**Analysis code EWAS in smoking discordant MZ twin pairs**

**######STEP 1 RESIDUALIZE METHYLATION DATA ############**

#==============================================================================================================================================================================================#

# \* script uses 2 data objects: 1=methylation data. 2=Phenotype/covariate data.

# \* methylation data is on the beta-value scale (0-1). Bad probes and bad samples have been removed. Data have been normalized.

# \* methylation data contains numeric values. Missing values are denoted by NA.

# \* methylation data has the following format: rows=samples, columns= methylation sites. The rownames are the sample identifiers.

# \* methylation data has the following column names: Illumina probe ID (e.g. "cg1234567")

# \* missing values in your phenotype and covariates are denoted by NA.

# \* phenotype/covariate data has the following format: rows=individuals, columns=variables (covariates/phenotype). The rownames are the sample identifiers.

# \* in code below, the methylation data object (1) is called "MethBetavalues" and is stored in Blood.450k.Betas.RData, the phenotype/covariate data object (2) is called "dat" and is stored in Covariates.RData

#==============================================================================================================================================================================================#

rm(list = ls(all = TRUE))

options(stringsAsFactors = FALSE)

# load data

load("Blood.450k.Betas.RData")

load("Covariates.RData")

#put objects in same order

dat <- dat[rownames(MethBetavalues),]

#extract necessary covariates

Sample\_Plate <- as.factor(dat$Sample\_Plate)

Array\_rownum <- as.numeric(dat$Array\_rownum)

Neut\_Perc <- as.numeric(dat$Neut\_Perc)

Mono\_Perc <- as.numeric(dat$Mono\_Perc)

Eos\_Perc <- as.numeric(dat$Eos\_Perc)

CpGi <- rep(NA,times=dim(dat)[1])

data <- data.frame(Sample\_Plate,Array\_rownumNeut\_Perc,Mono\_Perc,Eos\_Perc, CpGi)

rownames(data) <- rownames(dat)

# select samples without missing data

select <- which(!is.na(data$Mono\_Perc) & !is.na(data$Eos\_Perc) & !is.na(data$Neut\_Perc))

data <- data[select,])

# Linear model

LMJ <- function (data) { lm(CpGi ~ Mono\_Perc+Eos\_Perc+Neut\_Perc+Array\_rownum+Sample\_Plate, na.action=na.omit,data=data)}

# check if objects are in the same order

table(rownames(data)==rownames(MethBetavalues))

#create output object “residuals”

Ncpgs<-ncol(MethBetavalues)

Ncpgs

Nsubjects <- nrow(data)

Nsubjects

residuals <- matrix(NA,Ncpgs,Nsubjects)

colnames(residuals) <- rownames(data)

rownames(residuals) <- colnames(MethBetavalues)

# Run linear model for all CpGs and save residuals

for (i in 1:Ncpgs)

{

data$CpGi <- MethBetavalues[rownames(data),i]

linearmodel <- LMJ(data)

for (k in 1:length(linearmodel$residuals))

{ residuals[i,names(linearmodel$residuals)[k]] <- linearmodel$residuals[k]

}

}

**######STEP 2 Paired T-test #########**

# script to compare DNA methylation within smoking discordant pairs (twin 1= never smoker, twin 2=current smoker)

# \* script uses 2 data objects: 1= M\_nonsmokingtwin=Methylation data residuals (residualized for covariates) of the non-smoking twin. 2= M\_smokingtwin= Methylation data residuals (residualized for covariates) of the current smoking co-twins. Rows=samples, Columns=CpGs. Column names=CpG identifiers. The 2 objects are in the same order (that is, ordered by twin pairs, i.e. row 1=twin pair 1, row 2=twin pair 2 etc.).

NCpGs <- colnames(M\_nonsmokingtwin)

# create object to store output

results <- matrix(nrow= NCpGs, ncol=5)

for (i in 1: NCpGs)

{

res <- t.test(M\_nonsmokingtwin [,i], M\_smokingtwin [,i], paired=T)

results[i,1] <- res$estimate

results[i,2] <- res$p.value

results[i,3] <- res$conf.int[1]

results[i,4] <- res$conf.int[2]

results[i,5] <- res$statistic

}

results <- data.frame(colnames(NCpGs),results)

colnames(results) <- c("cgid","MeanDifference","pvalue","95confint\_L","95confint\_H","statistic")