Gene-Lifestyle Interactions (CHARGE)

Data Preparation for Replications

DC Rao (rao@wubios.wustl.edu); Karen Schwander (karen2@wubios.wustl.edu)

Patricia Munroe (p.b.munroe@qmul.ac.uk); Ioanna Ntalla (i.ntalla@qmul.ac.uk)

12/23/15; 10/26/16; 06/16/17

As you know, the CHARGE Gene-Lifestyle Interactions Working Group (WG) is carrying out a series of genome-wide interaction analyses for blood pressure (BP) and lipids using several domains of lifestyle measures [smoking, alcohol consumption, education, physical activity (PA), psychosocial variables, and sleep]. Thank you for providing replication results for several projects. Those projects are now preparing the results for publications.

Several new projects, each devoted to BP or lipids and one lifestyle domain, have identified multiple significant and promising novel loci which need replications. We are writing now to invite your continued participation in this next set of projects. Please create the BP and lipid traits exactly as created for the previous projects. You may use this updated "Data Preparation for Replications" document to prepare the lifestyle data for these "replication" analyses.

The attached project-specific "Replication Analysis Plan" provides the analysis model, lists of SNPs to be analyzed/replicated, format for organizing your results files, and instructions for uploading the results to a central location (a Google Drive).

We look forward to your continued collaboration, thank you!

DC Rao, Karen Schwander, Patricia Munroe, and Ioanna Ntalla

Subjects & Race/Ethnic Groups

1. Men and women between 18 and 80 years of age with existing data.
2. For longitudinal studies, use the one visit with the most data.
3. Four race-ethnic groups are considered for replication analysis: EA (European Ancestry), AA (Africans Ancestry), HA (Hispanic Ancestry), and AS (ASian ancestry).
4. If your study contains individuals from multiple race/ethnicity groups, or separate cohorts, analyze each group/cohort separately and provide as many sets of results.

Data Adjustments for Phenotypes

DATA ADJUSTMENTS FOR BLOOD PRESSURE PHENOTYPES (4 traits):

1. Resting/sitting Systolic Blood Pressure (SBP) (mmHg) and Diastolic Blood Pressure (DBP) (mmHg).
2. Multiple readings: When multiple BP readings are available at the same visit (typically 3 readings), take the average of all SBP readings and average of all DBP readings.
3. For subjects taking any anti-hypertensive (BP lowering) medications, adjust their SBP and DBP values as follows:
    a. Add 15 mmHg to SBP from step 2 above.
    b. Add 10 mmHg to DBP from step 2 above.
4. Derive Mean Arterial Pressure (MAP) and Pulse Pressure (PP) using SBP and DBP values obtained from step 3 above:
    a. MAP = DBP + (SBP – DBP)/3, and
    b. PP = SBP – DBP
5. For each of the 4 BP variables (SBP, DBP, MAP, and PP), Winsorise very extreme values: If a BP value is more than 6 standard deviations (SD) above or below the mean, set it exactly at 6 SDs from the mean; do so in both tails, and separately for each of the 4 BP variables. **Document how many values were Winsorised for each BP variable, if any.**

DATA ADJUSTMENTS FOR LIPID PHENOTYPES (3 TRAITS):

1. Phenotypes to be analyzed:
    a. High-density lipoprotein cholesterol (HDL, mg/dL)

b. Triglycerides (TG, mg/dL)

c. Low-density lipoprotein cholesterol (LDL, mg/dL), either directly assayed (LDL_{da}) or derived using the Friedewald equation (LDL_F). Otherwise, set LDL to missing.

   i. Note: For TG > 400 mg/dL, only LDL_{da} should be used (if you only have LDL_F, set LDL values to missing in those subjects).

