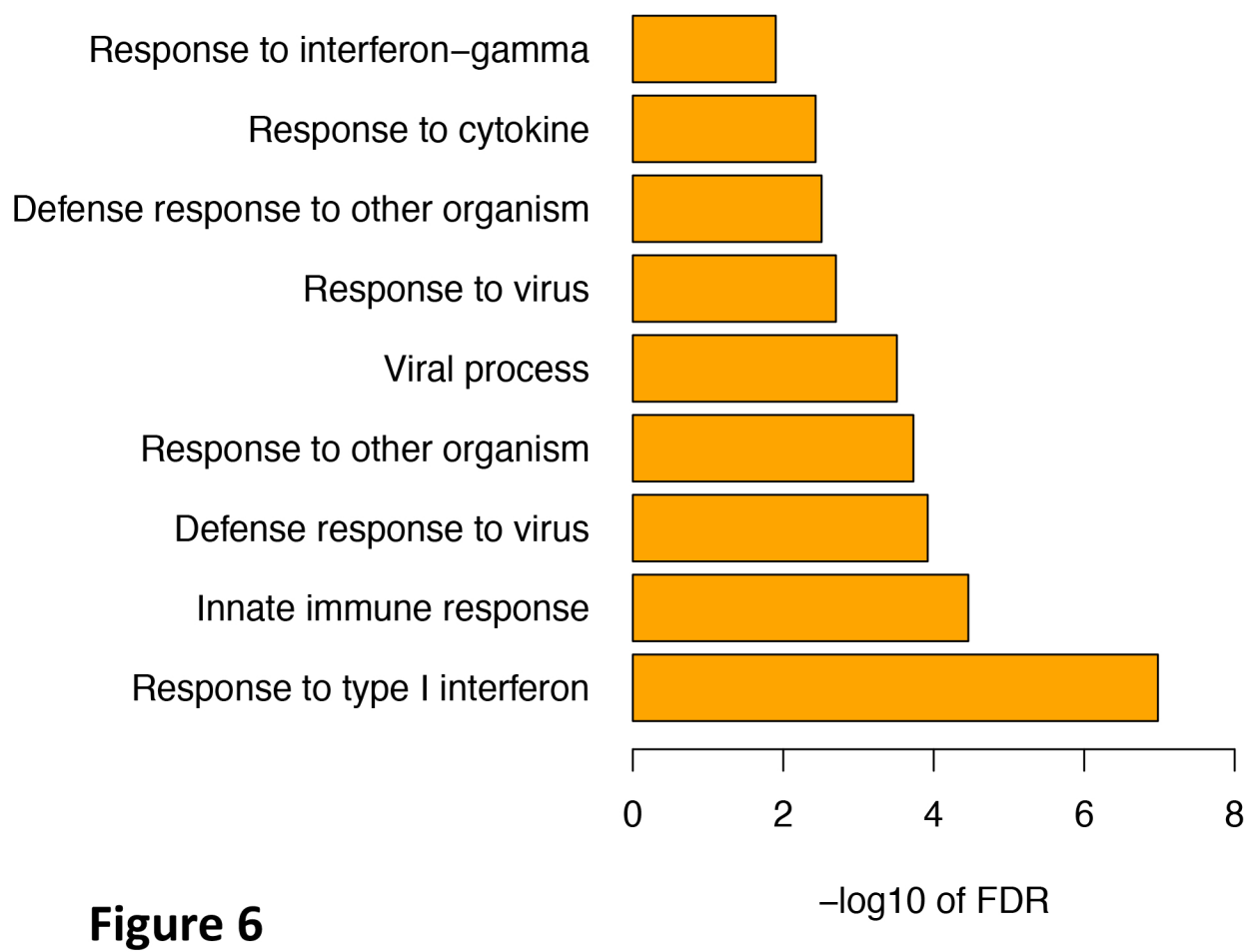
Response to interferon−gamma
Response to cytokine
Defense response to other organism
Response to virus
Viral process
Response to other organism
Defense response to virus
Innate immune response
Response to type I interferon
0
2
4
6
8
−log10 of FDR

Figure 6