DNA methylation signatures analysis with Illumina Infinitum MethylationEPIC and Infinium Human Methylation 450K BeadChip

Xinyu Zhang (xinyu.zhang@yale.edu)
Ke's Lab, Yale University
Ke Xu (ke.xu@yale.edu)
Department of Psychiatry, Yale School of Medicine, 300 George Street, New Haven, CT 06511
Ying Hu
NCI
Amy Justice
VA Connecticut Healthcare System, 950 Campbell Ave, West Haven, CT 06516
Boyang Li
Department of Biostatistics, Yale School of Public Health, New Haven, CT, 06511
Zuoheng Wang
Department of Biostatistics, Yale School of Public Health, New Haven, CT, 06511
Hongyu Zhao
Department of Biostatistics, Yale School of Public Health, New Haven, CT, 06511
John Krystal
Department of Psychiatry, Yale School of Medicine, 300 George Street, New Haven, CT 06511

Method Article

Keywords: IDU HCV 450K EPIC signature CpG island

Posted Date: June 12th, 2018

DOI: https://doi.org/10.1038/protex.2018.080

Abstract

This protocol provides detailed instructions on Quality Control, EWAS, and CpG island signature discovery with Illumina Infinitum MethylationEPIC and Infinium Human Methylation 450K BeadChip data. R package minfi and caret was used in this analysis.

Introduction

This protocol provides a pipeline for a two-stage genome-wide DNA methylation analysis for HIV-infected IDU in African American male. All subjects were from a well-established longitudinal cohort, Veteran Aging Cohort Study \(VACS). DNA samples were isolated from whole blood. DNA methylation profile: Illumina HumanMethylation 450k Beadchip. The replication analysis was based on Illumina EPIC BeadChip. Analysis pipeline:
 Figure 1 Analysis workflow

Reagents

Computer with at least 2 GB of RAM \(Linux server required for large data set, but PC and Mac are also OK with the data set used in this protocol) High performance computer system with Linux system.

Equipment

R ver >= 3.3.3 minfi ver >= 1.18.1 caret ver >= 6.0-77

Procedure

1. Rawdata processing 1 a) download raw data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100264 1 b). Detection Pvalue determination A total of 416 probes on chromosome Y were selected. A set of p-values, 5e-2, 1e-2, 1e-5, 1e-8, 1e-10, 1e-12, … 1e-30, was used to calculate the call rate of each female subject using the 416 probes. A detection p-value was chosen considering ratio of non-0-call-rate subjects, and not too stringent to exclude too many signals. Finally, the intensity values with detection P ≥ 1e-12 were set to missing data.  Figure 2 450K

Pvalue determination  Figure 3 EPIC Pvalue determination 1 c) Data QC The data quality was

checked by using minfi R package. Probe QC: All the probes located in sex chromosome probes and within 10 bp SNP were removed. Sex chromosome: 11,648 probes on sex chromosomes Within 10 bp SNP: 36,142 probes within 10 bp SNPs. A total of 437,722 probes remained after removing the above probes. Sample QC: We determined the proportion of missing data per sample, enabling calculation of the sample call rate, and excluded samples with sample call rate < 98%. We also compared the predicted sex with the self-reported sex. One sex mismatched sample and three samples with lower call rate than 0.98 were found and excluded in the further analysis.  Figure 4 sex prediction results

 Figure 5 whole beta value distribution 

Figure 6 sample data quality check R code: ############################## qc <- getQC\(MSet) head\(qc) #plot QC plotQC\(qc) #check sex plotSex\(getSex\(gRatioSet.quantile, cutoff = -2)) ############################## 1 d) Intensity data processing and normalization Background correction and within array normalization was conducted with preprocessIllumina function in minfi package. The original Green/Red channel intensity data was also transformed to Meth/Unmeth data. All probes were divided to 3 types: Type I Green, Type I Red, and Type II. Meth/Unmeth data were subsequently divided into 6 groups: Meth and Unmeth for each type as above. Each group was normalized independently by using normalizeBetweenArrays function in Limma R package \(version 3.26.2), and then were merged. Then Beta value of each probe was generated using getBeta function in minfi. R code: ############################## #background correction & within-array normalization MSet.illumina <- preprocessIllumina\(RGSet, bg.correct = TRUE, normalize = "controls") #between-array normalization after dividing the signals into 6 parts meth.1 = normalizeBetweenArrays\(my.meth\[typeIProbesGrn,]) meth.2 = normalizeBetweenArrays\(my.meth\[typeIProbesRed,]) meth.3 = normalizeBetweenArrays\(my.meth\[typeIIProbes,]) unmeth.1 = normalizeBetweenArrays\(my.unmeth\[typeIProbesGrn,]) unmeth.2 = normalizeBetweenArrays\(my.unmeth\[typeIProbesRed,]) unmeth.3 =

normalizeBetweenArrays\(my.unmeth\[typeIIProbes,]) ############################## 2. EWAS The EWAS pipeline mainly contains four steps: Principal component analysis \(PCA) on intensities of positive control probes \(to remove batch effect) Blood cell type proportion was estimated \(to adjust cell type confounders) PCA on intermediary residuals \(to control global biological confounders) General linear model of IDUs 2 a) PCA on intensities of control probes PCA on the intensities of 237 positive control probes. Prior-adjustment: Top PCs were correlated with some technical biases. Post-adjustment: batch effect was significantly eliminated.  Figure 7 Pre and post PCA

adjustment on intensities of control probes 2 b) Blood cell type proportions estimation Six blood cell sub-populations were estimated using the approach described by Houseman et al. 600 probes highly correlated with blood cell types were applied to estimate each cell type composition. Minfi R package was used. R code: ############################## cellCounts <- estimateCellCounts\(RGSet) ##############################
 Figure 8 CD4 estimated vs. lab measured 

