HEc v0.9

Contents

[Introduction 1](#_Toc108262166)

[Inputs 1](#_Toc108262167)

[Outputs 2](#_Toc108262168)

[Required packages 3](#_Toc108262169)

[Pipeline 3](#_Toc108262170)

[Step 1 - Prepare a configuration file 3](#_Toc108262171)

[Step 2 - Generate Blocked-Bootstrap helper file. 3](#_Toc108262172)

[Step 3 - Convert genotype and phenotype files to GCTA format 4](#_Toc108262173)

[Step 4 – Calculate Genetic Correlations in batches on the cluster 5](#_Toc108262174)

[Step 5 – Combine Genetic Correlations batches 5](#_Toc108262175)

[Step 6 – Generate Figures 5](#_Toc108262176)

[Data Files Example 5](#_Toc108262177)

[Configuration File 6](#_Toc108262178)

## Introduction

This package is an HE implementation to calculate genetic and/or any other correlation between pairs of variables that are related via some relatedness matrices. Essentially this calculation will answer the question: “What fraction of the correlation between variables is due to the relatedness”.

For any significant number of traits these computations require a computing cluster. This version works with the LSF scheduler although any other can be easily implemented. Because of the cluster-related steps, this is organized as a pipeline with several steps that need to be executed one after another.

The general approach is you construct one config file for the whole process and then pass it to the different R scripts in this package. Some of them will generate run-files which are needed to be run on the cluster.

## Inputs

1. R dataframe with phenotypes, covariates and ID field
2. An R matrix (or multiple ones) with colnames=rownames=ID and representing some similarity between subject *i* and *j*
3. Various Parameters (See “Configuration File” section for specific description)

## Outputs

There is one file which ends with “longformat.txt”, it has all the resulting data – the following fields:

1. First Phenotype name
2. Second Phenotype name
3. Total (Phenotypic ) model correlation.
4. Total model correlation pvalue
5. Number of people in each phenotype pair
6. Fractional Genetic Correlation
7. Genetic correlation
8. Genetic correlation p-value (Blocked Bootstrap, SD method)
9. Genetic correlation confidence interval – low bound (Blocked Bootstrap, SD method)
10. Genetic correlation confidence interval – high bound (Blocked Bootstrap, SD method)
11. Genetic correlation p-value, FDR corrected (Blocked Bootstrap, SD method)
12. Genetic correlation 95% confidence interval – low bound (Blocked Bootstrap, quantiles method)
13. Genetic correlation 95% confidence interval – high bound (Blocked Bootstrap, quantiles method)
14. Is Genetic correlation p-value significant for 95% confidence quantiles method (TRUE/FALSE)
15. Fractional Household Correlation
16. Household correlation
17. Household correlation p-value (Blocked Bootstrap, SD method)
18. Household correlation confidence interval – low bound (Blocked Bootstrap, SD method)
19. Household correlation confidence interval – high bound (Blocked Bootstrap, SD method)
20. Household correlation p-value, FDR corrected (Blocked Bootstrap, SD method)
21. Household correlation 95% confidence interval – low bound (Blocked Bootstrap, quantiles method)
22. Household correlation 95% confidence interval – high bound (Blocked Bootstrap, quantiles method)
23. Is Household correlation p-value significant for 95% confidence quantiles method (TRUE/FALSE)
24. Genetic correlation p-value (Fisher method)
25. Genetic correlation confidence interval – low bound (Fisher method)
26. Genetic correlation confidence interval – high bound (Fisher method)
27. Genetic correlation p-value, FDR corrected (Fisher method)
28. Household correlation p-value (Fisher method)
29. Household correlation confidence interval – low bound (Fisher method)
30. Household correlation confidence interval – high bound (Fisher method)
31. Household correlation p-value, FDR corrected (Fisher method)
32. Residual correlation

This is the main results file, however the software will also generate each of these 32 as a matrix file with rows and columns being phenotypes and the corresponding values. These are used for figures generations later on.

In addition, three more are generated with

1. Heritabilities (Genetic)
2. Household variance components (analogue for heritabilities for Household distance matrix)
3. PDF with figures for all of the above

## Required packages

Make sure you have these libraries installed:  
R – 3.6.3 or higher; RColorBrewer, corrplot, plotrix, ggchicklet, qgraph, corpcor, gdata, reshape2, ggpattern, igraph, ggplot2, config, gtools, tidyr, tidyverse, hash, R.utils, GENESIS, psych, dplyr

## Pipeline

The main steps are:

1. Prepare data files (phenotypes and relatedness matrices), decide upon covariates.
2. Prepare the configuration file
3. Convert the data to GCTA format and run variance components estimation (Heritability in the case of genetics) in batches on the cluster
4. Run the Genetic Correlation estimation in batches on the cluster
5. Combine the Genetic Correlation batches and Heritability estimates to the final dataset
6. Generate Figures

### Step 1 - Prepare a configuration file

This configuration file is used by all scripts in the pipeline. See “Configuration File” section for specific description.

