**Gene-Lifestyle Interactions (CHARGE)**

**Psychosocial-Lipids Replication Analysis Plan**

DC Rao ([rao@.wustl.edu](mailto:rao@.wustl.edu)); Patricia Munroe ([p.b.munroe@qmul.ac.uk](mailto:p.b.munroe@qmul.ac.uk));

Solomon K. Musani ([smusani@umc.edu](mailto:smusani@umc.edu)); Myriam Fornage ([myriam.fornage@uth.tmc.edu](mailto:myriam.fornage@uth.tmc.edu)); Mario Sims ([msims2@umc.edu](mailto:msims2@umc.edu))

**Date of release: July 25, 2017**

Timeline for completing cohort- and ancestry-specific replication analyses and uploading the results for this project

**November 4, 2017**

*Thank you again for your participation as a replication study in the Gene-Lifestyle Interactions Working Group of the CHARGE consortium. An updated “Data Preparation for Replications” document is attached separately which provides details about the data preparation for replication analyses.*

*Herewith we are providing three additional documents (in addition to the document mentioned above): (1) Psychosocial-Lipids Replication Analysis Plan (this document), which specifies a time line for completing and uploading your results to a central location (a Google Drive); (2) An Excel spreadsheet for collecting basic information about your study/cohort; and (3) A list of SNPs for replication analysis; the list includes all genome-wide significant and suggestive SNPs. The SNP list may seem large but this includes all significant and suggestive SNPs from analysis using 1000G imputations.*

*Due to the special nature of the lifestyle measure considered here (psychosocial), sample size for our Discovery phase is much smaller than for other commonly measured lifestyle measures such as smoking. Accordingly, to maximize the chances of novel discoveries, we invite you to consider full participation by carrying out genome-wide analysis of interactions using 1000G imputations (or using as many variants as your study has). However, if that is not convenient for you at this time, we are happy to receive only replication results for the subset of SNPs attached. Please upload your results following the specific instructions given in this analysis plan. If you do not have the data needed for this project, please let us know ASAP.*

*We look forward to your continued collaboration, thank you!*

*DC Rao, Patricia Munroe, Solomon K. Musani, Myriam Fornage, and Mario Sims*

**Goal of this Replication Project**

The primary goal is to replicate significant and promising SNPs found in the Psychosocial-Lipids interactions discovery phase by performing analysis of main and interaction effects for each of three lipid phenotypes (HDL, LDL and TG) with any of three psychosocial measures available (depression, anxiety, and social support; see the “Data Preparation for Replications” document for definitions)for a specified list of SNPs.

To run the analysis and upload the results, please follow this Plan, which provides:

1. Overview of data preparation, the analysis model, and recommended software
2. Instructions for preparing and uploading the results
3. A separate attachment providing the list of SNPs; this list includes all SNPs in novel lipid loci that are genome-wide significant or promising (at P < 10-5). Some of the known lipid loci are also included to examine the role of interactions.
4. A separate attachment (Excel spreadsheet) seeking summary statistics on psychosocial measures and lipids, and other related information.

**Data Preparation & Analysis Methods**

For this analysis project, you will need the following variables:

* Phenotypes: HDL, LDL, TG
* At least one psychosocial trait: DEPR, ANXT, SOCS
* Covariates: age, sex, field centers (for multi-center studies), and PCs (if necessary)

For complete details on how these variables should be prepared, please refer to the updated “Data Preparations for Replications” document (provided along with this document).

PLEASE NOTE:

1. Analyze men and women between 18 and 80 years of age with existing data.
2. Four ancestry groups are considered for replication analysis: European ancestry (EA), African ancestry (AA), Hispanic ancestry (HA), and Asian (AS). Please analyze and report each ancestry group separately.
3. For longitudinal studies, please choose a single visit for each ancestry group that maximizes the sample size. Once chosen, these studies can then follow the instructions as for a cross-sectional study.
4. Family studies require adequate statistical correction for dependencies among family members, while avoiding potential deflation. For more details, see the Methods section below.
5. For case/control studies, run all analyses within case and control samples separately.

