

2025 STT/EPI 855: HW3 (GWAS) Due Nov. 4, 2025

Problem Statement

For this HW, you are asked to perform a Genome-Wide Association Study (GWAS) for HDL cholesterol using a mice data set from the Wellcome Trust.

Objective

Identify genomic regions harboring SNPs significantly associated with HDL in mice.

Expected analyses

You should conduct a comprehensive study, covering all the critical steps needed to achieve the objective of the study—please refer to the class materials, including the papers provided, to define what would be needed to perform a good GWA study.

Expected report

Present a professional report. Limit the main report to 6 pages. Include in your report the following sections

- Introduction (<1 page)
- Summary of data and methods (brief, no more than 1. pages, use Supplementary Methods if needed)
- Results (this should take most of the report, highlight the main results, use Supplementary Results, more below, to present results that are important but not central to the main conclusions)
- Discussion (no more than 1 page)
- Conclusions
- References (not counted within the 6 page limit)
- Supplementary materials, including Supplementary Methods and Supplementary Results (i.e., supplementary tables, and supplementary figures) won't be counted within the 6 limit. Include in the Supplementary Results only tables and figures that are cited in the main report.

Data

You can use the following script to load the data into an R-environment

```
library(BGLR)

data(mice)

X=mice.X
A=mice.A
PHENO=mice.pheno
rownames(PHENO)=rownames(X)

tmp=!is.na(PHENO$Biochem.HDL)
```

```

X=mice.X[tmp,]
PHENO=PHENO[tmp,]

PHENO$y=scale(PHENO$Biochem.HDL,scale=TRUE,center=TRUE) # I suggest you scale the phenotype

# Computing a genomic relationship matrix
G=tcrossprod(scale(X,center=TRUE,scale=FALSE))
G=G/mean(diag(G))

```

Models

I suggest you consider performing GWAS accounting for sex, age at which the HDL was measured, and confounding due to family relationships.

Software

You can use the software of your choice (no need to use more than one software for the same task, for the same model, results should be relatively similar across software).

Below you have a few examples of software than can be used to estimate variance components (needed for GLS) and to perform genome-wide association analyses.

```

# Four ways to estimate heritability using REML

# GENESIS
library(GENESIS)
fm0.Genesis<-fitNullModel(PHENO,outcome='y',covar=c('Biochem.Age','GENDER'),cov.mat=G, family = "gauss")

```

Examples of R-software to estimate variance components

Computing Variance Component Estimates...

```

## Sigma^2_A      log-lik      RSS
## [1]      0.5000000      0.5000000 -1760.6303605      0.7325131
## [1]      0.2806522      0.3277879 -1760.6799240      1.1555619
## [1]      0.3340990      0.3669587 -1760.5207423      1.0186994
## [1]      0.3460109      0.3719476 -1760.5078493      1.0003662
## [1]      0.3465541      0.3719599 -1760.5077887      1.0000003

```

```

# check
fm0.Genesis$varComp

```

```

##          V_A V_resid.var
## 0.3465541 0.3719599

```

```

# RR-BLUP
library(rrBLUP)
Z=model.matrix(~Biochem.Age+factor(GENDER),data=PHENO)
fm0.rrBLUP<-mixed.solve(y=PHENO$y, X=Z, K=G, method="REML",SE=TRUE, return.Hinv=FALSE)
#check
c(fm0.rrBLUP$Ve , fm0.rrBLUP$Vu)

```

```
## [1] 0.3719566 0.3465655
# Custom code
source('https://raw.githubusercontent.com/gdgc/STAT_GEN/refs/heads/main/HW/HW3_GWAS/h2_ML_REML.R')
fm0.Custom<- fitREML(y=PHENO$y,X=Z,K=G,computeHessian=TRUE)
#check
fm0.Custom$estimates

##      varE      varU      h2
## 0.3725158 0.3420305 0.4786680

# Bayesian (takes a bit longer)
LP=list(fixed=list(X=Z,model='FIXED'),grm=list(K=G,model='RKHS'))
fm0.Bayes=BGLR(y=PHENO$y,ETA=LP,nIter=12000,burnIn=2000,verbose=FALSE)
# check
c(fm0.Bayes$varE,fm0.Bayes$ETA$grm$varU)

## [1] 0.3738455 0.3492640

unlink('*.dat')
```

Examples of R-software to perform GWAS

To conduct your GWA analyses you can choose the software you like (note: we will evaluate your choice of models, not the software used).

Here are a few examples GLS single-SNP association testing

Custom code (Please do not distribute it, this code was developing for instruction purposes)

```
source('https://raw.githubusercontent.com/gdgc/STAT_GEN/refs/heads/main/HW/HW3_GWAS/GWAS_OLS_GLS.R')

# y: phenotype, Z: incidence matrix for fixed effects, X: SNPs, G: GRM or pedigree-relationship,
# varComp: a vector with varU and varE from a mixed effect model.
# y, Z, X, and G are assumed to be match (i.e., same subjects and in order). This is not checked by the
system.time(GWAS.GLS<-SMR.GLS(y=PHENO$y,Z=Z,X=X,G=G,varComp=fm0.Genesis$varComp))

##      user      system elapsed
## 41.750      0.729    42.677

# for varComp, one can use any of the above estimates but varComp must have varU and varE, in this ord
```

Genesis

This package can be used for estimation of variance components and GWAS using GLS and OLS (see third web address). This package has several dependencies.

- **Bioconductor site:** <https://bioconductor.org/packages/release/bioc/html/GENESIS.html>
- **Github:** <https://github.com/UW-GAC/GENESIS>
- **Examples:** https://bioconductor.org/packages/release/bioc/vignettes/GENESIS/inst/doc/assoc_test.htm

rrBLUP

This package can be used for estimation of variance components and GWAS using GLS and OLS

<https://cran.r-project.org/web/packages/rrBLUP/refman/rrBLUP.html>

See `mixed.solve()` for variance components estimation and `GWAS()` for GWA analyses.

plink

<https://www.cog-genomics.org/plink/>

GCTA can be used for both variance components estimation and GWAS.

<https://yanglab.westlake.edu.cn/software/gcta/>

qqman

To display your GWAS results I recommend using **qqman** R-package. You can find a vignette here

``https://cran.r-project.org/web/packages/qqman/vignettes/qqman.html``