Practical 4 - Genome wide complex trait analysis (GCTA)

Data

We will again use the genotype and phenotype data from practical 3, which includes the height and serum transferrin levels. These data will come again in PLINK binary format. We will attempt to use the SNP marker data to build a genetic relationship matrix and estimate the proportion of phenotypic variance explained by genome wide SNPs. Chapters 26 and 27 of Lynch et al. (1998) go over the underlying methodology of the method used in this practical.

GCTA introduction

In this practical we will use the GCTA software (Yang et al., 2011, 2010) (developed at Peter Visscher's lab <cnsgenomics.com/>) to estimate variance components (from the model on slide 14 of lecture 3), where our data will be SNP markers from the whole genome or a chromosome. This 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 (perhaps for five minutes). GCTA has many functions but one of its primary uses is variance component estimation via restricted maximum likelihood (REML).

Today we will take a look at some of its primary functions – building the SNP genetic relationship matrix (GRM), and variance component analysis. In this practical we will use GCTA to do genomic restricted maximum likelihood to estimate variance components (GREML). This will be done for the height and serum transferrin levels data seen in the previous practical. We will interrogate these results using R.

We begin with building the GRM for the data we used in the previous practical. If you are interested in the conceptual basis for the GRM built with GCTA please refer to Yang et al. (2010). Peak RAM usage for building the GRM is high and thus it is likely that your process may fail if your computer does not have enough resources (around 8GB of RAM). If your system does not have this amount of RAM, we will use a much smaller subset of the data to build the GRM as an exercise; the real GRM is already stored in your practical4/data folder and will be used in subsequent analyses. Many of the flags that you have seen in PLINK are present in GCTA. Remember that if you need help with the syntax for GCTA please take a look at the url <http://cnsgenomics.com/software/gcta/>. We will not attempt to build the GRM as the run time on a good computer is approximately ten minutes. If you would like to build the GRM in your spare time listings 1 and 2 should work on an 8GB RAM machine and a 2GB RAM machine respectively. IMPORTANT - Remember to export your path again.

```
$ gcta --bfile practical_4/data/QIMRX_cleaned --make-grm --autosome --out practical_4/results/QIMRX
```

Listing 1 Building a GRM with GCTA

```
$ gcta --bfile practical_4/data/QIMRX_cleaned_small --make-grm --autosome
$          --out practical_4/results/QIMRX_small
```

Listing 2 Building a smaller GRM with GCTA

GCTA prints output to the console whilst processing the GRM and saves this information to a .log file. Some of the key summary statistics of the GRM built above are seen below.

```
$ Summary of the GRM:
$ Mean of diagonals = 1.00083
$ Variance of diagonals = 9.96466e-05
$ Mean of off-diagonals = -0.000206112
$ Variance of off-diagonals = 4.47574e-05
```

Listing 3 Summary of GRM

This process generates two files QIMRX.grm.gz and QIMRX.grm.id (for older versions on Mac and Windows) or a binary version QIMRX.grm.bin with auxiliary file QIMRX.grm.N.bin (on Linux). Let's take a look at these files in R. Alternatively you can just use head practical_4/data/QIMRX.grm.gz from the command line.

```
> # Read in the gzipped GRM file
> grm <- read.table("practical_4/data/QIMRX.grm.gz")
> head(grm)
  V1 V2    V3          V4
  1  1 265805 0.982474300
  2  1 265771 0.430762300
  2  2 265791 0.997155800
  3  1 261504 0.001788883
  3  2 261492 0.001014439
  3  3 261529 1.000038000
```

Listing 4 Take a look at the GRM

The gzipped GRM is stored in row form with each row having four elements. The first two columns correspond to the (i, j) position of the lower triangular matrix, the third column is the number of non missing SNPs for this row-column calculation, and the fourth contains an estimate of the genetic relatedness.

Estimating proportion of phenotypic variation due to additive genetic factors using GCTA

Let's use the GRM matrix to estimate the proportion of phenotypic variance explained by additive genome-wide SNPs for height and serum transferrin. Open the terminal or command prompt and execute the following command

```
$ gcta --grm practical_4/data/QIMRX  --pheno practical_4/data/HT_T_X.pheno --mpheno 1
$          --reml --out practical_4/results/QIMRX_1
$ gcta --grm practical_4/data/QIMRX  --pheno practical_4/data/HT_T_X.pheno --mpheno 2
$          --reml --out practical_4/results/QIMRX_2
```