Analysis Model

**Joint analysis of main and interaction effects**: For each combination of the Lifestyle variable (E) and Phenotype (Y), fit the following standard linear regression model

**Y = β0 + βE E + βG SNP + βGE E \* SNP + βC C**

1. Y is the (final) phenotype value (each of three serum lipid phenotypes prepared following the Data Preparation for Replications)
2. β0 is the intercept
3. E is the lifestyle exposure variable (each psychosocial variable available)
4. SNP is the dosage of the genetic variant, coded additively accounting for genotype imputation uncertainty
5. C is the vector of all covariates, including age, sex, field centers (for multi-center studies), and PCs, as described in the “Covariates” section of the Data Preparation for Replication document. Do not adjust for BMI.

**What to report**: Provide results as shown in the “RESULTS FORMAT” section below. Provide the **βG** (“main effect” of the SNP) and its SE, **βGE** (interaction effect) and its SE, and the covariance between **βG** and **βGE**.To control potential inflation of false positives, we obtained the **robust estimates** of thecovariance matrix (**robust SE**s as well as the **robust covariance** between **βG** and **βGE**) from all discovery cohorts. If at all possible, we would greatly appreciate receiving the robust estimates from all cohorts participating in replication. If that is not possible, we will accept model-based estimates, but please let us know which is provided (model-based or robust).

Software for robust SEs & covariance

Our discovery cohorts used the following packages to obtain robust SEs and robust covariance:

1. ProbABEL version 0.1.3 or newer (<http://www.genabel.org/>)
2. MMAP (<http://edn.som.umaryland.edu/mmap/>)
3. The R package sandwich in R (<https://cran.r-project.org/web/packages/sandwich/>)

We therefore recommend using the above or other suitable packages.

Methods/Software for Analysis

1. **For longitudinal studies** (i.e. if more than one visit/measurement per participant is available): Use data from the one visit with the maximum information and follow the same methods as proposed for cross-sectional studies (below).
2. **For cross-sectional studies** (i.e. if only one visit/measurement per participant is available): For **unrelated studies** (with unrelated subjects), you may use linear regression (with robust estimates of standard errors and covariance, if possible): you may use packages appropriate for interaction analysis (as shown above; the Appendix provides code for ProbABEL). For **family studies** (with related subjects), you may use linear mixed model using a kinship matrix to account for family relationships, or other packages which deal with relatedness in families appropriately. We discourage using GEE for analysis of family studies. Please clarify which method and package were used for your analysis.

File Preparation and Upload Details

**DATA EXCHANGE:**

Results should be uploaded to the CHARGE Google Drive. Follow the ‘FILE NAMING SCHEME’ section below for naming the results files and the ‘RESULTS FORMAT’ section below for naming and including variables. Only summary results will be transferred, not individual level genotype or phenotype data. You may want to compress results files before uploading.

Please select one person from your study to upload the files. This person must have a valid Gmail email address. Please contact Solomon K. Musani at [smusani@umc.edu](mailto:smusani@umc.edu) with the name and email address of the person who will be uploading the files, and he will facilitate access to Google Drive. Some studies have chosen to create a new Gmail email address dedicated solely for data uploads.

You will be given access to the Google Drive “staging” folder, called ***upload\_CHARGE – Psychosocial and Lipids: REPLICATIONS***. Once you are ready to upload, please follow the instructions on the Google Drive website and upload your results here. Once complete, your files will automatically be copied within a few minutes to the “main” folder, called *CHARGE – Psychosocial and Lipids: REPLICATIONS*. Once you have confirmed that the system has copied your files into the “main” folder, you can then remove your files from the “staging” folder. Users do not have permission to change, edit, or remove files from the “main” folder. If you have accidentally uploaded the wrong files and need them removed, please contact Solomon K. Musani. Again, please make sure to upload your files to the folder called ***upload\_CHARGE – Psychosocial and Lipids: REPLICATIONS***.

