Obesity Metabolome Scripts Documentation
(December 1, 2020)

Yu-Han Hsu (yuhanhsu@broadinstitute.org)

Overview

The ObesityMetabolome GitHub provides example scripts used to perform the main analyses in the paper:

> Hsu, YH, Astley, CM *et al*. Combining untargeted metabolomics, human genetics, causal inference, and pathway enrichment to define the obesity metabolome. *Int J Obes* (2020). [link]

A brief description of all scripts and examples of how to run them are provided in this documentation. All input and output files are tab- or space-delimited plain text files, except for plots that are stored in PDF format. All scripts were tested using R v3.6.2 or Python v2.7.16; additional packages required to run each script are indicated within the script documentation below. We refer to **Figure 1** in the paper to indicate where the scripts fit into our overall analysis plan. For analyses performed using pre-existing or external software, please refer to their respective publication and/or documentation: PAIRUP-MS, EPACTS, METAL, PLINK, BOLT-LMM, and MR-PRESSO.

This GitHub also hosts meta-analyzed GWAS summary statistics with $p < 1 \times 10^{-5}$ for the BMI-associated metabolites in the *DataForPaper* sub-directory.

Table of contents

Identifying BMI-associated metabolites (Figure 1a)

R script: IdentifyBmiMetabolites.r

Description:

(1) Calculate association between BMI and shared known or matched metabolites across two datasets using linear regression and inverse variance-weighted meta-analysis
(2) Generate clustered correlation heat map of the BMI-associated metabolites

Required packages (with tested versions): meta (v4.11-0), RColorBrewer (v1.1-2), gplots (v3.0.3)

Example input files:

File Name	Description
example/SharedKnowns.txt	Names of shared known metabolites in two datasets
example/MatchedMets.txt	Names of PAIRUP-MS matched metabolites in two datasets
example/MetData1.txt	Sample x metabolite abundance z-score data for first dataset
example/BmiData1.txt	Sample x BMI z-score data for first dataset
example/MetData2.txt	Sample x metabolite abundance z-score data for second dataset
example/BmiData2.txt	Sample x BMI z-score data for second dataset

Example command:

```
$ Rscript IdentifyBmiMetabolites.r example/SharedKnowns.txt example/MatchedMets.txt \
example/MetData1.txt example/BmiData1.txt example/MetData2.txt example/BmiData2.txt \
out_MetBmiStats.txt out_MetCorrHeatmap.pdf
```

Example output files:

File Name	Description
out_MetBmiStats.txt	Metabolite-BMI association statistics (i.e. effect size β, standard error SE, and *p*-value from linear regression of BMI on metabolite) in the two datasets + inverse variance-weighted meta-analysis
out_MetCorrHeatmap.pdf	Clustered correlation heat map of BMI-associated metabolites in the meta-analysis (at Bonferroni significance of $p < 0.05/m$, where m = total number of shared known + matched metabolites)

Calculating BMI genetic risk score for use as G_B (Figure 1b)

Python script: CalculateBmiGrs.py

Description:

Calculate effect size-weighted BMI genetic risk score (GRS) for samples with imputed genotype data

Required modules: sys

Example input files:

File Name	Description
example/BmiSnps.txt	Table containing BMI SNPs (with chromosome, position, BMI-increasing allele, other allele, and effect size information)
example/SampleIDs.txt	List of sample IDs (corresponding to samples in genotype data below)
example/BmiSnpGenotypes.txt	GEN format imputed genotype data for each BMI SNP: chromosome, SNP ID, position, allele1, and allelle2 columns, followed by genotype probability columns for each sample (3 columns per sample)

Example command:

```
$ python CalculateBmiGrs.py example/BmiSnps.txt example/SampleIDs.txt \
example/BmiSnpGenotypes.txt out_BmiGrsValues.txt
```

Example output files:

File Name	Description
out_BmiGrsValues.txt	Sample ID and corresponding BMI GRS value for all samples

Calculating association between G_B and BMI or BMI-associated metabolites (Figure 1b)

R script: CalculateBmiGrsAssocStats.r

Description:

Calculate association between G_B and BMI or BMI-associated metabolites across two datasets using linear regression and inverse variance-weighted meta-analysis

Required packages (with tested versions): meta (v4.11-0)

Example input files:

File Name	Description
example/BmiMets.txt	Names of BMI-associated metabolites in two datasets
example/GrsData1.txt	Sample x BMI GRS data for first dataset
example/MetData1.txt	Sample x metabolite abundance z-score data for first dataset
example/BmiData1.txt	Sample x BMI z-score data for first dataset
example/GrsData2.txt	Sample x BMI GRS data for second dataset
example/MetData2.txt	Sample x metabolite abundance z-score data for second dataset
example/BmiData2.txt	Sample x BMI z-score data for second dataset

Example command:

```
$ Rscript CalculateBmiGrsAssocStats.r example/BmiMets.txt \
example/GrsData1.txt example/MetData1.txt example/BmiData1.txt \
example/GrsData2.txt example/MetData2.txt example/BmiData2.txt \
out_BmiGrsAssocStats.txt
```

Example output files:

File Name	Description
out_BmiGrsAssocStats.txt	BMI-GRS or metabolite-GRS association statistics (i.e. effect size β, standard error SE, and *p*-value from linear regression of GRS on BMI or GRS on metabolite) in the two datasets + inverse variance-weighted meta-analysis

Calculating Wald ratio IV effect estimate for G_M and G_B (Figure 1b)