2. Fasting Status

   a. If have fasting lipids (fasting ≥ 8 hours), use HDL, TG, and either LDL_{da} or LDL_F.

   b. If have non-fasting lipids only (fasting < 8 hours):

      i. Use LDL_{da} and HDL

      ii. Do not use LDL_F or TG

3. Transformations:

      i. Use natural log for TG and HDL;

      ii. No transformation on LDL

4. Adjustment for statin use:

   a. Only LDL will be adjusted (not TG or HDL)

   b. Modelled after CHARGE Lipids Exome Chip Analysis Plan

   c. Assumptions: Before 1994, commonly used lipid therapy drugs (which were mostly non-statin) were relatively ineffective at lowering lipids. The benchmark "4S" study published in 1994 led to more wide-spread statin use as they were very effective. Therefore, if somebody was on an unspecified lipid lowering drug after 1994, it can be safely assumed to be statin. The following lipid adjustments reflect these assumptions (as implemented in the CHARGE Pharmacogenetics WG).

   d. Perform LDL adjustment as shown in the Table:

Lipid Lowering Drug used	Lipids measured before 1994	Lipids measured during or after 1994
Statin	**Adjust**	**Adjust**
Unspecified	No adjustment	**Adjust**
Non-statin (eg, fibrate and fibric acid derivatives, niacin, binding agents)	No adjustment	No adjustment

   e. When adjustment is indicated per the table above, perform LDL adjustment as follows:

      i. If LDL was derived from Friedewald and TG < 400 mg/dL,

         a) First adjust total cholesterol (TC) for statin use: adjTC = TC / 0.8

         b) Then adjust LDL using adjusted TC: $adjLDL_F$ = adjTC – HDL – (TG/5)

      ii. If LDL is directly assayed (not derived as above),

a) Adjust LDL directly: $adjLDL_{da} = LDL_{da} / 0.7$

f. No adjustments for use of any other lipid lowering medications

Lifestyle Domains (denoted by 'E')

SMOKING STATUS:

Define 2 dichotomous smoking variables.

1. Current smoking status (**CURSMK**) (yes/no)

   CURSMK = 1 if a current smoker (i.e., E=1)

   CURSMK = 0 if not a current smoker (including never and former smokers) (E=0)

2. Ever smoking status (**EVERSMK**) (yes/no)

   EVERSMK = 1 if a current or former smoker (E=1)

   EVERSMK = 0 if a never smoker (E=0)

ALCOHOL CONSUMPTION:

Define three dichotomous alcohol consumption variables. The justification for the second dichotomous variable comes from the fact that many current drinkers may drink no more than 1 drink per week. The potential beneficial effects of current alcohol consumption may occur at 2 or more drinks per week.

1. **Current drinking status (CURDRINK):**

   CURDRINK = 1 if a current drinker (i.e., E=1)

   CURDRINK = 0 if not a current drinker (E=0)

2. **Current regular alcohol use status (REGDRINK):**

   REGDRINK = 1 if a current drinker who drinks ≥2 drinks per week (E=1)

   REGDRINK = 0 if a never or former drinker *or* a current drinker consuming <2 drink**s** per week or less (E=0)

3. <u>ONLY FOR STUDIES WITH ASSESSMENT OF AMOUNT OF ALCOHOL CONSUMPTION</u>: **Quantity of Drinks (QUDRINK) only for subjects who currently drink:**

   QUDRINK = 1 if drinks <u>≥</u> 8 drinks (>100 g ethanol) per week

   QUDRINK = 0 if 1-7 drinks (or <u>≤</u> 100 g ethanol) per week

---

EDUCATION (SOMECOL AND GRADCOL):

Define 2 dichotomous education variables. If your study has information on only one (say, SOMECOL) but not both variables, you may contribute analyses using that variable only. Studies lacking any data on education will not be able to participate. For questions or if you are not sure of how to create the education variables (especially in non-US cohorts), please contact Karen Schwander (karen2@wubios.wustl.edu).

1. **Some College (SOMECOL):**

   SOMECOL = 1 if subject attended any education beyond high school (i.e., E=1)

   SOMECOL = 0 if subject has no education beyond high school/GED (E=0)

2. **Graduated College (GRADCOL):**

   GRADCOL = 1 if subject completed at least a 4 year college degree (BA/BS) (E=1)

   GRADCOL = 0 if subject has not completed a 4 year college degree (E=0)

PHYSICAL ACTIVITY (PA):

Define one *dichotomous physical activity variable, active vs. inactive.*

PA = 0 if physically active (i.e. the exposure variable E=0)

PA = 1 if inactive (E=1)

Definition of 'inactive' is doing <3.75 MET-hours of moderate or vigorous leisure-time or commuting PA per week. The remaining individuals with valid PA data are defined as 'active'. In studies that do not have MET-hours/week measures available, contact Tuomas O. Kilpeläinen (tuomas.kilpelainen@sund.ku.dk) and Ruth J.F. Loos (ruth.loos@mssm.edu) for defining the PA variable in a customized manner.

