**Prospective cohort study in the UK Biobank**

setwd("D:\\ 1\_data and codes")

library(data.table)

load("UKB\_CD\_incident\_time.Rdata") #total=482596,cases=1770#

head(UKB\_CD\_incident\_time)

table(UKB\_CD\_incident\_time$site\_new)

#Cox Regression#

#somking1#

table(UKB\_CD\_incident\_time$f.20116.0.0)

UKB\_CD\_incident\_time$smoking1=ifelse(UKB\_CD\_incident\_time$f.20116.0.0 == -3, NA,

ifelse(UKB\_CD\_incident\_time$f.20116.0.0 == 0, 0, 1))

table(UKB\_CD\_incident\_time$smoking1)

#smoking#

a=c("0","1","2")

UKB\_CD\_incident\_time$smoking=ifelse(UKB\_CD\_incident\_time$f.20116.0.0%in%a, UKB\_CD\_incident\_time$f.20116.0.0, NA)

table(UKB\_CD\_incident\_time$smoking)

#age stopped smoking#

table(UKB\_CD\_incident\_time$f.2897.0.0)

UKB\_CD\_incident\_time$age\_stop\_smk <- factor(ifelse(UKB\_CD\_incident\_time$f.2897.0.0 <= 30, 1,

ifelse(UKB\_CD\_incident\_time$f.2897.0.0 > 30 & UKB\_CD\_incident\_time$f.2897.0.0<=40, 2,

ifelse(UKB\_CD\_incident\_time$f.2897.0.0 > 40 & UKB\_CD\_incident\_time$f.2897.0.0<= 50, 3,

ifelse(UKB\_CD\_incident\_time$f.2897.0.0 > 50, 4, 0)))))

table(UKB\_CD\_incident\_time$age\_stop\_smk)

#pack years of smoking)

table(UKB\_CD\_incident\_time$f.20161.0.0)

UKB\_CD\_incident\_time$pack\_year <- factor(ifelse(UKB\_CD\_incident\_time$f.20161.0.0 <= 20, 1,

ifelse(UKB\_CD\_incident\_time$f.20161.0.0 > 20 & UKB\_CD\_incident\_time$f.20161.0.0 <= 40, 2,

ifelse(UKB\_CD\_incident\_time$f.20161.0.0 > 40 & UKB\_CD\_incident\_time$f.20161.0.0 <= 60, 3,

ifelse(UKB\_CD\_incident\_time$f.20161.0.0 > 60, 4, 0)))))

table(UKB\_CD\_incident\_time$pack\_year)

#alcohol drinking#

UKB\_CD\_incident\_time$f.1568.0.0=as.numeric(UKB\_CD\_incident\_time$f.1568.0.0)

UKB\_CD\_incident\_time$f.1578.0.0=as.numeric(UKB\_CD\_incident\_time$f.1578.0.0)

UKB\_CD\_incident\_time$f.1588.0.0=as.numeric(UKB\_CD\_incident\_time$f.1588.0.0)

UKB\_CD\_incident\_time$f.1598.0.0=as.numeric(UKB\_CD\_incident\_time$f.1598.0.0)

UKB\_CD\_incident\_time$f.1608.0.0=as.numeric(UKB\_CD\_incident\_time$f.1608.0.0)

UKB\_CD\_incident\_time$red=UKB\_CD\_incident\_time$f.1568.0.0\*125\*0.135

UKB\_CD\_incident\_time$white=UKB\_CD\_incident\_time$f.1578.0.0\*125\*0.12

UKB\_CD\_incident\_time$beer=UKB\_CD\_incident\_time$f.1588.0.0\*570\*0.055

UKB\_CD\_incident\_time$spirits=UKB\_CD\_incident\_time$f.1598.0.0\*30\*0.415

UKB\_CD\_incident\_time$fortified=UKB\_CD\_incident\_time$f.1608.0.0\*58\*0.19

UKB\_CD\_incident\_time$alcohol\_g\_day = (UKB\_CD\_incident\_time$f.1568.0.0\*125\*0.135+UKB\_CD\_incident\_time$f.1578.0.0\*125\*0.12 +

UKB\_CD\_incident\_time$f.1588.0.0\*570\*0.055+UKB\_CD\_incident\_time$f.1598.0.0\*30\*0.415+

UKB\_CD\_incident\_time$f.1608.0.0\*58\*0.19)\*0.789/7

UKB\_CD\_incident\_time$alcohol\_10g\_day=UKB\_CD\_incident\_time$alcohol\_g\_day/10