Example:  
/data/myProject/Code/runGenCor\_ALL\_Base\_v1.config

### Step 2 - Generate Blocked-Bootstrap helper file.

Run: “/data/myProject/Code/1\_Generate\_blocked\_bootstrap\_distmatr\_v2.R”  
with the config file as a parameter. Very fast. Two ways you can do it.

First way - from interactive shell :  
[user@eris1n3 ~]$ bsub -Is -q interact-big -R "rusage[mem=128000]" -n 8 csh

Job <341649> is submitted to queue <interact-big>.

<<Waiting for dispatch ...>>

<<Starting on celeste>>

[user@celeste ~]$ Rscript /data/myProject/Code/1\_Generate\_blocked\_bootstrap\_distmatr\_v2.R /data/myProject/Code/runGenCor\_ALL\_Base\_v1.config

Or create a script file: “/data/myProject/Code/runBootstrap\_v1.sh”

#!/bin/sh

#BSUB -o /data/myProject/DataProc/output\_bootstrap\_cluster.out

#BSUB -e /data/myProject/ DataProc/output\_bootstrap\_cluster.err

#BSUB -q normal

#BSUB -R "rusage[mem=30000]"

#BSUB -J bootstrap1

module load R/4.0.2

Rscript /data/myProject/Code/1\_Generate\_blocked\_bootstrap\_distmatr\_v2.R /data/myProject/Code/runGenCor\_ALL\_Base\_v1.config

And run it:

[user@eris1n3 ~]$ bsub < /data/myProject/Code/runBootstrap\_v1.sh

### Step 3 - Convert genotype and phenotype files to GCTA format

This is done to calculate heritabilities. Run from command line: “/data/myProject/Code/2\_convert\_grm\_RData\_GCTA\_v2.R”  
with the config file as a parameter (Same as in Step 2).  
\*\*Note – you need to this for every phenotype file, but different phenotype files can use the same genotype file (if the genotype file includes all the participants from the phenotype file). The genotype conversion is slow - multiple minutes. So you can run it once for the combined model and then just use it for the sex-specific modules.

1. Step (3) will generate a shell-script file which name you define in the config file under:  
   “run\_GCTA\_script\_filename”  
   You now need to run it on the cluster from command line. So if the file is:  
   /data/myProject/DataProc/run\_GCTA\_VCs\_reml\_2mat\_Base\_All.sh  
   run this line via the command line.  
   [user@eris1n3 ~]$ /data/myProject/DataProc/run\_GCTA\_VCs\_reml\_2mat\_Base\_All.sh  
     
    This will submit as many jobs as you have phenotypes in the file. Need to wait for it to finish. May take up to several hours depending on cluster availability. Wait till all finish.
2. Some runs may fail. You will need to check which failed and re-run them. Failing, meaning a missing “.hst” file. The way you do it is via   
   /data/myProject/Code/3\_rerun\_miising\_GCTA\_v1.R  
   with the config file as a parameter.
   1. Step (3-2) will generate a shell-script file which name you define in the config file under:  
      “rerun\_GCTA\_script\_filename”  
      You now need to run it on the cluster from command line (Same as Step 3-1). So if the file is:  
      /data/myProject/DataProc/rerun\_GCTA\_VCs\_reml\_2mat\_Base\_All.sh  
      run this line via the command line. This will submit as many jobs as have failed in the previous step. Need to wait for it to finish. May take up to several hours depending on cluster availability. Wait till all finish.

### Step 4 – Calculate Genetic Correlations in batches on the cluster

Now we get to the actual genetic correlation calculation. First run   
/data/myProject/Code/4\_HEc\_distrib\_Kinship\_Household\_v9.R  
with the config file as a parameter (Same as Step 2). Very fast.

