Aug 20, 2019

DNA Methylation Signatures Predict HIV Prognosis and Mortality V.1

In 1 collection

Xinyu Zhang[1], Ying Hu[2], Ke Xu[1]

[1]Yale University, [2]National Cancer Institute

1 Works for me dx.doi.org/10.17504/protocols.io.smdec26

Xinyu Zhang

ABSTRACT

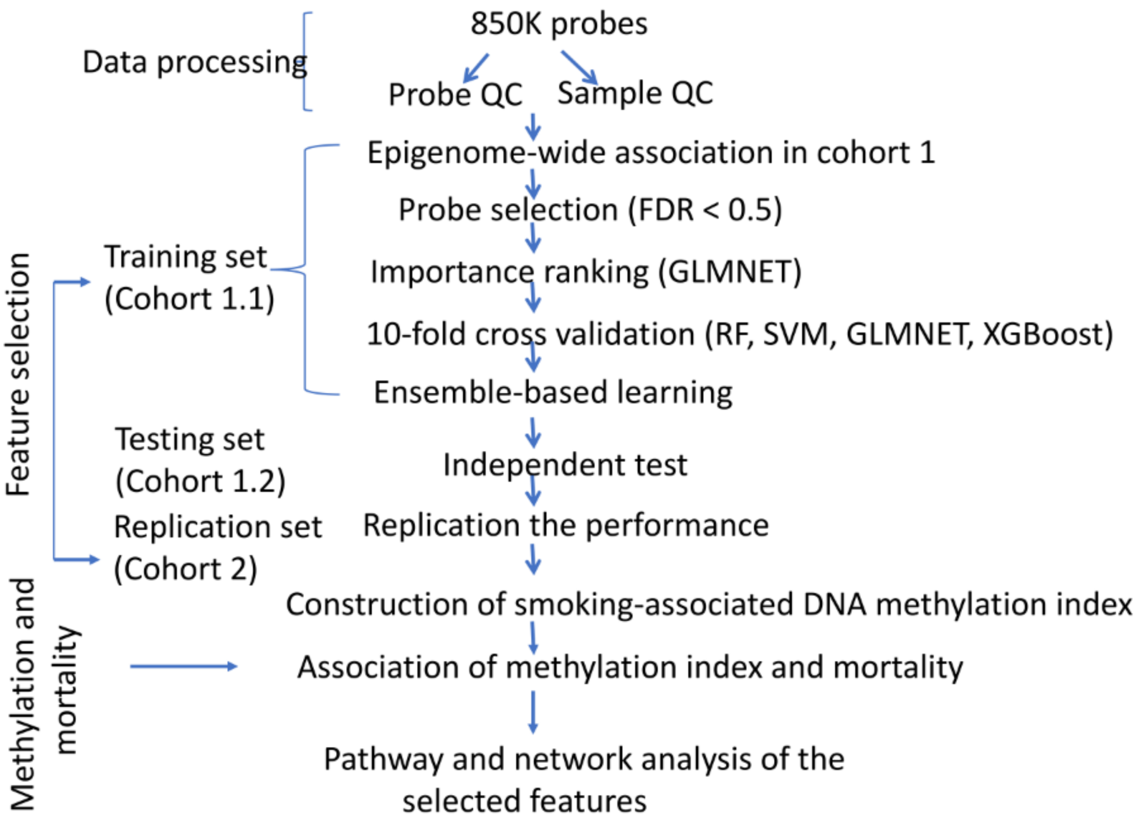
Background The effects of tobacco smoking upon epigenome-wide methylation signatures in white blood cells (WBCs) collected from persons living with HIV may have important implications for their immune-related outcomes, including frailty and mortality. The application of a machine learning approach to the analysis of CpG methylation in the epigenome enables the selection of phenotypically relevant features from high-dimensional data. Using this approach, we now report that a set of smoking-associated DNA-methylated CpGs predicts HIV prognosis and mortality in an HIV-positive veteran population.