Listing 5 Estimating variance components via GREML

We will use R to take a look at the output files that GCTA has calculated. Follow the listing below to read in the files and answer the following questions

```
> # Read in GREML result files
> hsq.1 <- read.table("practical_4/results/QIMRX_1.hsq",
                      header = T, fill = T)
> hsq.2 <- read.table("practical_4/results/QIMRX_2.hsq",
                      header = T, fill = T)
> head(hsq.1)
  Source  Variance  SE
  V(G)    0.795498  0.051003
  V(e)    0.238877  0.041478
  Vp      1.034375  0.028005
  V(G)/Vp 0.769062  0.040840
  logL  -1416.193
  logL0 -1466.335
  LRT 100.283
  df  1
  Pval  0
  n 2836
```

Listing 6 GCTA `.hsq` file in R

If you prefer the command line you can do this in one line at the command line

```
$ # Look at the heritability file provided you are in SISG_AQG_2015 folder
$ cat practical_4/results/QIMRX_1.hsq
  Source  Variance  SE
  V(G)    0.795498  0.051003
  V(e)    0.238877  0.041478
  Vp      1.034375  0.028005
  V(G)/Vp 0.769062  0.040840
  logL  -1416.193
  logL0 -1466.335
  LRT 100.283
  df  1
  Pval  0
  n 2836
```

Listing 7 Command line GCTA `.hsq` file

In the above listing 7 `G` represents genetic, `e` residual, and `p` phenotype.

Exercise 1

- What is the percentage of phenotypic variance that is explained by common SNPs for both traits?
- Are the heritability estimates significant?
- Are the heritability values what you expect?

We will now take a closer look at some of the properties of the GRM by reading using R

```
> # Name the columns of the GRM
> names(grm) <- c("IND_1", "IND_2", "SNP_NUM", "REL")
> dim(grm)
  11817091        4
# Take out the diagonal elements
> grm.diag <- grm[which(grm$IND_1 == grm$IND_2),  ]
> dim(grm.diag)
  4861    4
> head(grm.diag)
  IND_1 IND_2 SNP_NUM       REL
      1     1  265805 0.9824743
      2     2  265791 0.9971558
      3     3  261529 1.0000380
      4     4  265743 1.0015680
      5     5  265796 1.0079730
      6     6  265615 1.0066640
> # Take out the GRM off-diagonal elements
> grm.off.diag <- grm[which(grm$IND_1 != grm$IND_2),  ]
> # Make a histogram of the diagonals
> hist(grm.diag[, 4], breaks = 2500, freq = F,
>      xlab = "GRM diagonals", xlim = c(0.95, 1.2), main = "")
> # Make a histogram of the GRM off-diagonal relatedness estimates
> par(mfrow = c(2, 1))
> hist(grm.off.diag[, 4], breaks = 2500, freq = F,
>      xlab = "GRM off-diagonals", xlim = c(-0.1, 0.1), main = "")
> hist(grm.off.diag[which(grm.off.diag[, 4] > 0.1), 4],
>      breaks = 200, freq = F,
>      xlab = "GRM off-diagonals", xlim = c(0.1, 1.1), main = "")
```

Listing 8 Looking at GRM diagonals and off-diagonals

In the above results the relatedness may have affected the estimate of heritability (Fig. 2 panel 2). We will remove the relatedness and see whether the results change. This is done with `GCTA` via the `--grm-cutoff 0.025` flag and the line in listing 11.

Execise 2

- Repeat the REML analyses as in exercise 1 but with relatedness removed
- Compare the results with those in exercise 1 (using listing 9) and from Yang et al. (2010) (for trait 1 height)

```
> # Read in GREML result files without relatedness.
> grm.nr <- read.table("practical_4/data/QIMRX_nr.grm.gz")
> # Create the same figures as above
> # Note that these are only example file extensions
> # and may change depending on what you want to call the files
> hsq.1.nr <- read.table("practical_4/results/QIMRX_nr_1.hsq",
>                        header = T, fill = T)
> hsq.2.nr <- read.table("practical_4/results/QIMRX_nr_2.hsq",
>                        header = T, fill = T)
```

