2022 Genetics and Genomics Winter School Module 4: Practical 10

Valentin Hivert

Genome wide complex trait analysis (GCTA)

GCTA software is a command line tool that has the most support on the Linux operating system. The GCTA program uses an argument interface similar to PLINK. It is advised that you visit GCTA's website http://cnsgenomics.com/software/gcta/ and familiarise yourself a little with the software. GCTA has many functions but one of its primary uses is variance component estimation via restricted maximum likelihood (REML). GCTA is available for Linux, MAC and Windows environments.

Similarly to previous practicals, you will most likely use GCTA in a Linux environment (e.g. HPC), thus all command lines in this material are for Linux users.

Data Use Agreement

- To maximize your learning experience, we will be working with genuine human genetic data.Access to this data requires agreement to the following in to comply with human genetic data ethics regulations.
- Please email pctgadmin@imb.uq.edu.au to confirm that you agree with the following: "I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts."

Cluster Access

You have all been provided with login details to computing resources needed for the practical component * An SSH terminal is needed to connect to the computing:

- Windows: Install PuTTY
  - Hostname: as provided (203.101.228.xxx)
  - User: as provided
  - Check Connection > SSH > X11 > Enable X11 forwarding
- Mac/Linux: Use the terminal
  - ssh -X @203.101.228.xxx

Data and Objective of the practical

We will use the genotype and phenotype data stored in **/data/module4/prac10**. These data will be in PLINK binary format. We will first attempt to use the SNP marker data to build genetic relationship matrices and estimate the proportion of phenotypic variance explained by genome wide SNPs additive and dominance variance in a sample of unrelated individuals. Finally, we will estimate the genetic correlation between two traits using bivariate GREML.

PART I: Estimation of non-additive genetic variance from a sample of unrelated individuals

In this part, you will perform a multi-component GREML analysis to jointly estimate the additive and dominance genetic variance of a trait from a sample of unrelated individuals. The different steps to perform the analysis are detailed below.

Since you will most likely use GCTA in a Linux environment (e.g. HPC), all command lines in this material are for Linux users.

Because of time constraints, we already performed steps 1 to 4 and you should only run the commands from step 5.

Step 1: Data QC

Standard Data QC (see Module 1: GWAS) must be done. The data used in this practical are already QCed and consist of 11,780 individuals genotyped at 273,469 SNPs.

Step 2: Build the additive GRM and extract unrelated individuals (Please do not perform this step in practical)

In this practical, we use 0.05 as a grm cut-off to retain unrelated individuals. In practice, we use a more stringent 0.025 cut-off when estimating non-additive genetic effects.

```
input_file="/data/module4/prac10/Data/ToyExample"
gcta --bfile ${input_file} --make-grm --out ToyExample
```

The Additive GRMs is now created and consist of three output files:

- ToyExample.grm.bin (it is a binary file which contains the lower triangle elements of the GRM)
- ToyExample.grm.N.bin (it is a binary file which contains the number of SNPs used to calculate the GRM)
- ToyExample.grm.id (no header line; columns are family ID and individual ID)

We now identify unrelated individuals (at GRM threshold 0.05) from the GRM values:

```
gcta --grm ToyExample --grm-singleton 0.05 --out ToyExample_GRM05
```

11,730 unrelated individuals at GRM threshold 0.05 have been identified and the list of unrelated IDs is stored in the ToyExample_GRM05.singleton.txt file. We then use this list to recompute the additive GRMs with unrelated individuals only.

```
input_file="/data/module4/prac10/Data/ToyExample"
gcta --bfile ${input_file} --keep ToyExample_GRM05.singleton.txt --make-grm --out ToyExample_GRM05
```

Note: it is possible to directly extract a subset of the GRM with unrelated individuals only using the –grm-cut-off option of gcta. However, this can be problematic if sample allele frequencies are significantly changed after removing related individuals. Therefore, it is recommended to recompute the GRM with the subset of unrelated individuals only.

Step 3: Build the dominance GRM (Please do not perform this step in practical)

Compute the dominance GRM for unrelated individuals.

```
input_file="/data/module4/prac10/Data/ToyExample"
gcta --bfile ${input_file} --keep ToyExample_GRM05.singleton.txt --make-grm-d --out ToyExample_GRM05
```

The three dominance GRMs output files have prefix ToyExample_GRM05.d.

Step 4: Write the mgrm file that contain paths and prefix of the GRMs to analyse (Please do not perform this step in practical)

```
echo -e $(pwd)"/ToyExample_GRM05\n"$(pwd)"/ToyExample_GRM05.d" > mgrm.txt
```

format of mgrm.txt: no headline; each line represents the prefix of a GRM file with its complete path. Note that the previous command assume that your current working directory is the folder where both GRMs are stored.

Step 5: GREML analysis and joint estimation of the proportion of phenotypic variance explained by the additive and dominance genetic variance

```
mkdir prac10 # Create a directory where you will save the analysis output
cd prac10
mgrm="/data/module4/prac10/GRMs/mgrm.txt "
pheno="/data/module4/prac10/Data/ToyExample.pheno"
gcta --mgrm ${mgrm} --pheno ${pheno} --reml --reml-no-constrain --threads 3 --out analysis_output
```

This GREML analysis will generate two files:

- analysis_output.log: the log file with all information about the GREML analysis that you performed
- analysis_output.hsq: rows are:
  - header line;
  - name of genetic variance, estimate and standard error (SE);
  - residual variance, estimate and SE;
  - phenotypic variance, estimate and SE;
  - **ratio of genetic variance to phenotypic variance, estimate and SE**;
  - log-likelihood;
  - sample size.

