The following codes were written by Tianye Jia and Sylvane Desrivières. If you have any questions related to these analyses, please contact us at enigma.epihelpdesk@gmail.com.

**Files required for these analyses**

**Methylation data** – The following data, generated by our ENIGMA-Epigenetics QC protocol will be used:

"./Quan-norm.rda" (Beta values after quantile normalization)

"./fast\_svd.rda" (Principle components of beta values)

"./cellcount.rda" (Estimated cell type proportion)

"./RGset.rda" (Raw Dataset after QC)

**Subcortical volumes** ­– These should be the **LandRvolumes.csv** file containing the subcortical brain volumes (after quality control) used for ENIGMA2. Please, provide **brain volumes extracted with FSL** if possible. The **LandRvolumes.csv** file should look like this:



**Covariates file**–The same files used for ENIGMA2 will also be used. This comma delimited (.csv) text file should contain the following columns: SubjID, Age, Sex. Additional columns for dummy covariates (i.e. a covariate to control for different acquisitions sites, if applicable), etc is optional. In addition, the first 4 principal components of the beta value, i.e. from "./fast\_svd.rda", and the first two components of estimated cell-type proportion, i.e. from "./cellcount.rda", will also be included as control variables. The relevant code of combining data will be provided.

* If your cohort has only healthy controls, only patients and twin study, this file should be named **SubCortCovs.csv** and look like this:



* For your cohort has both patients and healthy controls, you should include a covariate called "AffectionStatus", coded as a binary indicator variable where Controls = 0 and Patients = 1. The final file, saved as **SubCortCovs.csv**, should have the following columns at a minimum: SubjID, Age, Sex, AffectionStatus. Additional columns for dummy covariates (i.e. a covariate to control for different acquisitions sites, if applicable) is optional. Note we will have three outputs for the whole sample, the case individuals only and the control individuals only.
* **Note 1:** !! Please make sure that missing data has been recoded as NA in these files!! Self-coded missing data should be transformed to NA through Data== -9 <- NA, where Data should be replaced with the name of your data file in question and -9 should be replaced with your self-coded missing data value.
* **Note 2:** Sex must be specified as follows: (Males=1, Females=2), and "FID" and "IID" should be named exactly the same in all files.
* **Note 3:** If population stratification is of reasonable concern, the self-reported ethnicity information should be included in the covariates file, i.e. as dummy variables. As an alternative, the first 4 MDS of genetic data can also be included.

##################

#####Preparing files#

##################

#Provide cohort name

cohort = "IMAGEN" #change for the name of your cohort

###Probe Quality Check####

library(minfi)

library(minfiData)

load("./Quan-norm.rda") #variable object will be loaded

load("./RGset.rda") #variable RGset will be loaded

#Adding SNP info to the data

objectWithSNPinfo <- addSnpInfo(object)

#Dropping probes that contain either a SNP at the CpG interrogation or at the single nucleotide extension

objectSNPQCed<- dropLociWithSnps(objectWithSNPinfo, snps=c("SBE","CpG", "Probe"), maf=0.05)

detP <- detectionP(RGset)

Match1 <- match(colnames(objectSNPQCed),colnames(detP))

Match2 <- match(rownames(objectSNPQCed),rownames(detP))

detPSNPQCed <- detP[Match2[!is.na(Match2)],Match1[!is.na(Match1)]]

failed <- detPSNPQCed >0.01

beta <- getBeta(objectSNPQCed)

###Drop probes that failed quality control via the detection p-value in greater than 20% of samples

failedCG02 <- rowMeans(failed)>0.2

####Get the list non-variable CpG sites i.e. those where beta values for all samples are ≤20% or ≥80%

ProbeInvar <- (rowSums(beta<=0.2)==ncol(beta))|(rowSums(beta>=0.8)==ncol(beta)) #Mark probes with either all beta value <=0.2 or all beta value>=0.80

ListInvarProbe <- rownames(beta)[which(ProbeInvar)] #Generate a list of probes marked as invariant.

#This list will be included in the output file that you will send us

###Remove sex chromosome probes#####

keepIndex=!seqnames(objectSNPQCed)%in%c("chrX","chrY") #mark probes on X and Y chromosome

keepIndex <- keepIndex&(!failedCG02) #Combine with failed probes

beta[failed] <- NA #Remove all probes with detected P-value >0.01.

betaQC <- beta[which(keepIndex),] #Remove probes

##Reformat beta value to match the format suggested above

Methy <- as.data.frame(t(betaQC)) #load quantile normalized methylation data

Methy$Subject <- colnames(betaQC) #add subject ID to the methylation data as the last column

Methy <- Methy[,c(ncol(Methy),1:(ncol(Methy)-1))] #re-order the column and making the subject ID as the first column

MethyName <- colnames(Methy)[-c(1)] #output probe names

load("./fast\_svd.rda") #load principal components of beta value

PC\_Beta <- as.data.frame(ss$v[,1:4]) #generate a variable with first four principle components of methylation

PC\_Beta$Subject <- Methy$Subject #add subject ID to the component variable

load("./cellcount.rda") #load estimated cell type proportion

tmp <- prcomp(cellcount) #generate principle components of cell count

pc\_cell <- as.data.frame(tmp$x[,1:2]) #Generate a variable with subject ID and first 2 PCs of cellcount

pc\_cell$Subject <- rownames(pc\_cell) #Add row names

###Read and format Subcortical volumes ####

Raw\_Struc <- read.csv("./LandRvolumes.csv", header=T) #Please ensure the format of the ‘LandRvolumes.csv’ file is as illustrated above.