UKB\_CD\_incident\_time$drinking=cut(UKB\_CD\_incident\_time$alcohol\_g\_day, breaks=c(0, 12.5, 50, Inf),

labels = c(0, 1, 2),

na.rm = TRUE)

#Physical activity#

UKB\_CD\_incident\_time$f.884.0.0=as.numeric(UKB\_CD\_incident\_time$f.884.0.0)

UKB\_CD\_incident\_time$f.894.0.0=as.numeric(UKB\_CD\_incident\_time$f.894.0.0)

UKB\_CD\_incident\_time$f.904.0.0=as.numeric(UKB\_CD\_incident\_time$f.904.0.0)

UKB\_CD\_incident\_time$f.914.0.0=as.numeric(UKB\_CD\_incident\_time$f.914.0.0)

UKB\_CD\_incident\_time$pp=ifelse((UKB\_CD\_incident\_time$f.884.0.0>=0|UKB\_CD\_incident\_time$f.894.0.0>=0|UKB\_CD\_incident\_time$f.904.0.0>=0|UKB\_CD\_incident\_time$f.914.0.0>=0), 0,NA)

UKB\_CD\_incident\_time$PA[UKB\_CD\_incident\_time$pp==0]="0"

UKB\_CD\_incident\_time$PA[UKB\_CD\_incident\_time$f.884.0.0>=5|UKB\_CD\_incident\_time$f.904.0.0>=1|(UKB\_CD\_incident\_time$f.894.0.0>=150&UKB\_CD\_incident\_time$f.894.0.0<1440)|(UKB\_CD\_incident\_time$f.914.0.0>=75&UKB\_CD\_incident\_time$f.914.0.0<1440)]="1"

#BMI

UKB\_CD\_incident\_time$BMI <- ifelse(UKB\_CD\_incident\_time$f.21001.0.0 < 18.5,1,

ifelse(UKB\_CD\_incident\_time$f.21001.0.0 >= 18.5&UKB\_CD\_incident\_time$f.21001.0.0 < 25.0, 2,

ifelse(UKB\_CD\_incident\_time$f.21001.0.0 > 25.0&UKB\_CD\_incident\_time$f.21001.0.0 < 30.0, 3,

ifelse(UKB\_CD\_incident\_time$f.21001.0.0 > 30.0,4,0))))

table(UKB\_CD\_incident\_time$BMI)

#Processed meat#

UKB\_CD\_incident\_time$processed.meat=as.numeric(UKB\_CD\_incident\_time$f.1349.0.0)

UKB\_CD\_incident\_time$processed.meat[UKB\_CD\_incident\_time$processed.meat<0]=NA

UKB\_CD\_incident\_time$processed.meat[UKB\_CD\_incident\_time$processed.meat>0&UKB\_CD\_incident\_time$processed.meat<=1]="0"

UKB\_CD\_incident\_time$processed.meat[UKB\_CD\_incident\_time$processed.meat>1]="1"

#Enviromental factors only#

UKB\_CD\_incident\_time$f.21001.0.0=as.numeric(UKB\_CD\_incident\_time$f.21001.0.0)

UKB\_CD\_incident\_time$f.48.0.0=as.numeric(UKB\_CD\_incident\_time$f.48.0.0)

UKB\_CD\_incident\_time$f.50.0.0=as.numeric(UKB\_CD\_incident\_time$f.50.0.0)

UKB\_CD\_incident\_time$f.30710.0.0=as.numeric(UKB\_CD\_incident\_time$f.30710.0.0)

UKB\_CD\_incident\_time$CRP\_10=UKB\_CD\_incident\_time$f.30710.0.0/10

table(UKB\_CD\_incident\_time$f.30710.0.0)

#crude model#

library("survival")

UKB\_CD\_incident\_time$f.21022.0.0=as.numeric(UKB\_CD\_incident\_time$f.21022.0.0)

UKB\_CD\_incident\_time$time=as.numeric(UKB\_CD\_incident\_time$time)