* 1. This will generate a shell-script file which name you define in the config file under:  
     “run\_Gencor\_script\_filename”  
     You now need to run it on the cluster from command line. If the file is   
     "/data/myProject/DataProc/runRscript\_distrib\_Kinship\_Household\_BaseModel\_All\_v9.sh"  
     then you run it like this:  
     [user@eris1n3 ~]$ bsub < /data/myProject/DataProc/runRscript\_distrib\_Kinship\_Household\_BaseModel\_All\_v9.sh  
     This will submit as many jobs as you have genetic correlations in the file. Need to wait for it to finish. May take minutes to days depending on cluster availability. Wait till all finish

### Step 5 – Combine Genetic Correlations batches

Run from command line: “/data/myProject/Code/5\_HEc\_combine\_Kinship\_Household\_v9.R”  
with the config file as a parameter (Same as Step 2). Pretty fast

### Step 6 – Generate Figures

Run from command line: “/data/myProject/Code/7\_make\_figure\_gencor\_twomat\_v9.R”  
with the config file as a parameter (Same as Step 2). Very fast

## Data Files Example

Phenotypes file:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | SEX | AGE | Phenotype1 | Phenotype2 |
| 1523 | Female | 57 | 1 | 32 |
| 1100 | Male | 48 | 3 | 24 |
| 1052 | Female | 56 | 13 | 25 |
| 1102 | Female | 45 | 15 | 25 |
| 1973 | Female | 50 | 19 | 24 |
| 1026 | Female | 53 | 18 | 30 |
| 1558 | Male | 58 | 5 | 22 |
| 1268 | Male | 46 | 3 | 33 |
| 1553 | Female | 56 | 6 | 21 |
| 1130 | Female | 67 | 6 | 15 |

Relatedness Matrix:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1523** | **1100** | **1052** | **1102** | **1973** | **1026** | **1558** | **1268** | **1553** | **1130** |
| **1523** | 1 | 0.06 | -0.41 | -0.11 | -1.24 | 0.32 | 1.38 | 2.2 | -0.28 | -1.09 |
| **1100** | 0.05 | 1 | 1.65 | -0.1 | 0.86 | 0.3 | 0.49 | -0.36 | -1.48 | 0.48 |
| **1052** | 0.23 | -0.31 | 1 | -1.65 | -1.05 | 0.12 | 0.33 | 0.25 | -1.8 | -0.38 |
| **1102** | -0.96 | -0.15 | -1.61 | 1 | 1.06 | -0.58 | -0.17 | 1.6 | 1.18 | 1.46 |
| **1973** | 1 | -0.53 | -0.44 | -0.22 | 1 | -0.18 | 0.82 | 0.53 | 1.5 | -0.9 |
| **1026** | -1.11 | 1.58 | -0.68 | -0.33 | 0.1 | 1 | -1 | 0.38 | 2.04 | 1.58 |
| **1558** | -1.18 | -0.35 | 1.4 | 0.48 | -0.84 | -0.04 | 1 | 0.55 | -0.01 | -1.06 |
| **1268** | 1.01 | 1.97 | -0.28 | -2.08 | -1.65 | -0.53 | -0.34 | 1 | -0.92 | -1.67 |
| **1553** | -0.07 | -1.01 | -0.31 | -0.05 | -0.56 | 0.54 | -0.31 | -1.1 | 1 | 0.82 |
| **1130** | 1.1 | -1.29 | 0.4 | -0.22 | -0.02 | -2.07 | 1.42 | -0.74 | -2.26 | 1 |

## Configuration File

# please keep this structure

default:

# Name of this config file

this\_config\_filename : "/myProject/Code/runGenCor\_ALL\_Base\_v1.config"

# cluster parameters - name of the queue you want to use

LSF\_QUEUE\_NAME : "normal"

# cluster parameters - amount of memory available in queue

LSF\_MEMORY : "30000"

# relatedness matrices names

kinship\_matrix\_filename : "/myProject/data/genetic\_cor\_kinship\_matrix.RData"

household\_matrix\_filename : "/myProject/data/genetic\_cor\_household\_matrix.RData"

# phenotypes file name – has to include ID, covariates and the phenotypes themselves

phenotypes\_filename : /myProject/data/genetic\_cor\_phenotype\_baseline.RData

# list of non-quantitative covariates

factor\_covariates : ["SEX","CENTER","Education"]

# list of quantitative covariates

numeric\_covariates : ["AGE","weights\_baseline\_log","EV1","EV2","EV3","EV4","EV5"]

# Name of ID field

idfield\_name : "SUBJECT\_ID"