Struc <- matrix(data=0, ncol=8, nrow=nrow(Raw\_Struc)) #Generate a new matrix for the mean of L/R subcortical brain regions

colnames(Struc) <- c("Subject","Mthal”, "Mcaud", "Mput", "Mpal", "Mhippo", "Mamyg", "Maccumb") #Provide column names for the mean subcortical brain volumes

Struc <- as.data.frame(Struc) #set data as data.frame

Struc$Subject = Raw\_Struc$SubjID #assign subject ID

Struc$Mthal <- rowMeans(Raw\_Struc[,c("Lthal","Rthal")]); #calculate mean Thalamus volume

Struc$Mcaud <- rowMeans(Raw\_Struc[,c("Lcaud","Rcaud")]); #calculate mean Caudate volume

Struc$Mput <- rowMeans(Raw\_Struc[,c("Lput","Rput")]); #calculate mean Putamen volume

Struc$Mpal <- rowMeans(Raw\_Struc[,c("Lpal","Rpal")]); #calculate mean Pallidum volume

Struc$Mhippo <- rowMeans(Raw\_Struc[,c("Lhippo","Rhippo")]); #calculate mean volume Hippocampus

Struc$Mamyg <- rowMeans(Raw\_Struc[,c("Lamyg","Ramyg")]); #calculate mean Amygdala volume

Struc$Maccumb <- rowMeans(Raw\_Struc[,c("Laccumb","Raccumb)]); #calculate mean Nucleus Accumbens volume

StrucName <- colnames(Struc)[-1] #column names of subcortical brain volumes

ICV <- Raw\_Struc[,c(1,ncol(Raw\_Struc))] #generate a separate variable for ICV

colnames(ICV)[1] <- "Subject" #rename the column for easier merging

###Read and Combine Covariates Data#####

Raw\_Cov <- read.csv("./**SubCortCovs.csv**", header=T) #load covariates file as illustrated above, please remember to include the extra column of ‘AffectStatus’ for case/control sample.

Raw\_Cov<-Raw\_Cov[,-c(1)] #Assume it is the IID that is equal to the Subject ID from Structure and Methylation data

Raw\_Cov$Age\_Square <- (Raw\_Cov$Age)^2 #Add age^2 as the last column of Raw\_Cov

colnames(Raw\_Cov)[1] <- "Subject"

Cov <- merge(Raw\_Cov, PC\_Beta, by="Subject", all=F) #combine covariates and PCs of methylation data

Cov <- merge(Cov, pc\_cell, by="Subject", all=F) #combine covariates and PCs of cell count

Cov < merge(Cov, ICV, by="Subject", all=F) #combine covariates and ICV

CovName <- colnames(Cov)[-c(1)] #names of covariates

###Generating Files with Complete Data across All Data#####

Data <- merge(Cov, Struc, by="Subject", all=F) #combine covariates and structure data

Match <- match(Methy$Subject, Data$Subject) #match methylation sample with combined covariates and structure sample

## generate new covariates, structure and methylation data, which have had their individuals matched

Cov <- Data[Match[!is.na(Match)],2:(length(CovName)+1)] #generate covariates with matched individual

Struc <- Data[Match[!is.na(Match)],-c(1:(length(CovName)+1))] #generate structure data with

matched individual

Methy <- Methy[!is.na(Match),-c(1)] #generate methylation data with matched individual

#####Linear regression Analysis######

Num\_Methy <- ncol(Methy)

Num\_Cov <- ncol(Cov)

Num\_Struc <- ncol(Struc)

##Preparing the output files

Origin\_Beta <- matrix( data= NA, nrow=Num\_Methy, ncol= Num\_Struc, byrow=F, dimnames=NULL)

colnames(Origin\_Beta) <- colnames(Struc)

rownames(Origin\_Beta) <- colnames(Methy)

Origin\_SD <- matrix( data= NA, nrow=Num\_Methy, ncol= Num\_Struc, byrow=F, dimnames=NULL)

colnames(Origin\_SD) <- colnames(Struc)

rownames(Origin\_SD) <- colnames(Methy)

Origin\_P <- matrix( data= NA, nrow=Num\_Methy, ncol= Num\_Struc, byrow=F, dimnames=NULL)

colnames(Origin\_P) <- colnames(Struc)

rownames(Origin\_P) <- colnames(Methy)

##################################################################

**##Association analysis: the following section should be run by ALL cohorts except for Twin Studies ####**

###A different code based on GEE for Twin Study will be circulated when it is ready#####

**##################################################################**

##To minimise workload at each site, the following codes are written in such a way that association analyses with all 7 subcortical brain volumes will be performed, but only the outputs from the association analyses with 3 brain volumes (Thalamus, hippocampus and nucleus accumbens) will be selected for our initial meta-analyses.