Methods and Findings A total of 1,137 DNA samples were extracted from WBCs in HIV-positive male veterans. Here, HIV prognosis was measured by Veteran Aging Cohort Study Index that was highly correlated with HIV disease frailty. We first conducted two epigenome-wide association studies (EWAS) for tobacco smoking in two HIV-positive sample sets. We then selected smoking-associated DNA methylation features using a machine learning approach to differentiate good and poor HIV prognosis and to predict mortality. The performance of the prediction was validated in an independent sample set. We identified137epigenome-widesignificant CpGs in meta-EWAS analysis for smoking in HIV-positive samples ($p < 1.70E-07$). Using an in-house developed bioinformatic pipeline, we selected 698 CpG sites that predicted high HIV frailty [(Area Under Curve (AUC) = 0.73, (95%CI: 0.63 ~ 0.83)] and was replicated in an independent sample [(AUC = 0.78, 95%CI: 0.73 ~ 0.83)]. We further found an association of a DNA methylation index constructed from the 698-CpGs that were associated with a 5-year survival rate [HR =1.46; 95%CI: 1.06 ~ 2.02, p = 0.02]. Interestingly, the 698-CpGs located on 445 genes were enriched on the integrin signaling pathway (p=9.55E-05, False discovery rate=0.036), which is responsible for regulation of the cell cycle, differentiation, and adhesion.

Conclusion We demonstrated that smoking-associated DNA methylation features in white blood cells predict HIV infection prognosis and mortality in a population living with HIV. Feature selection is a powerful approach to identify DNA methylation biomarkers for the prediction of medical outcomes.

Workflow

1


Data processing
850K probes
Probe QC
Sample QC
Feature selection
Training set (Cohort 1.1)
Epigenome-wide association in cohort 1
Probe selection (FDR < 0.5)
Importance ranking (GLMNET)
10-fold cross validation (RF, SVM, GLMNET, XGBoost)
Ensemble-based learning
Testing set (Cohort 1.2)
Independent test
Replication set (Cohort 2)
Replication the performance
Methylation and mortality
Construction of smoking-associated DNA methylation index
Association of methylation index and mortality
Pathway and network analysis of the selected features

Epigenome-wide Association Analysis

2 Following protocols of https://www.nature.com/protocolexchange/protocols/6335

Meta-analysis

3 We conducted an EWAS meta-analysis by combining the data from the discovery and replication samples. Effect size and p-values for each probe were obtained from analyses in cohort 1 and cohort 2 samples, respectively. We performed fixed-effects, inverse-variance meta-analysis, with scheme parameters of sample size and standard error by implementing the METAL (ver: 2010-02-08) program, combining summary statistics in two sample sets. We investigated heterogeneity in two sample sets using the I2-statistic.

Machine Learning Selection of Smoking-association DNA Methylation for HIV Prognosis using an In-house Developed Protocol

4 We developed an in-house ensemble-based feature selection R package, *SmartFeatureSelction* (*https://bitbucket.org/starrcofly/smartfeatureselection.git*)to link smoking-associated CpG sites to HIV outcomes. *SmartFeatureSelection*used a supervised machine learning approach and merges EWAS with step-wise feature selection. The protocol included: 1) filtering of irrelevant probes based on methylation association analysis; 2) consideration of the relative attribution of each probe to rank the importance of the probes; 3) application of multiple machine learning methods to optimize the model; 4) building a model based on greedy ensemble by applying a weighted average of 4 machine learning models; 5) validation of the model in two independent samples.

Considering the samples were processed at different times and platforms, batch effects were removed using Function removeBatchEffect in R limma (ver. 3.32.10) library before performing machine learning prediction. To reduce redundant DNA methylation signals and noise for improving the prediction accuracy of HIV frailty, CpG sites with FDR < 0.5 from EWAS in cohort 1 were selected for machine learning. The samples in cohort 1 were randomly divided into a training set and a test setwith ratio of 8:2. We first built a model using the training set, in which each sample was labeled poor (VACS index >50) or good prognosis (VACS index <=50). We then tested the model by performing 10-fold cross validation in the testing set, and the best-performed model was tested in an independent replication set.

Association of Smoking-associated DNA Methylation Index and Mortality

5 To examine whether the select CpG site methylation was associated with mortality, we constructed a methylation index from the selected smoking-associated CpG sitesfollowing the previous formula[34]. A separate index was constructed for hypomethylated and hypermethylated CpG sites, respectively.

The association of the DNA methylation risk index with all-cause mortality was examined by Kaplan–Meier plots and log-rank tests in all samples. Cox regression model was then used to adjust for age, antiretroviral therapy, HIV-1 loads, and CD4 counts. In the Cox regression model, the DNA methylation index score was a categorical variable (using the highest quartiles as the reference category) or a continuous variable (calculating HR for a decrease in DNA methylation by one standard deviation). Index$_{hypo}$and index$_{hyper}$were evaluated for the prediction of mortality separately.

Gene Enrichment Analysis

6 Pathway and network analysis was conducted for the selected CpG sites on the nearest genes by employing Ingenuity Pathway Analysis (IPA). For genes with multiple CpG sites, the lowest p-value at the CpG site within a gene was used to represent the gene level significance. Significant pathways were defined at a FDR <0.05.