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ smoking1+f.31.0.0.x+f.21022.0.0, data=UKB\_CD\_incident\_time))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ smoking+f.31.0.0.x+f.21022.0.0, data=UKB\_CD\_incident\_time))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ pack\_year+f.31.0.0.x+f.21022.0.0, data=UKB\_CD\_incident\_time))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ age\_stop\_smk+f.31.0.0.x+f.21022.0.0, data=UKB\_CD\_incident\_time))

cox.HLS1\_category

#Multivariate model#

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ smoking1+f.31.0.0.x+f.21022.0.0+drinking+PA+processed.meat+education+f.21001.0.0+pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10, data=UKB\_CD\_incident\_time))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ smoking+f.31.0.0.x+f.21022.0.0+drinking+PA+processed.meat+education+f.21001.0.0+pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10, data=UKB\_CD\_incident\_time))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ pack\_year+f.31.0.0.x+f.21022.0.0+drinking+PA+processed.meat+education+f.21001.0.0+pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10, data=UKB\_CD\_incident\_time))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ age\_stop\_smk+f.31.0.0.x+f.21022.0.0+drinking+PA+processed.meat+education+f.21001.0.0+pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10, data=UKB\_CD\_incident\_time))

cox.HLS1\_category

#statified analysis based on age#

UKB\_CD\_incident\_early<-subset(UKB\_CD\_incident\_time,UKB\_CD\_incident\_time$age<60)

table(UKB\_CD\_incident\_early$site\_new)

UKB\_CD\_incident\_late<-subset(UKB\_CD\_incident\_time,UKB\_CD\_incident\_time$age>=60)

table(UKB\_CD\_incident\_late$site\_new)

#crude model#

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ smoking1+f.31.0.0.x+f.21022.0.0, data=UKB\_CD\_incident\_early))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ smoking+f.31.0.0.x+f.21022.0.0, data=UKB\_CD\_incident\_early))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ pack\_year+f.31.0.0.x+f.21022.0.0, data=UKB\_CD\_incident\_early))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ age\_stop\_smk+f.31.0.0.x+f.21022.0.0, data=UKB\_CD\_incident\_early))

cox.HLS1\_category

#Multivariate model#

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ smoking1+f.31.0.0.x+f.21022.0.0+drinking+PA+processed.meat+education+f.21001.0.0+pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10, data=UKB\_CD\_incident\_early))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ smoking+f.31.0.0.x+f.21022.0.0+drinking+PA+processed.meat+education+f.21001.0.0+pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10, data=UKB\_CD\_incident\_early))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ pack\_year+f.31.0.0.x+f.21022.0.0+drinking+PA+processed.meat+education+f.21001.0.0+pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10, data=UKB\_CD\_incident\_early))

cox.HLS1\_category

cox.HLS1\_category <- summary(coxph(Surv(time,site\_new == 1)~ age\_stop\_smk+f.31.0.0.x+f.21022.0.0+drinking+PA+processed.meat+education+f.21001.0.0+pc1+pc2+pc3+pc4+pc5+pc6+pc7+pc8+pc9+pc10, data=UKB\_CD\_incident\_early))

cox.HLS1\_category

**Note:** These are the example codes for the Cox regression of Crohn’s diease in the UK Biobank. When people perform the analysis for ulcerative colitis, they only need to replace the "CD" with "UC". The dataset for CD and UC were named as “UKB\_CD\_incident\_time.Rdata” and “UKB\_UC\_incident\_time.Rdata”, respectively.

**Two-sample Mendelian randomization analysis**

setwd("D:\\1\_data and codes")

library(utils)

library(data.table)

IBD\_summary <- read.table("D:\\1\_data and codes\\CD\_summary",header=T,sep=" ")

head(IBD\_summary)

smoking\_age <- read.table("D:\\1\_data and codes \\smoking\_age\_snp\_clumping.csv",sep=",",header=T)

head(smoking\_age)

smokingage\_CD <- merge(smoking\_age, IBD\_summary,by="snp")

head(smokingage\_CD)

write.table(smokingage\_CD,file="smokingage\_CD.csv",row.names=FALSE,sep=",")

library(TwoSampleMR)

exp\_dat <- read\_exposure\_data(

filename = "smokingage\_CD.csv",

sep = ",",

snp\_col = "snp",

beta\_col = "beta.exposure",

se\_col = "se.exposure",

effect\_allele\_col = "effect\_allele.exposure",

other\_allele\_col = "other\_allele.exposure",

eaf\_col = "eaf.exposure",

pval\_col = "pval.exposure",

)

outcome\_dat <- read\_outcome\_data(

snps = exp\_dat$SNP,

filename = "smokingage\_CD.csv",

sep = ",",

snp\_col = "snp",

beta\_col = "beta",

se\_col = "standard\_error",

#eaf\_col = "A1FREQ",

effect\_allele\_col = "effect\_allele",

other\_allele\_col = "other\_allele",

pval\_col = "p\_value",

)