PSYCHOSOCIAL VARIABLES (DEPR, ANXT, SOCS):

We will focus on 3 dichotomous psychosocial variables derived from existing instrument data. If your study has information on only some (say, DEPR) but not all variables, you may contribute analyses using those variables only. Studies lacking any data on these psychosocial variables will not be able to participate.

1. **Depression (DEPR):**

   Depressed individuals may experience lack of interest and pleasure in daily life, insomnia, lack of energy, loss of focus, and excessive feelings of worthlessness or guilt. Standard cut points for depression-screening scales will be used to identify individuals exhibiting a high proportion of these characteristics.

   DEPR = 1 if subject screened positive for depressive symptoms (E=1)

   DEPR = 0 if subject screened negative for depressive symptoms (E=0)

Assessment	# Items	Score Range	E=1	E=0
CESD	20	0 – 60	≥ 16	< 16
CESD	12	0 – 36	≥ 12	< 12
CESD	10	0 – 30	≥ 10	< 10
SF-36 MH score	36	0 – 100	≤ 52	> 52
SF-8 or SF-36 MCS score	8 or 36	0 – 100	≤ 42	> 42
HADS-D	7	0 – 21	≥ 8	< 8
IDS	28	0 – 84	≥ 14	< 14
Beck (BDI)	21	0 – 63	≥ 10	< 10
Geriatric Depression (GDS)	15	0 – 15	≥ 6	< 6
Zung SDS	20	20 – 80	≥ 50	< 50
GWB-D	3	0 – 25	≤ 13	> 13
PHQ-9	9	0 – 27	≥ 10	< 10

2. **Trait Anxiety (ANXT):**

Trait anxiety is a general measure of one's personal disposition to stress, whereas state anxiety is a measure of one's fear, discomfort, or nervousness induced by temporary situations. The anxiety-assessment cut points provided are indicative of having at least mild to moderate anxiety as a personality trait (general anxiety, but not necessarily as a disorder).

ANXT = 1 if subject has a high index of trait anxiety (E=1)

ANXT = 0 if subject has a low index of trait anxiety (E=0)

Assessment	# Items	Score Range	E=1	E=0
Spielberger (STAI Form Y2)	20	20 – 80	≥ 45	< 45
Beck (BAI)	21	0 – 63	≥ 10	< 10
HADS-A	7	0 – 21	≥ 8	< 8
GWB-A	5	0 – 25	≤ 13	> 13

3. **Social Support (SOCS):**

Social support is a composite of supports thought to assist individuals in handling stress: appraisal/emotional support, tangible/instrumental support, self-esteem/informational support, and belonging/companionship support. SOCS should be created based on an overall score for all types of social support measured by the assessment.

SOCS = 1 if subject has a low index of social support (E=1)

SOCS = 0 if subject has a high index of social support (E=0)

Assessment	# Items	Score Range	E=1	E=0
ISEL	40	40 – 160	Lowest quartile	Highest three quartiles
ISEL	6	6 – 24		
MOS Social Support	19	19 – 95		
SSQ6 S-score	6	1 – 6		

Additional psychosocial measures not listed above may harmonize. For questions about scoring or use of any of the psychosocial measures, please contact Melissa Richard (melissa.a.lee@uth.tmc.edu).

LIFESTYLE VARIABLES: Total Sleep Time (STST & LTST):

Most cohorts have collected information on sleep by using various questionnaires (e.g., PSQI questionnaire). Nevertheless, the questions on "total sleep time" are all similar and can be used for this project. Questions asked to the participant should be similar to "On an average day, how long do you sleep?". Questions can be either "open" or "multiple choice".