Listing 9 GCTA GREML `.hsq` result files with no relatives

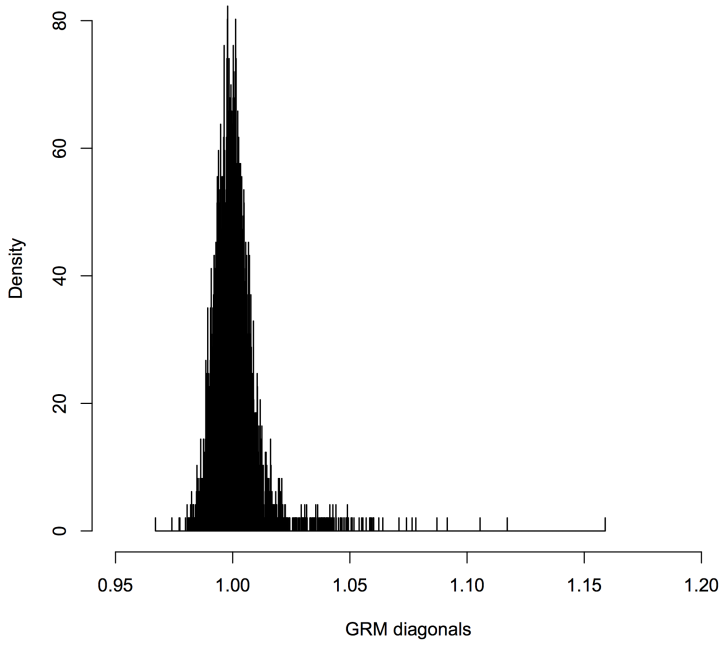

Figure 1 Plot of QIMR GRM diagonals.

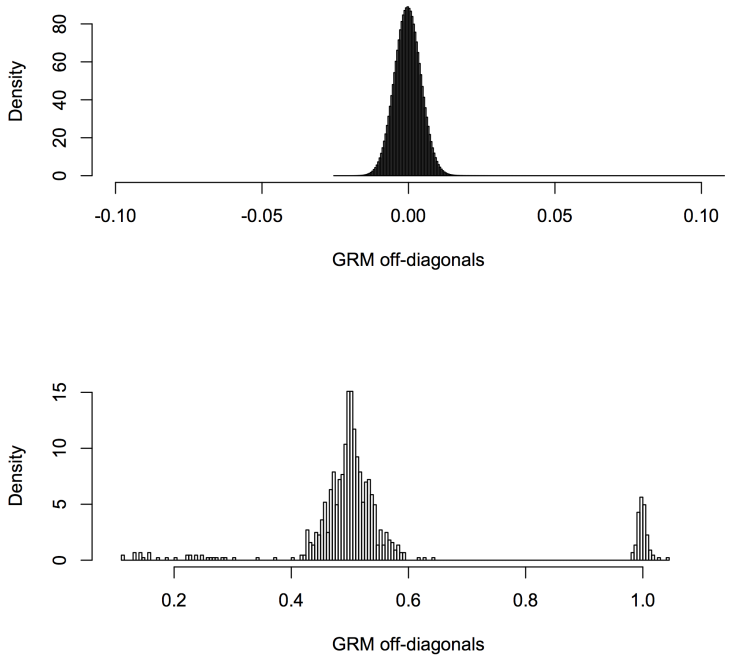

Figure 2 Plot of QIMR GRM off diagonals.

Partitioning the variance via minor allele frequency

We will now investigate partitioning variance components by creating two GRM matrices. One will be created with SNPs with lower MAFs and another with those that have higher MAFs. This will allow us to investigate (in a very imprecise way) the genetic architecture of the trait by trying to understand whether rare or common variants contribute more or less to the proportion of phenotypic variance explained by additive genetic variance (tagged by genome wide SNPs). The SNP files used are located in the `practical_4/data` directory; these files along with the `--extract` flag were used to build the two GRMs (listing 10). In order to filter on relatedness we will keep the individuals from the relatedness thresholded GRM built above and the flag `--keep` (listing 10). The GRMs were built with the command and it is best if you try to build these in your own time as the process is computationally expensive.

```
$ gcta --bfile practical_4/data/QIMRX_cleaned --extract practical_4/data/bot_maf_snps.txt --autosome
$          --make-grm  --keep QIMRX_nr.grm.id --out practical_4/data/QIMRX_bot_maf_snps
$ gcta --bfile practical_4/data/QIMRX_cleaned --extract practical_4/data/top_maf_snps.txt --autosome
$          --make-grm  --keep QIMRX_nr.grm.id --out practical_4/data/QIMRX_top_maf_snps
```

Listing 10 Preparing the GRMs for variance partitioning

Given these two GRMs we can estimate the proportion of phenotypic variance that can be explained by additive genetic variants from low MAF SNPs versus higher MAF SNPs. Attempt to run this listing and answer the following questions

```
$  gcta --mgrm practical_4/data/QIMRX_multi.txt --pheno practical_4/data/HT_T_X.pheno
$           --mpheno 1 --reml --out practical_4/results/QIMRX_1_multi_nr
```

Listing 11 Partitioning variance components via GREML

Exercise 3

- Repeat for phenotype 2
- What do you observe for the different variance component estimates from the GRMs of low and high MAF SNPs?
- Is this what we expect?

References

Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.