dat <- harmonise\_data(

exposure\_dat = exp\_dat,

outcome\_dat = outcome\_dat,

action=3

)

res <- mr(dat)

res\_OR<-generate\_odds\_ratios(res)

write.table(res,file="smoking\_CD\_results.csv",row.names=FALSE,col.names=TRUE,sep=",")

res\_heterogeneity=mr\_heterogeneity(dat)

res\_pleiotropy=mr\_pleiotropy\_test(dat)

res\_single <- mr\_singlesnp(dat)

save(dat, res, res\_heterogeneity, res\_pleiotropy, res\_single, file="smoking\_CD.Rdata")

mr\_report(dat)

#MR PRESSO#

library(MRPRESSO)

mr\_presso(BetaOutcome = "beta.outcome", BetaExposure = "beta.exposure", SdOutcome = "se.outcome", SdExposure = "se.exposure",

OUTLIERtest = TRUE, DISTORTIONtest = TRUE, data = dat, NbDistribution = 1000, SignifThreshold = 0.05)

**Note:** These are the example codes for the two-sample MR analysis of smoking behaviors with Crohn’s disease. The instrument variables could be found in the Supplementary Table 12, and the summary-level data of CD and UC could be accessed through the web link which was provided in the Data availability section. When people run these analyses, they may need to rename and consolidate some of the data columns.

**Epigenetic MR analysis**

setwd("D:\\1\_data and codes ")

library(data.table)

library(httr)

library(dplyr)

library(openxlsx)

library(stringr)

library(xml2)

library(tibble)

cpgs<-read.xlsx("D:\\1\_data and codes\\smoking CpG sites.xlsx",sheet=1)

#Because there are too many CpG sites, we need to split the data#

cpg1=cpgs[1:800,]

cpg2=cpgs[801:1600,]

cpg3=cpgs[1601:2400,]

cpg4=cpgs[2401:2623,]

#find mqtl for the CpG site from GoDMC#

query <- list(

cpgs = cpg1$CpG,

pval\_are = 1e-08)

res <- POST("http://api.godmc.org.uk/v0.1/query", body = query, encode = "json")

dat <- content(res) %>% lapply(., as\_tibble) %>% bind\_rows

head(dat)

##data merge#

IBD\_summary <- read.table("D:\\1\_data and codes\\CD\_summary",header=T,sep=" ")

head(IBD\_summary)

colnames(IBD\_summary)[1]="rsid"

head(IBD\_summary)

mergedata <- merge(dat,IBD\_summary,by="rsid")

write.csv(mergedata,file = "smoking\_mqtl\_CD\_merge.csv",row.names = F)

mqtl <- read.csv("smoking\_mqtl\_CD\_merge.csv",header = T)

head(mqtl)

submqtl <- subset(mqtl,pval\_are<1e-08&cistrans==1)

head(submqtl)

write.table(submqtl,file = "smoking\_cis\_mqtl\_CD.csv",sep = ",",row.names = F)

#MR#

library(TwoSampleMR)

exp\_dat <- read\_exposure\_data(

filename = "smoking\_cis\_mqtl\_CD.csv",

sep = ",",

phenotype\_col = "cpg",

snp\_col = "rsid",

beta\_col = "beta\_are\_a1",

se\_col = "se\_are",

effect\_allele\_col = "a1",

other\_allele\_col = "a2",

eaf\_col = "freq\_a1",

pval\_col = "pval\_are",

)

exp\_dat <- clump\_data(exp\_dat,clump\_r2 = 0.01)

outcome\_dat <- read\_outcome\_data(

snps = exp\_dat$SNP,

filename = "smoking\_cis\_mqtl\_CD.csv",

sep = ",",

phenotype\_col = "cpg",

snp\_col = "rsid",

beta\_col = "beta",

se\_col = "standard\_error",

effect\_allele\_col = "effect\_allele",

other\_allele\_col = "other\_allele",

#eaf\_col = "Freq1",

pval\_col = "p\_value",

)

dat <- harmonise\_data(

exposure\_dat = exp\_dat,

outcome\_dat = outcome\_dat,

action=1

)

dat1 <- subset(dat,dat$exposure==dat$outcome)

res <- mr(dat1)

sin <- mr\_singlesnp(dat1)

het<- mr\_heterogeneity(dat1)

plt <- mr\_pleiotropy\_test(dat1)

write.table(dat1,file = "1harmonise\_smoking\_CD.csv",sep = ",",row.names = F)

write.table(res,file = "1result\_smoking\_CD.csv",sep = ",",row.names = F)

write.table(sin,file = "1single\_smoking\_CD.csv",sep = ",",row.names = F)

write.table(het,file = "1heterogeneity\_smoking\_CD.csv",sep = ",",row.names = F)

write.table(plt,file = "1pleiotropy\_smoking\_CD.csv",sep = ",",row.names = F)