**FILE NAMING SCHEME:**

Please use the following file naming scheme:

STUDY.ANCESTRY.PHENOTYPE.PSY.txt

**STUDY** is a short (14 characters or less) identifier for the population studied.

**ANCESTRY** options: “EA” or “AA” or “HA” or “AS”

**PHENOTYPE** options: “HDL” or “LDL” or “TG”

**PSY** options: “DEPR” or “ANXT” or “SOCS”

To better illustrate the file naming scheme, here is an example:

JHS.AA.HDL.DEPR.txt

**RESULTS FORMAT:**

1. Provide the results in tab- or space-delimited text files, including a single header line with all columns in the order listed below.
2. Please keep at least 5 digits after the decimal place for all statistics (the use of more precision is recommended as shown in the Tables below).
3. Integers should be single numbers without decimals.
4. No quotes should be used around any data cells or headers.
5. No other extra columns should be provided.
6. Please code missing values in any columns as a single period character (“.”).

Please provide the results of each of the analyses in a separate file, named as described in the ‘FILE NAMING SCHEME’ section above. **Following the requested format and naming scheme for your results will greatly assist us in collecting and processing the data from many different groups while minimizing errors.**

***Please provide results with all columns listed below***

|  |  |  |  |
| --- | --- | --- | --- |
| **Column header** | **Description** | **Recommended format** | **Examples** |
| rsID | The rs‐number of the SNP analyzed | rs‐number | rs3845291 |
| CHR | Chromosome Number | Numeric, integer | 1 |
| POS | Position of the variant | Numeric, integer | 132146 |
| STRAND | Orientation of the site to the human genome strand used | A single character, ‐ or +. Strong preference for + strand orientation | + |
| EFFECT\_ALLELE | Allele to which the effect has been estimated | A single upper‐case character (A, C, G, or T) | A |
| NON\_EFFECT\_ALLELE | Allele which is not the effect allele | A single upper‐case character (A, C, G, or T) | T |
| EAF | Allele frequency of the EFFECT\_ALLELE | At least 5 digits to the right of the decimal and use scientific E notation. | 0.35412 |
| IMPUTATION | A value (range 0‐1) corresponding to the imputation quality measure (Rsq from MACH/Minimac or info from IMPUTE2) | Numeric fraction 0 to 1 | 0.954565 |
| BETA\_SNP | Beta‐coefficient for the association of SNP with DEPENDENT VARIABLE (main effect) | At least 6 digits to the right of the decimal. | 0.045228 |
| SE\_SNP | Robust standard error for the association of SNP with the DEPENDENT VARIABLE (main effect) | At least 6 digits to the right of the decimal. | 0.018343 |
| P\_SNP | P value for the association of SNP with the DEPENDENT VARIABLE (main effect) | At least 6 digits and use scientific E notation | 6.219424E‐10 |
| BETA\_INT | Beta‐coefficient for the SNPxE interaction. | At least 6 digits to the right of the decimal. | 0.045228 |
| SE\_INT | Robust standard error for the SNPxE interaction. | At least 6 digits to the right of the decimal. | 0.018343 |
| P\_INT | P value for the SNPxE interaction. | At least 6 digits and use scientific E notation | 6.212423E‐10 |
| COVAR\_SNP\_INT | Robust covariance between BETA\_SNP and BETA\_INT. | At least 6 digits to the right of the decimal. | 0.002343 |

**SIMPLE LOOKUPS FOR MAIN EFFECTS:**

If your study already has pre-existing “main effects” results for any of the 3 phenotypes (HDL, LDL, TG), without consideration of interactions, please provide these results for the same attached list of SNPs in separate datasets. **We are not asking for additional analyses here; only simple look-ups of existing results, if available.** Please use the following file naming scheme:

STUDY.ANCESTRY.PHENOTYPE.LOOKUPS.txt

To better illustrate the file naming scheme, here is an example:

JHS.AA.HDL.LOOKUPS.txt

**LOOKUPS: If you are providing lookup results, please include all columns listed below for the list of SNPs provided. In lieu of this, we will also accept complete GWAS results.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column header** | **Description** | **Recommended format** | **Examples** |
| rsID | The rs‐number of the SNP analyzed | rs‐number | rs3845291 |
| CHR | Chromosome Number | Numeric, integer | 1 |
| POS | Position of the variant | Numeric, integer | 132146 |
| STRAND | Orientation of the site to the human genome strand used | A single character, ‐ or +. Strong preference for + strand orientation | + |
| EFFECT\_ALLELE | Allele to which the effect has been estimated | A single upper‐case character (A, C, G, or T) | A |
| NON\_EFFECT\_ALLELE | Allele which is not the effect allele | A single upper‐case character (A, C, G, or T) | T |
| EAF | Allele frequency of the EFFECT\_ALLELE | At least 5 digits to the right of the decimal and use scientific E notation. | 0.35412 |
| IMPUTATION | A value (range 0‐1) corresponding to the imputation quality measure (Rsq from MACH/Minimac or info from IMPUTE2) | Numeric fraction 0 to 1 | 0.954565 |
| BETA\_SNP | Beta‐coefficient for the association of SNP with DEPENDENT VARIABLE (main effect) | At least 6 digits to the right of the decimal. | 0.045228 |
| SE\_SNP | Robust standard error for the association of SNP with the DEPENDENT VARIABLE (main effect) | At least 6 digits to the right of the decimal. | 0.018343 |
| P\_SNP | P value for the association of SNP with the DEPENDENT VARIABLE (main effect) | At least 6 digits and use scientific E notation | 6.219424E‐10 |

**README FILE:  EXCEL template DiSTRIBUTED WITH this replication analysis PLAn**

When uploading the result files to the CHARGE Google Drive (following instructions provided above in “DATA EXCHANGE”), please fill in, rename, and upload the accompanying Excel file “**STUDY.ANCESTRY.PSYCHOSOCIAL.LIPIDS.README.xls**”. This Excel file asks for information about the lipid phenotypes and the psychosocial variables used in the analyses. In particular, the following information is required:

**1. Contacts.** List the contact information for the study: name of the Principal Investigator (PI) and the Contact Analyst, and their email and telephone number.

**2. Study characteristics.** Provide information about the characteristics of your study and genotype data, including:

* Whether the study is unrelated (UN), family-based (FB), or case-control (CC)
* How many principal components (PCs) or additional covariates were used
* Information about genotyping platforms and imputation software
* Analysis software, and whether a robust covariance matrix was used (yes/no). For family studies, also provide details on how relationships were handled.

**3. Descriptive Statistics.** Summary statistics for HDL, LDL and TG in the total sample and within the exposed and unexposed groups separately for each psychosocial variable.

**4. Demographic Statistics**. Demographic statistics for age, sex, and statin use.

**5. Simple Lookups**. If you are providing the simple lookup results (described above in “SIMPLE LOOKUPS FOR MAIN EFFECTS”), please complete this tab, which asks for basic information about how those analyses were done.

**6. Psychosocial Statistics.** Summary statistics for each continuous psychosocial measure within categories of that psychosocial measure.

**7. Psychosocial Measures.** Description of the questionnaire used for each psychosocial trait.

Please report the required information in each tab in the Excel file for each ancestry group separately, as indicated. Most information (except for 1) will be used in publication(s). Please remember to rename the file.

**Contact Information**

**Lead Investigators for the Psychosocial-Lipids project:**

Solomon K. Musani ([smusani@umc.edu](mailto:smusani@umc.edu))

Myriam Fornage ([myriam.fornage@uth.tmc.edu](mailto:myriam.fornage@uth.tmc.edu))

Mario Sims ([msims2@umc.edu](mailto:msims2@umc.edu))

Appendix

**Example Code for ProbABEL**

**DISCLAIMER: The code shown below was originally developed for the discovery analyses (for smoking-BP). Although a separate program containing the code is not provided here, you may be able to use this code as a template/guide for your analyses.**

This appendix includes 3 steps. They are:

1. Prepare phenotype file for ProbABEL – This step creates an input file for ProbABEL that includes the trait and covariates (e.g., CURSMK, AGE, SEX).
2. Run ProbABEL – It describes how to run the model in ProbABEL.
3. Post-process ProbABEL result files – This step uses ProbABEL output, selects/renames required variables, and calculates p-values.

How to Use the Code Below: All R codes below are in courier new font. Any code commented in #red below is part of the function. It is provided here for your reference, but should not need to be edited. Any code commented in #green inside a box is an example of using a function; this is the code that you will need to copy and edit for your analysis.