# list of phenotype fields

phenotypes\_fieldnames : ["SLPA54","WHIIRS","SLPDUR","ESS","EDS","global\_cog\_score","TOTAL\_6ITEM","SEVLT\_3TRIALS","SEVLT\_RECALL","WORDFREQ","DIGITSYMBOL","short\_slp","long\_slp","AHI\_GE15","SDB5","Avg\_event\_length","SLPA92","SLPA91","SLPA97","SLEA4","SLEA5","SLEA6","SLEA7","SLEA8"]

#relatedness parameters - used for blocked\_bootstrap, heritabilities

relatedness\_distance : 0.0625 #3rd degree relatives

#what phenotypes compare to what other ones. Group1 <-> Group2

phenotypes\_group1 : ["SLPA54","WHIIRS","SLPDUR","ESS","EDS","short\_slp","long\_slp","AHI\_GE15","SDB5","Avg\_event\_length","SLPA92","SLPA91","SLPA97","SLEA4","SLEA5","SLEA6","SLEA7","SLEA8"]

phenotypes\_group1\_Name : "Sleep"

phenotypes\_group2 : ["global\_cog\_score","TOTAL\_6ITEM","SEVLT\_3TRIALS","SEVLT\_RECALL","WORDFREQ","DIGITSYMBOL"]

phenotypes\_group2\_Name : "Cognition"

# Genetic Correlation - provide directory name for output files

workdirname : "/myProject/output/"

# Genetic Correlation - provide basis for file names to be generated

outfileprefix : "GenCor\_CogSleep\_newKinship\_Household\_Model2\_All\_v8"

# Number of genetic correlations to calculate per batch

ONEBATCH : 2

# Number of repeats for Bootstrap per correlation. determines final possible pvalue. If 0 then only Fisher pvalues will be calculated

BOOTSTRAP\_REPEATS : 100

####filtering and figures

# calculate order by clustering for corrplots. If not will use same order as in "phenotypes"

CALC\_ORDER : False

# calculate display order for qgraphs

CALC\_ORDER\_QGRAPH : True

QGRAPH\_LAYOUT\_FILE : ""

# do not display any correlations with pvalue less than this

pvalueThreshold : 0.05

# do not display any correlations calculated for less people than this

peopleNumThreshold : 1000

# do not display any correlations with absolute value less than this

correlationThreshold : 0

#Filename for script to run on the cluster - GCTA Heritability calculations

run\_GCTA\_script\_filename : "/myProject/main/run\_GCTA\_VCs\_reml\_2mat\_Base\_All.sh"

#Filename for script to run on the cluster - Distributed Genetic Correlation calculation

run\_Gencor\_script\_filename : "/myProject/main/runRscript\_distrib\_Kinship\_Household\_BaseModel\_All\_v8.sh"

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

## These are mainly internal filenames

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

# blocked lookup generation - provide file names to be generated

kinship\_matrix\_filename\_blocked\_lookup : "/myProject/main/genetic\_cor\_kinship\_matrix\_Base\_All\_lookup.RData"

household\_matrix\_filename\_blocked\_lookup : "/myProject/main/genetic\_cor\_household\_matrix\_Base\_All\_lookup.RData"

#convert data to GCTA format from R data - provide file names to be generated

CONVERT\_PHENOTYPES\_toGCTA : True

phenotypes\_GCTA\_prefix : "/myProject/main/genetic\_cor\_phenotype\_baseline\_GCTA"

rerun\_GCTA\_script\_filename : "/myProject/main/rerun\_GCTA\_VCs\_reml\_2mat\_Base\_All.sh"

GCTA\_cluster\_filename\_prefix : "/myProject/main/GCTA\_VCs\_pheno\_NoRel3rdDeg\_PCs5\_2mat\_Base\_All"

CONVERT\_MATRICES\_toGCTA : False #if this is false, matrices generation will be skipped

Unrelated\_individuals\_filename : "/myProject/main/genetic\_cor\_kinship\_matrix\_Unrelated3rdDeg.RData"

kinship\_matrix\_filename\_GCTA : "/myProject/main/genetic\_cor\_kinship\_matrix\_NoRel3rdDeg.GCTA"

household\_matrix\_filename\_GCTA : "/myProject/main/genetic\_cor\_household\_matrix\_NoRel3rdDeg.GCTA"

two\_matrix\_GCTA\_filename : "/myProject/main/20210511\_genetic\_cor\_phenotype\_baseline\_GCTA\_mgrm\_K\_HH.txt"

# Prtial correlations - need to provide precalculated pairs and group file. These files can be created after regular run

PARTIAL : False

pairfile : ""

grpfile : ""