#The above analysis was performed on the split data separately#

#combine the results#

data1 <- read.csv("1pleiotropy\_smoking\_CD.csv",header = T)

data2 <- read.csv("2pleiotropy\_smoking\_CD.csv",header = T)

data3 <- read.csv("3pleiotropy\_smoking\_CD.csv",header = T)

data4 <- read.csv("4pleiotropy\_smoking\_CD.csv",header = T)

bind <- rbind(data1,data2,data3,data4)

write.table(bind,file="pleiotropy\_smoking\_CAD.csv",sep = ",",row.names = F)

#FDR correction#

result <- read.table("result\_smoking\_CD\_IVW\_wald ratio.csv",header=T,sep=",")

result$FDR <- p.adjust(result$pval, method = "BH")

write.table(result,file="result\_smoking\_CD\_FDR.csv",sep=",",row.names = F)

#The effect direction was corrected based on the beta value of the original EWAS#

data <- read.table("result\_smoking\_CD\_FDR.csv",header=T,sep=",")

Smoking\_EWAS <- read.xlsx("D:\\1\_博士\\1\_Xue Group\\1\_IBD\\2-IBD\\1methylation-IBD analysis\\2IBD methylation MR\\CpG sites-v1.xlsx",sheet=2)

merge <- merge(data,Smoking\_EWAS,by="CpG")

head(merge)

df <- mutate(merge,beta\_corrected=ifelse(beta<0,-b,b))

head(df)

write.table(df,file="smoking\_CD.csv",sep=",",row.names=F)

**Note**: These are the example codes for the two-sample MR analysis of smoking-related CpG sites with Crohn’s disease. The smoking-related CpG sites could be found in the Supplementary Table 13, and the summary-level data of CD and UC could be accessed through the web link which was provided in the Data availability section. When people run the analysis for UC, they just need to change the summary-level data. If you are having problems running the codes, please do not hesitate to contact us.

**Colocalization analysis**

setwd("D:\\1\_data and codes")

library(coloc)

library(dplyr)

library(locuscomparer)

IBD\_summary<- read.table("D:\\1\_data and codes\\CD\_summary",header=T,sep=" ")

head(IBD\_summary)

colnames(IBD\_summary)[colnames(IBD\_summary)=="snp"] <- "rsid"

dat=read.csv("cg03599224\_mqtl.csv",header=T)

head(dat)

data=merge(dat,IBD\_summary,by="rsid",all=F,suffixes=c("\_mqtl","\_gwas"))

input=data

head(input)

input$varbeta=input$se\*input$se

summary(input$freq\_a1)

input$MAF=ifelse(input$freq\_a1<0.5,input$freq\_a1,1-input$freq\_a1)

summary(input$MAF)

result <- coloc.abf(dataset1=list(pvalues=input$p\_value, type="cc", s=0.3, N=40266),

dataset2=list(pvalues=input$pval\_are, type="quant", N=input$samplesize), MAF=input$MAF)

need\_result=result$results %>% filter(SNP.PP.H4 > 0.80)

single\_results=result$results

##Visualization##

gwas\_fn=input[,c("rsid","p\_value")]

mqtl\_fn=input[,c("rsid","pval\_are")]

gwas\_fn=rename(gwas\_fn,c("rsid"="rsid"),c("pval"="p\_value"))

mqtl\_fn=rename(mqtl\_fn,c("rsid"="rsid"),c("pval"="pval\_are"))

tiff(file="cg03599224\_rs1799964.tiff",units='cm',width=20,height=20,res=300,

compression = "lzw")

locuscompare(in\_fn1 = gwas\_fn, in\_fn2 = mqtl\_fn,marker\_col1 = "rsid", pval\_col1 = "pval", title1 = 'GWAS',

marker\_col2 = "rsid", pval\_col2 = "pval",title2 = 'mQTL',snp="rs1799964",population="EUR")

dev.off()

**Note**: Before performing colocalization analysis, you need to find all mqtls of the target CpG sites from the previous step and the significant eqtl of IBD.