**Step 1: Prepare phenotype input files**

This step creates a phenotype input file that includes the trait and needed covariates (e.g., SBP, CURSMK, AGE, SEX).

Note: In order for these R scripts to work, the subject ID must be unique in the dataset, not just within families. For example, there can only be one subject with ID=1.

The phenotype file must have at least two columns (first for id, second for the trait, in that order), then the remaining columns are considered as covariates (see Step 2 for examples of input files). Please note that for ProbABEL, the phenotype and genotype data have to be in the same order of subjects (one row per subject) and therefore should have the same number of rows. If a subject has partly missing phenotype data, they should still be included in the phenotype file; ProbABEL will automatically exclude them from the analysis as appropriate. In addition, the order of the SNPs in the info files should be the same as the order of the SNPs in the dose files. The function prep.pheno will take care of this.

# This function creates phenotype input file that includes trait,

# the environmental covariate and other covariates to the file.

**prep.pheno** = function(phen.outfile, phen, trait, env.cov, add.cov, geno.id){

### minimal checks for pheno file

if (!any(colnames(phen)=='ID'))

stop('pheno file does not have the column for ID')

if (!any(colnames(phen)==trait))

stop(paste('pheno file does not have the column for',trait))

if (!any(colnames(phen)==env.cov))

stop(paste('pheno file does not have the column for',env.cov))

col.out=c('ID',trait,env.cov,add.cov)

index=match(col.out,colnames(phen))

if (any(is.na(index)))

stop('pheno file dose not have the column for additional covariates')

out=phen[,index]

### reorder rows of out to match with geno.id

### geno.id is a text file containing a list of IDs (with no header), that are ### in the dosage files, in the same order that they appear in the dosage files.

### This code reorders the phenotype data, if needed, to match the dosage files.

index=match(geno.id,phen$ID)

out=out[index,]

out$ID=geno.id

colnames(out)[1]='id'

write.table(out, file=phen.outfile, row.names=F, quote=F)

print(paste('created',phen.outfile))

}

# Example code to use prep.pheno

phen = read.csv("pheno-white.csv", head=T)

geno.id = scan('geno.id')

# Prepare pheno file for ProbABEL

add.cov = c('AGE','SEX','FC1','FC2')

**prep.pheno**('pheno-SBP-CURSMK.txt', phen, 'SBP', 'CURSMK', add.cov,

geno.id)

**Step 2: Run ProbABEL**

Here are examples of what your input files for ProbABEL should look like:

**Input files:**

**pheno-SBP-CURSMK.txt:**

id SBP CURSMK AGE SEX

1 145.1 1 68 2

3 150.2 0 42 1

4 132.4 1 40 2

5 141.0 1 21 2

**ch22.info:**

SNP Al1 Al2 Freq1 MAF AvgCall Rsq

rs131526 G A 0.92785 0.07215 0.93295 0.23028

rs131527 C T 0.9301 0.0699 0.93304 0.22309

rs131531 A G 0.92464 0.07536 0.93179 0.2331

rs131538 G A 0.92667 0.07333 0.93431 0.24482

**ch22.dose:**

1 DOSE 1.999 1.989 1.999 1.999 1.478

3 DOSE 1.999 1.986 1.999 1.999 1.731

4 DOSE 1.997 1.983 1.999 1.996 1.593

5 DOSE 1.999 1.990 1.999 1.999 1.718

See the ProbABEL\_manual.pdf (available from <http://www.genabel.org/sites/default/files/pdfs/ProbABEL_manual.pdf>) for more details.

Below is the code that shows how to run ProbABEL while getting robust standard errors. To calculate robust standard errors in ProbABEL, use the --robust option.

**Commands to run ProbABEL:**

# Example code for running ProbABEL

palinear --pheno pheno-SBP-CURSMK.txt --info ch22.info --dose ch22.dose --robust –-interaction=1 –-out out-ch22-SBP-CURSMK.txt

In the above code, --interaction=1 tells ProbABEL to use the 1st listed covariate as the interaction covariate, because in the phenotype file (see above input file for example), CURSMK is the 1st listed covariate, i.e., the 1st column after the subject ID and trait. It is crucial that the column containing your lifestyle covariate in your phenotype file is correctly identified in the interaction statement above. For example, if the covariates were in the order AGE SEX CURSMK, then you would use --interaction=3. An example output file is shown below; the bold-faced column names correspond to the beta estimates and their robust standard errors.