  If there are multiple GRMs included in the GREML analysis, there will be multiple rows for the genetic variance (as well as their ratios to phenotypic variance) with the names of V(1), V(2), … .

Questions

Question 1: Given the sample size N=11,730 and the observed GRMs, what are the theoretical expectations of the standard errors (SEs) of the estimates for $\widehat{h^2_{SNP}}$ and $\widehat{\delta^2_{SNP}}$?

Note: You can read the GRM binary file in R using the following R function:

```
# R script to read the GRM binary file
ReadGRMBin=function(prefix, AllN=F, size=4){
  sum_i=function(i){
    return(sum(1:i))
  }
  BinFileName=paste(prefix,".grm.bin",sep="")
  NFileName=paste(prefix,".grm.N.bin",sep="")
  IDFileName=paste(prefix,".grm.id",sep="")
  id = read.table(IDFileName)
  n=dim(id)[1]
  BinFile=file(BinFileName, "rb");
  grm=readBin(BinFile, n=n*(n+1)/2, what=numeric(0), size=size)
  NFile=file(NFileName, "rb");
  if(AllN==T){
    N=readBin(NFile, n=n*(n+1)/2, what=numeric(0), size=size)
  }
  else N=readBin(NFile, n=1, what=numeric(0), size=size)
  i=sapply(1:n, sum_i)
  return(list(diag=grm[i], off=grm[-i], id=id, N=N))
}
```

The function return a list of 4 objects:

- diag (the diagonal elements of the GRM)
- off (the off-diagonal elements of the GRM)
- id (individual and family IDs)
- N (the number of markers used in the GRM)

Question 2: What are the estimated h^2_{SNP} and δ^2_{SNP} from the GREML analysis?

Question 3: Are the reported SEs of the estimates in accordance with theoretical expectations? What are the 95% confidence intervals (CIs) of the estimates?

Answers

Question 1: On the server, open R and load the GRMs binary files and compute the variance of the off-diagonal elements. Theoretical standard errors of the narrow-sense and dominance heritability can be computed given these values and the sample size N=11,730:

```
#In R: load the function ReadGRMBin.

ReadGRMBin=function(prefix, AllN=F, size=4){
  sum_i=function(i){
    return(sum(1:i))
  }
  BinFileName=paste(prefix,".grm.bin",sep="")
  NFileName=paste(prefix,".grm.N.bin",sep="")
  IDFileName=paste(prefix,".grm.id",sep="")
  id = read.table(IDFileName)
  n=dim(id)[1]
  BinFile=file(BinFileName, "rb");
  grm=readBin(BinFile, n=n*(n+1)/2, what=numeric(0), size=size)
  NFile=file(NFileName, "rb");
  if(AllN==T){
    N=readBin(NFile, n=n*(n+1)/2, what=numeric(0), size=size)
  }
  else N=readBin(NFile, n=1, what=numeric(0), size=size)
  i=sapply(1:n, sum_i)
  return(list(diag=grm[i], off=grm[-i], id=id, N=N))
}

#Load the two GRMs
GRMA=ReadGRMBin(prefix="/data/module4/prac10/GRMs/ToyExample_GRM05")
GRMD=ReadGRMBin(prefix="/data/module4/prac10/GRMs/ToyExample_GRM05.d")

#Calculate the variance of the off-diagonal elements of the two GRMs
var_GRMA=var(GRMA$off)
var_GRMD=var(GRMD$off)

#Compute the theoretical standard error of the estimates given these values and the sample size N=11,73
```

```
N=11730
theo_se_h2=sqrt(2/(var_GRMA * N^2)) #0.02824041
theo_se_d2=sqrt(2/(var_GRMD * N^2)) #0.04019642
```

Theoretical expectation of the standard error of the estimates are 0.028 for $SE(\widehat{h^2_{SNP}})$ and 0.040 for $SE(\widehat{\delta^2_{SNP}})$.

Question 2: the estimated SNP-based heritability from GCTA is 0.46 and dominance heritability is 0.13.

Question 3: The reported standard errors of the narrow sense and dominance heritability are respectively 0.032 and 0.040, in accordance with their theoretical expectations. The 95% CIs are computed as $estimate + -1.96 \times SE$, resulting in $CI95\ \widehat{h^2_{SNP}} = [0.41; 0.53]$ and $CI95\ \widehat{\delta^2_{SNP}} = [0.05; 0.20]$.

PART II: Estimation of genetic correlation between two traits: Bivariate-GREML

In this part, you will perform a bivariate GREML analysis with GCTA to estimate the genetic correlation between two simulated traits using a sample of unrelated individuals.

Repeat step 1 and 2 from PART I

Perform bivariate GREML analysis to estimate the genetic correlation between two traits

```
#Run bivariate GREML
gcta --reml-bivar --reml-bivar-no-constrain --pheno /data/module4/prac10/Data/bivariateREML.pheno --grm
```

Questions

Question 1: How many individuals are included in the analysis for traits 1 and 2?

Question 2: What are the estimated SNP-based heritability and genetic correlation of the two traits? What are the 95% confidence intervals?

Answers

Question 1: From the gcta log file, 3000 individuals are phenotyped for traits 1 and 2 with a complete sample overlap.

Question 2: The estimated SNP-based heritability are 0.59 and 0.39 for trait 1 and 2 respectively. The estimated genetic correlation between the two traits is 0.66. The 95% confidence intervals are $CI95\ \widehat{h^2_{SNP\_Trait1}} = [0.36; 0.83]$, $CI95\ \widehat{h^2_{SNP\_Trait2}} = [0.15; 0.62]$ and $CI95\ rg = [0.44; 0.88]$