**G-E interaction analysis**

setwd("D:\\1\_data and codes")

#load("incident\_UC\_idata.Rdata")#481525,incident 1658,control=479867#

#head(idata)

load("D:\\1\_data and codes\\UKB\_UC\_incident\_time.Rdata")

head(UKB\_UC\_incident\_time)#case=2889,n=481743#

load("ukb66354.UC.mqtl.snps.wGRS.487409.dosage.Rdata")

head(ukb66354\_UC.mqtl\_dosage)#n=487254#

colnames(ukb66354\_UC.mqtl\_dosage)[48]="f.eid"

UC\_idata <- merge(UKB\_UC\_incident\_time,ukb66354\_UC.mqtl\_dosage,by="f.eid")

head(UC\_idata)

table(UC\_idata$site\_new)#case=2886,n=481902#

save(UC\_idata, file="UC\_idata.Rdata")

load("UC\_idata.Rdata")

#smoking#

a=c("0","1","2")

UC\_idata$smoking=ifelse(UC\_idata$f.20116.0.0%in%a, UC\_idata$f.20116.0.0, NA)

table(UC\_idata$smoking)

#age stopped smoking#

table(UC\_idata$f.2897.0.0)

UC\_idata$age\_stop\_smk <-ifelse(UC\_idata$f.2897.0.0 <= 10, 1,

ifelse(UC\_idata$f.2897.0.0 > 10 & UC\_idata$f.2897.0.0<= 20, 2,

ifelse(UC\_idata$f.2897.0.0 > 20 & UC\_idata$f.2897.0.0<= 30, 3,

ifelse(UC\_idata$f.2897.0.0 > 30 & UC\_idata$f.2897.0.0<= 40, 4,

ifelse(UC\_idata$f.2897.0.0 > 40 & UC\_idata$f.2897.0.0<= 50, 5,

ifelse(UC\_idata$f.2897.0.0 > 50 & UC\_idata$f.2897.0.0<= 60, 6,

ifelse(UC\_idata$f.2897.0.0 > 60 & UC\_idata$f.2897.0.0<= 70, 7, 0)))))))

table(UC\_idata$age\_stop\_smk)

#pack years of smoking

table(UC\_idata$f.20161.0.0)

UC\_idata$pack\_year <- factor(ifelse(UC\_idata$f.20161.0.0 <= 10, 1,

ifelse(UC\_idata$f.20161.0.0 > 10 & UC\_idata$f.20161.0.0 <= 20, 2,

ifelse(UC\_idata$f.20161.0.0 > 20 & UC\_idata$f.20161.0.0 <= 30, 3,

ifelse(UC\_idata$f.20161.0.0 > 30 & UC\_idata$f.20161.0.0 <= 40, 4,

ifelse(UC\_idata$f.20161.0.0 > 40 & UC\_idata$f.20161.0.0 <= 50, 5,

ifelse(UC\_idata$f.20161.0.0 > 50 & UC\_idata$f.20161.0.0 <= 60, 6,

ifelse(UC\_idata$f.20161.0.0 > 60, 7, 0))))))))

table(UC\_idata$pack\_year)

#alcohol drinking#

UC\_idata$f.1568.0.0=as.numeric(UC\_idata$f.1568.0.0)

UC\_idata$f.1578.0.0=as.numeric(UC\_idata$f.1578.0.0)

UC\_idata$f.1588.0.0=as.numeric(UC\_idata$f.1588.0.0)

UC\_idata$f.1598.0.0=as.numeric(UC\_idata$f.1598.0.0)

UC\_idata$f.1608.0.0=as.numeric(UC\_idata$f.1608.0.0)

UC\_idata$red=UC\_idata$f.1568.0.0\*125\*0.135

UC\_idata$white=UC\_idata$f.1578.0.0\*125\*0.12

UC\_idata$beer=UC\_idata$f.1588.0.0\*570\*0.055

UC\_idata$spirits=UC\_idata$f.1598.0.0\*30\*0.415

UC\_idata$fortified=UC\_idata$f.1608.0.0\*58\*0.19

UC\_idata$alcohol\_g\_day = (UC\_idata$f.1568.0.0\*125\*0.135+UC\_idata$f.1578.0.0\*125\*0.12 +