**Output file: out-ch22\_add.out**

name A1 A2 Freq1 MAF Quality Rsq n Mean\_predictor\_allele **beta\_SNP\_add sebeta\_SNP\_add beta\_SNP\_CURSMK sebeta\_SNP\_CURSMK cov\_SNP\_int\_SNP\_CURSMK** loglik

rs149201999 T C 0.93864 0.06136 0.93864 2045 0.0129584 3.65995 4.12266 -11.6668 4.73242 -16.999 -7230.14

rs146752890 C G 0.91575 0.08425 0.91575 2045 0.0867971 1.27578 1.57563 -4.95806 2.14644 -2.47974 -7229.94

rs139377059 C T 0.94826 0.05174 0.94826 2045 0.452078 -0.561018 0.953585 0.491564 1.29279 -0.907429 -7232.15

**Step 3: Post-process ProbABEL result files**

Because the ProbABEL output doesn’t include the p-values, one can use the following function in R to obtain them. This code also renames and reorders the variables to match the analysis plan. Note prep.upload assumes that you have one ProbABEL result file for each chromosome. If you have chopped chromosomes, then you need to edit this function.

# This function reads ProbABEL output, renames variables, and computes p-values

**p.probabel** = function(probabel.outfile,include.int=TRUE) {

# read ProbABEL output file

out=read.table(probabel.outfile,head=T)

# match column headers to those in the table on page 7 of the analysis plan

new.name=c("rsID","EFFECT\_ALLELE","NON\_EFFECT\_ALLELE","EAF","IMPUTATION",

"BETA\_SNP","SE\_SNP")

old.name=c("name","A1","A2","Mean\_predictor\_allele","Rsq","beta\_SNP\_add",

"sebeta\_SNP\_add")

index=match(old.name,colnames(out))

new.out=out[,index]

colnames(new.out)=new.name

out=out[,-index]

if (include.int) {

index=grep("SNP",names(out))

new.names=c("BETA\_INT","SE\_INT","COVAR\_SNP\_INT")

out=out[,index]

names(out)=new.names

out=cbind(new.out,out)

} else out=new.out

# Compute p-value and create output

out$P\_SNP=pchisq((out$BETA\_SNP/out$SE\_SNP)^2,df=1,lower.tail=F)

if (include.int)

out$P\_INT=pchisq((out$BETA\_INT/out$SE\_INT)^2,df=1,lower.tail=F)

return(out)

}

### This function will create the output file ready to upload

## This function is modified to include position column in the output file

**prep.upload** = function(uploadfile,probabel.files,position.files,strand='+',include.int=T) {

chr=1:22

if (length(probabel.files)!=22)

stop('probabel files should be 22, one for each chromosome')

if (length(position.files)!=22)

stop('position files should be 22, one for each chromosome')

for (i in chr) {

out=p.probabel(probabel.files[i],include.int=include.int)

out$CHR=i

out$STRAND=strand

if (setdiff(colnames(position.files[i]), c(’name’,’position’)))

stop(sprintf('position file for chr %d does not have name and position column',i))

out = merge(out, position.files[i], by=’name’)

if (i==1) write.table(out,file=uploadfile,quote=F,row.names=F,na='.')

else write.table(out,file=uploadfile,quote=F,row.names=F,

col.names=F,append=T,na='.')

print(paste('processed',probabel.files[i]))

}

system(paste('gzip',uploadfile))

print(paste(uploadfile,'is ready to upload'))

}

# Example using the prep.upload function

# Separately, you need to prepare 22 position files.

# Each file(position-ch1.txt) must have only two columns: name and position

position.files=sprint("positions-ch%d.txt",1:22)

# Call this function for ProbABEL output with interaction effects

probabel.files=sprintf("EA-SBP-CURSMK-ch%d\_add.out.txt",1:22)

**prep.upload**(uploadfile="EA.SBP.CURSMK.txt",probabel.files,position.files)