Covar <- matrix(unlist(Cov), ncol=ncol(Cov), byrow=F)

for (i in 1:Num\_Struc) {

 for (j in 1:Num\_Methy) {

 Out <- summary(lm(Struc[,i]~Covar+Methy[,j]))

 Origin\_Beta[j,i] <- Out$coefficients[nrow(Out$coefficients),1]

 Origin\_SD[j,i] <- Out$coefficients[nrow(Out$coefficients),2]

 Origin\_P[j,i] <- Out$coefficients[nrow(Out$coefficients),4]

 }

}

###Save output file

save(Origin\_Beta, Origin\_SD, Origin\_P, ListInvarProbe, file=paste("./Output\_of\_",cohort, "\_Methylation\_and\_All\_Subcortical\_Structure.RData",sep=""),compress=T)

#This file should be saved for further reference (contains output for all subcortical volumes), please don’t send it to us

Select\_Beta <- Origin\_Beta[,c(1,5,7)] #column 1,5,7 are corresponding to ‘Mthal’, ‘Mhippo’ and ‘Maccumb’.

Select\_SD <- Origin\_SD[,c(1,5,7)]

Select\_P <- Origin\_P[,c(1,5,7)]

save(Select\_Beta, Select\_SD, Select\_P, ListInvarProbe, file=paste("./**Output\_of\_",cohort, "\_Methylation\_and\_Selected\_Subcortical\_Structure.RData**",sep=""),compress=T)

#We will need this file with selected beta, SD and P values for the Thalamus, Hippocampus and Nucleus Accumbens, as well as a list of invariant probes.

###########################################################################

##**Cohorts with patients should also run the following section, in addition to the above #**#

###########################################################################

##The script below will perform associations for cases and controls separately.

Affect **<-** colnames**(**Cov**)==**"AffectionStatus"

**if** **(**sum**(**Affect**)==**1**)** **{** #Check if there is a column named "AffectionStatus", i.e. a case-control study

 Status **<-** c**(**"Case","Control"**)**

 **for** **(**k **in** 1**:**2**)** **{**

 Ind **<-** Covar**[**,Affect**]==(**2**-**k**)** #AffectionStatus should be coded as 1 (Case) and 0 (Control)

 **for** **(**i **in** 1**:**Num\_Methy**)** **{**

 **for** **(**j **in** 1**:**Num\_Struc**)** **{**

 Out **<-** summary**(**lm**(**Methy**[**Ind,i**]~**Covar**[**Ind,Affect**]+**Struc**[**Ind,j**]))**

 Origin\_Beta**[**i,j**]** **<-** Out**$**coefficients**[**nrow**(**Out**$**coefficients**)**,1**]**

 Origin\_SD**[**i,j**]** **<-** Out**$**coefficients**[**nrow**(**Out**$**coefficients**)**,2**]**

 Origin\_P**[**i,j**]** **<-** Out**$**coefficients**[**nrow**(**Out**$**coefficients**)**,4**]**

 **}**

 **}**

 ###Save output file

 save**(**Origin\_Beta, Origin\_SD, Origin\_P, ListInvarProbe, file**=**paste**(**"./Output\_of\_",cohort,"\_",Status**[**k**]**,"Individuals\_Methylation\_and\_All\_Subcortical\_Structure.RData",sep**=**""**)**,compress**=**T**)**

 #This file should be saved for further reference contains output for all subcortical volumes), please don’t send it to us

 Select\_Beta **<-** Origin\_Beta**[**,c**(**1,5,7**)]**

 Select\_SD **<-** Origin\_SD**[**,c**(**1,5,7**)]**

 Select\_P **<-** Origin\_P**[**,c**(**1,5,7**)]**

 save**(**Select\_Beta, Select\_SD, Select\_P, ListInvarProbe, file**=**paste**(**"./**Output\_of\_",cohort,"\_",Status[k],"Individuals\_Methylation\_and\_Selected\_Subcortical\_Structure.RData**",sep**=**""**)**,compress**=**T**)**

 #We will need this file with selected beta, SD and P values for the Thalamus, Hippocampus and Nucleus Accumbens, as well as a list of invariant probes

 **}**

**}**

#**For case control studies, all three RData files** (i.e., corresponding to selected outputs from the full sample, cases only and controls only) for the selected subcortical structure should be sent to us. **For all other cohorts, one RData file**, corresponding to selected outputs from the full sample, should be sent to us.

#**Other Information Required**: Apart from the relevant output files, we will also need information of **sample size** and **covariates included**.

#Please contact enigma.epihelpdesk@gmail.com to obtain upload information for your group’s data.