UC\_idata$f.1588.0.0\*570\*0.055+UC\_idata$f.1598.0.0\*30\*0.415+

UC\_idata$f.1608.0.0\*58\*0.19)\*0.789/7

UC\_idata$alcohol\_10g\_day=UC\_idata$alcohol\_g\_day/10

UC\_idata$drinking=cut(UC\_idata$alcohol\_g\_day, breaks=c(0, 12.5, 50, Inf),

labels = c(0, 1, 2),

na.rm = TRUE)

#Basic model#

#Environmental factor only#

UC\_idata$f.21022.0.0=as.numeric(UC\_idata$f.21022.0.0)

UC\_idata$pack\_year=as.numeric(UC\_idata$pack\_year)

UC\_idata$smoking=as.factor(UC\_idata$smoking)

model=glm(site\_new~smoking+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #smoking status#

summary(model)

model=glm(site\_new~pack\_year+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #pack years of smoking#

summary(model)

model=glm(site\_new~age\_stop\_smk+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #age stopped smoking#

summary(model)

model=glm(site\_new~drinking+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #drinking status#

summary(model)

model=glm(site\_new~alcohol\_g\_day+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #drinking grams per day#

summary(model)

UC\_idata$f.24006.0.0=as.numeric(UC\_idata$f.24006.0.0)

model=glm(site\_new~f.24006.0.0+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #PM2.5 continuous#

summary(model)

UC\_idata$f.24003.0.0=as.numeric(UC\_idata$f.24003.0.0)

model=glm(site\_new~f.24003.0.0+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #NO2 continuous#

summary(model)

UC\_idata$f.30710.0.0=as.numeric(UC\_idata$f.30710.0.0)

model=glm(site\_new~f.30710.0.0+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #CRP continuous#

summary(model)

UC\_idata$f.22192.0.0=as.numeric(UC\_idata$f.22192.0.0)

model=glm(site\_new~f.22192.0.0+f.31.0.0.x+f.21022.0.0,data=UC\_idata,family=binomial) #LTL continuous#

summary(model)

#Genetic factor only#

#if (!requireNamespace("BiocManager", quietly = TRUE))

#install.packages("BiocManager")

#BiocManager::install("CGEN")

library(CGEN)

UC\_idata$rs1799964=round(UC\_idata$rs1799964)

head(UC\_idata)

ret1 =snp.logistic(UC\_idata, "site\_new", "rs1799964", main.vars=c("f.21022.0.0","f.31.0.0.x","pc1","pc2","pc3","pc4","pc5","pc6","pc7","pc8","pc9","pc10"), int.vars=NULL,

strata.var=NULL, op=NULL)

getSummary(ret1)

#G-E interaction#

UC\_idata$rs1799964=round(UC\_idata$rs1799964)

ret2 =snp.logistic(UC\_idata, "site\_new", "rs1799964", main.vars=c("f.21022.0.0","f.31.0.0.x","pc1","pc2","pc3","pc4","pc5","pc6","pc7","pc8","pc9","pc10"), int.vars="smoking",

strata.var=NULL, op=NULL)

getSummary(ret2)

UC\_idata$rs1799964=round(UC\_idata$rs1799964)

ret2 =snp.logistic(UC\_idata, "site\_new", "rs1799964", main.vars=c("f.21022.0.0","f.31.0.0.x","pc1","pc2","pc3","pc4","pc5","pc6","pc7","pc8","pc9","pc10"), int.vars="pack\_year",

strata.var=NULL, op=NULL)

getSummary(ret2)

UC\_idata$rs1799964=round(UC\_idata$rs1799964)

ret2 =snp.logistic(UC\_idata, "site\_new", "rs1799964", main.vars=c("f.21022.0.0","f.31.0.0.x","pc1","pc2","pc3","pc4","pc5","pc6","pc7","pc8","pc9","pc10"), int.vars="age\_stop\_smk",

strata.var=NULL, op=NULL)

getSummary(ret2)

#Multivariate model#

#Physical activity#

UC\_idata$f.884.0.0=as.numeric(UC\_idata$f.884.0.0)

UC\_idata$f.894.0.0=as.numeric(UC\_idata$f.894.0.0)

UC\_idata$f.904.0.0=as.numeric(UC\_idata$f.904.0.0)

UC\_idata$f.914.0.0=as.numeric(UC\_idata$f.914.0.0)

UC\_idata$pp=ifelse((UC\_idata$f.884.0.0>=0|UC\_idata$f.894.0.0>=0|UC\_idata$f.904.0.0>=0|UC\_idata$f.914.0.0>=0), 0,NA)