Figure 9 CD8 estimated vs. lab measured  Figure 10 pre and post cell type adjustment

2 c) PCA on intermediary residuals Regression model 1: adjusted for technical bias using Control probe PCA and some essential biological factors GLM model: Beta ~ Age + WBC_total + CD8T + CD4T + Gran + NK + Bcell + Mono + PC1-30ControlProbe 2 d) Final EWAS model Regression model 2: GLM model: Beta ~ ivdused + Age + logVL+ WBC_total + CD8T + CD4T + Gran + NK + Bcell + Mono + PC1-30ControlProbe + PC1-5residual PC1-30ControlProbe was used to control technical bias PC1-5residual was used to control other global confounders logVL: lg\(HIV Viral Load); to control HIV VL WBC_total, CD8T, CD4T, Gran, NK, Bcell, Mono: to control cell type 3 Clustering 3 a) hirarchical clustering R library heatmap3 was used. R code: ############################## result <- heatmap3\(dat, distfun = distfunc, hclustfun = hclustfunc, ColSideCut = ColSideCut, ColSideAnn = ColSideAnn, ColSideFun = function\(x) \{ showAnn\(x) }, ColSideWidth = ColSideWidth, ColSideColors = ColSideColors, RowAxisColors = RowAxisColors, breaks = breaks, l egendfun = legendfun, col = cols, verbose = T, scale = scale, showRowDendro = T, showColDendro = T, RowSideLabs = F, ColSideLabs = F)

##############################  Figure 11 hierarchical clustering 3 b) T-SNE

visualization R library Rtsne was used R code: ############################## train = dat library\(Rtsne) set.seed\(123456789) tsne <- Rtsne\(t\(train), dims = 2, perplexity = 50, verbose = TRUE, max_iter = 2000, theta = 0) ############################## 4 Genome-wide differential DNA methylation region \(DMR) analysis using bumphunter protocol. First define a regression model very similar with EWAS final model \(refre to 2 d), and then define gene clusters with R object gRatioSet. At last produce DMRs with bumphunter fucntions in minfi. R code: designMatrix_cov <- model.matrix\(~ logVL_new + adhmed+dcq1pot+dcq2coke+dcq3stim+dcq4opio+alcohol+ivdused+AGEBL+RACECOMG+WBC_new+CD8T+CD4T+Gran+NK+Bcell+Mono+Control_Probe_PC1 targets) R code: ############################## get_pos_chr_cluter_from_gRatio = function \(cc_gRatioSet, maxGap=200000) \{ annotation <- getAnnotation\(cc_gRatioSet); chr = annotation$chr; pos = annotation$pos; cl <- clusterMaker\(chr, pos, maxGap = maxGap) print\(table\(cl)) Indexes <- split\(seq_along\(cl), cl) ret = list\(chr=chr, pos=pos, cl=cl, Indexes = Indexes) } cluster_info = get_pos_chr_cluter_from_gRatio\(cc_gRatioSet, maxGap); dmrs_cov.cl <- bumphunter\(getBeta\(cov_gRatioSet), design = designMatrix_cov, chr=cluster_info$chr, pos=cluster_info$pos, cluster=cluster_info$cl, cutoff = NULL, pickCutoff=T, maxGap = maxGap, B=perm_num, type="Beta", smooth=F, nullMethod="bootstrap") ############################## 5 Machine learning analysis 5 a) Support vector machine \(SVM) algorithm and R package caret was used. The 386 samples in EWAS and 748 CpG sites were used as training set. A different set of 238 samples also from VACS was used as independent testing set. The training samples were divided into 5 equal number of subgroups with indexes of <16, 17–24, 25–34, 35–50, and >50. In the testing samples, high and low HIV frailty was defined as VACS index scores above 50 \(the upper 20% quantile in 386 training samples) and below 16 \(lower 20% quantile). We then tested prediction performance on subjects with high \(VACS index >50 vs. ≤50) and low HIV frailty \(VACS index <16 vs. ≥16), respectively. 5 b) 1000 times permutation test was conducted to study that if the 748 probes has a significant performance than random ones. AUROC was used in the analysis.  Figure 12 independent

test, TP \(true positive), FP \(false positive), TN \(true negative), FN \(false negative) distribution 6 Pathway enrichment analysis IPA was used on pathway

enrichment analysis of DMP and DMR genes  Figure 13 Pathway enrichment analysis

Anticipated Results

The pipeline should produce the following results: 1) QC figures and normalized raw methylation data 2) EWAS results and significant probes 3) Machine learning results with using data of the significant CpG sites

References

Lehne B, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol 16, 37 \(2015). Houseman EA, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13, 86 \(2012). Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 15, R31 \(2014).