Python script: CalculateWaldIvStats.py

Description:

Calculate Wald ratio IV effect estimate and corresponding SE and *p*-value for G_M and G_B (i.e. metabolite and BMI instrument, respectively)

Required modules: sys, math, scipy.stats

Example input files:

File Name	Description
example/ObsAssocStats.txt	Table containing observational association statistics between BMI, metabolites ("MET"), metabolite instrument G_M ("SNP"), and BMI instrument G_B ("GRS"); "BMIvMET" and "METvGRS" statistics were derived using R scripts described above; "METvSNP" and "BMIvSNP" statistics were derived from GWAS software

Example command:

```
$ python CalculateWaldIvStats.py example/ObsAssocStats.txt \
[BMIvGRS_Beta] [BMIvGRS_SE] out_WaldIvStats.txt
```

- **Note:** [BMIvGRS_Beta] and [BMIvGRS_SE] are numeric values corresponding to effect size and standard error of association between BMI and G_B (derived using "CalculateBmiGrsAssocStats.r").

Example output files:

File Name	Description
out_WaldIvStats.txt	Same table as "example/ObsAssocStats.txt", with additional columns containing Wald ratio IV effect estimate, standard error, and *p*-value for G_M ("BMIvSNP_Wald") and G_B ("METvGRS_Wald")

Identifying “cause”, “effect”, and “bidirectional” metabolite groups (Figures 1c)

R script: IdentifyMetaboliteGroups.r

Description:

(1) Generate plot of -$\log_{10}$ G_B IV *p*-value vs. -$\log_{10}$ G_M IV *p*-value for BMI-associated metabolites
(2) Identify "cause", "effect", and "bidirectional" metabolite groups using top and bottom quartile cutoffs of G_B and G_M IV *p*-values

Required packages (with tested versions): ggplot2 (v3.3.0), ggrepel (v0.8.1)

Example input files:

File Name	Description
example/WaldIvStats.txt	Same format as “out_WaldIvStats.txt” described above

Example command:

```
$ Rscript IdentifyMetaboliteGroups.r example/WaldIvStats.txt \
out_CauseMets.txt out_EffectMets.txt out_BidirectionalMets.txt out_MetGroupPlot.pdf
```

Example output files:

File Name	Description
out_CauseMets.txt	List of “cause” metabolites separated by “;” (using names in “Metabolite1” column in “example/WaldIvStats.txt”)
out_EffectMets.txt	List of “effect” metabolites separated by “;”
out_BidirectionalMets.txt	List of “bidirectional” metabolites separated by “;”
out_MetGroupPlot.pdf	Plot of metabolites in -$\log_{10}$ G_B IV *p*-value vs. -$\log_{10}$ G_M IV *p*-value space

Generating null metabolite groups for PAIRUP-MS pathway analysis (Figures 1d)

R script: CreateNullMetaboliteGroups.r

Description:

Perform permutations across two datasets to generate null "cause", "effect", and "bidirectional" metabolite groups for use in PAIRUP-MS pathway analysis

Required packages (with tested versions): meta (v4.11-0)

Example input files:

File Name	Description
example/WaldIvStats_toy.txt	Same format as "out_WaldIvStats.txt" described above
example/BmiData1.txt	Sample x BMI z-score data for first dataset
example/GrsData1.txt	Sample x BMI GRS data for first dataset
example/MetData1.txt	Sample x metabolite abundance z-score data for first dataset
example/SnpData1.txt	Sample x metabolite SNP genotype dosage data for first dataset
example/BmiData2.txt	Sample x BMI z-score data for second dataset
example/GrsData2.txt	Sample x BMI GRS data for second dataset
example/MetData2.txt	Sample x metabolite abundance z-score data for second dataset
example/SnpData2.txt	Sample x metabolite SNP genotype dosage data for second dataset

Example command:

```
$ Rscript CreateNullMetaboliteGroups.r example/WaldIvStats_toy.txt \
example/BmiData1.txt example/GrsData1.txt example/MetData1.txt example/SnpData1.txt \
example/BmiData2.txt example/GrsData2.txt example/MetData2.txt example/SnpData2.txt \
[nPerm] [BMIvGRS_Beta] [BMIvGRS_SE] \
out_NullCauseMets.txt out_NullEffectMets.txt out_NullBidirectionalMets.txt
```

- **Note:** [nPerm] is an integer indicating number of permutations to perform (i.e. number of lists to generate); [BMIvGRS_Beta] and [BMIvGRS_SE] are numeric values corresponding to effect size and standard error of observed association between BMI and G_B.

Example output files:

File Name	Description
out_NullCauseMets.txt	Lists (rows) of null "cause" metabolites separated by ";" (using names in "Metabolite1" column in "example/WaldIvStats_toy.txt")
out_NullEffectMets.txt	Lists (rows) of null "effect" metabolites separated by ";"
out_NullBidirectionalMets.txt	Lists (rows) of null "bidirectional" metabolites separated by ";"

- **Note:** The output files generated using the provided toy example contain empty rows due to small input data size; there are no empty rows when we performed the same analysis using our full datasets. We used the "out_*Mets.txt" output files of "IdentifyMetaboliteGroups.r" and "CreateNullMetaboliteGroups.r" to replace the signal lists generated by Step B-1 of the PAIRUP-MS *PathwayAnalysis* pipeline; other pathway analysis steps were performed as described in PAIRUP-MS documentation.