UC\_idata$PA[UC\_idata$pp==0]="0"

UC\_idata$PA[UC\_idata$f.884.0.0>=5|UC\_idata$f.904.0.0>=1|(UC\_idata$f.894.0.0>=150&UC\_idata$f.894.0.0<1440)|(UC\_idata$f.914.0.0>=75&UC\_idata$f.914.0.0<1440)]="1"

#Processed meat#

UC\_idata$processed.meat=as.numeric(UC\_idata$f.1349.0.0)

UC\_idata$processed.meat[UC\_idata$processed.meat<0]=NA

UC\_idata$processed.meat[UC\_idata$processed.meat>0&UC\_idata$processed.meat<=1]="0"

UC\_idata$processed.meat[UC\_idata$processed.meat>1]="1"

#Enviromental factors only#

UC\_idata$f.21001.0.0=as.numeric(UC\_idata$f.21001.0.0)

UC\_idata$f.48.0.0=as.numeric(UC\_idata$f.48.0.0)

UC\_idata$f.50.0.0=as.numeric(UC\_idata$f.50.0.0)

UC\_idata$CRP\_10=UC\_idata$f.30710.0.0/10

model=glm(site\_new~smoking+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #smoking status#

summary(model)

model=glm(site\_new~pack\_year+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #pack years of smoking#

summary(model)

model=glm(site\_new~age\_stop\_smk+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #age stopped smoking#

summary(model)

model=glm(site\_new~drinking+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #drinking status#

summary(model)

model=glm(site\_new~alcohol\_10g\_day+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #drinking 10 grams per day#

summary(model)

model=glm(site\_new~f.24006.0.0+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #PM2.5 continuous#

summary(model)

model=glm(site\_new~f.24003.0.0+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #NO2 continuous#

summary(model)

model=glm(site\_new~CRP\_10+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #CRP continuous#

summary(model)

model=glm(site\_new~f.22192.0.0+f.31.0.0.x+f.21022.0.0+PA+processed.meat+f.21001.0.0+f.48.0.0+f.50.0.0,data=UC\_idata,family=binomial) #LTL continuous#

summary(model)

#Genetic factor only#

library(CGEN)

UC\_idata$rs1799964=round(UC\_idata$rs1799964)

ret1 =snp.logistic(UC\_idata, "site\_new", "rs1799964", main.vars=c("f.21022.0.0","f.31.0.0.x","PA","processed.meat","f.21001.0.0","f.48.0.0","f.50.0.0","pc1","pc2","pc3","pc4","pc5","pc6","pc7","pc8","pc9","pc10"), int.vars=NULL,

strata.var=NULL, op=NULL)

getSummary(ret1)

#G-E interaction#

UC\_idata$rs1799964=round(UC\_idata$rs1799964)

ret2 =snp.logistic(UC\_idata, "site\_new", "rs1799964", main.vars=c("f.21022.0.0","f.31.0.0.x","PA","processed.meat","f.21001.0.0","f.48.0.0","f.50.0.0","pc1","pc2","pc3","pc4","pc5","pc6","pc7","pc8","pc9","pc10"), int.vars="smoking",

strata.var=NULL, op=NULL)

getSummary(ret2)

UC\_idata$rs1799964=round(UC\_idata$rs1799964)

ret2 =snp.logistic(UC\_idata, "site\_new", "rs1799964", main.vars=c("f.21022.0.0","f.31.0.0.x","PA","processed.meat","f.21001.0.0","f.48.0.0","f.50.0.0","pc1","pc2","pc3","pc4","pc5","pc6","pc7","pc8","pc9","pc10"), int.vars="pack\_year",

strata.var=NULL, op=NULL)

getSummary(ret2)

UC\_idata$rs1799964 =round(UC\_idata$rs1799964 )

ret2 =snp.logistic(UC\_idata, "site\_new", "rs1799964 ", main.vars=c("f.21022.0.0","f.31.0.0.x","PA","processed.meat","f.21001.0.0","f.48.0.0","f.50.0.0","pc1","pc2","pc3","pc4","pc5","pc6","pc7","pc8","pc9","pc10"), int.vars="age\_stop\_smk",

strata.var=NULL, op=NULL)

getSummary(ret2)

**Note:** These are the example codes for the G-E interaction analysis of ulcerative colitis. We have shared the datasets according to the request of the journal. Please find the details in the Data availability section. When people run these analysis for CD, the need to change the corresponding datasets. If you are having problems running the codes, please do not hesitate to contact us.