Before defining exposure, we request all cohorts (see note below) to obtain the age- and sex-adjusted residuals (separately for each ancestry), which can be obtained using the following R script:

```
df = read.table("yourPhenotypeFile", h=T) # read data with headers
# Note this data set should have column with TST, age and sex

TST.lm = lm(TST ~ age + sex, data = df)
df$TST.res = resid(TST.lm)  # age and sex-adjusted residuals

#Where:
#TST = total sleep time (continuesly in hours)
#age = age of the participant at the time of the sleep assessment
#sex = gender of the participant

#Define exposed participants to Short Total Sleep Time (STST) and Long
Total Sleep Time (LTST):

quantile(df$TST.res, c(0.20, 0.80)) # 20th and 80th percentiles
```

NOTE: this data can only be obtained in cohorts with information on total sleep time based on "open" questions and "multiple-choice" questions with more than 4 alternatives. Please contact **Raymond Noordam (r.noordam@lumc.nl)** in cases where information was obtained differently. For these cohorts, we will make **cohort-specific definitions** for short and long total sleep time.

We define two sleep variables (exposure variables, E) using the distribution of the residuals obtained above through regression analysis, depending on whether the residual is in the lower tail (STST) or in the upper tail (LTST).

1. **Short Total Sleep Time (STST):**

STST = 1 if TST.res ≤ 20th percentile (i.e., E = 1)

STST = 0 otherwise (E = 0)

2. **Long Total Sleep Time (LTST):**

   LTST = 1 if TST.res ≥ 80th percentile (i.e., E = 1)

   LTST = 0 otherwise (E = 0)

For questions relating to the creation of sleep variables, please contact Raymond Noordam (r.noordam@lumc.nl).

Genotypes

This section applies to genome-wide analysis of interactions. Please follow these instructions if you are willing to carry out genome-wide analysis using all the variants available in your study. If you are only providing replication analysis for the small subset of the SNPs provided, you may skip this section.

1. Use dosage of imputed SNPs using data from the 1000 Genomes Project (1000G).
2. If possible, the 1000G imputation should use the ALL ancestry panel (also referred to as the cosmopolitan panel or worldwide panel) from 1000G Phase I Integrated Release Version 3 Haplotypes (2010-11 data freeze, 2012-03-14 haplotypes); this reference panel contains haplotypes of 1,092 individuals of all ethnic backgrounds and excludes monomorphic and singleton sites. For MACH, it is "ALL GIANT.phase1_release_v3.20101123.snps_indels_svs.genotypes.refpanel.ALL.vcf.gz.tgz" available at http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G.2012-03-14.html. For IMPUTE2, it is "ALL_1000G_phase1integrated_v3_impute_macGT1.tgz" available at http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html.
3. Use dosage of imputed SNPs using HapMap Phase II reference panel if 1000G imputations are not available.
4. **SNP EXCLUSIONS**: BEFORE ANALYSIS *please exclude* the following SNPs to reduce the overall analysis burden (and file sizes):
   a. SNPs with very low imputation quality (r^2< 0.1 if using MACH and information metric < 0.1 if using IMPUTE2) and
   b. SNPs with MAF < 1% (the allele frequency of an imputed SNP can be computed as the average of dosage values for all subjects in the sample divided by 2; if this value is > 0.5, subtract it from 1 to get the MAF).
   c. Any SNPs mapping to *sex chromosomes or mitochondria*.

Covariates

Include the following covariates in each analysis:

1. **age**, **sex** (code male=0, female=1) **[do not adjust for BMI]**
2. **field center** (for multi-center studies; create n-1 dichotomous covariates where n= number of field centers), and
3. **principal components (PCs)** derived using genotyped SNPs:
    a. The first PC (and optionally more PCs, if appropriate) for African Ancestry (AA): Each AA cohort will derive the first 10 PCs and will always include the first PC; additional PCs (among the other 9) may also be included if appropriate.
    b. For other race/ethnic groups (optional): Each cohort will decide if a small number of PCs need to be included.
4. Additional **cohort-specific** covariates, if any, for controlling additional confounding.

Contact Information

CONTACT INFORMATION:

DC Rao (rao@wustl.edu); Karen Schwander (karen2@wustl.edu);

Patricia Munroe (p.b.munroe@qmul.ac.uk); Ioanna Ntalla (i.ntalla@qmul.ac.uk)