Lab 6: Calculating the Heritability of Complex Traits with GCTA

Created by Max Winston & Brandon Pierce; slight modifications by Ittai Eres & Arjun Biddanda

Genome-wide Complex Trait Analysis (GCTA) was originally designed to estimate the heritability of complex traits using genome-wide SNPs, but has now been extended for numerous other functionalities to better understand the genetic architecture of complex traits (Yang et al., 2011). Generally, GCTA estimates heritability using the proportion of phenotypic variance explained by the a genetic relationship matrix (GRM), which is calculated using the genome-wide SNP data. In today's lab we will become familiar with the GCTA software and some of its capabilities, as well as explore some of the conceptual issues dealt with in class with a large SNP dataset provided by Brandon. By the end of the lab, you should be able to:

- **Construct GRMs given BED, FAM, and BIM files.**
- **Run standard analysis in GCTA on large datasets (univariate REML).**
- **Run bivariate REML analysis in GCTA.**
- **Relate how increasing density of markers affects heritability estimate.**
- **Relate how subsetting markers may affect heritability estimate.**

Section 1: Basics of GCTA

1.1: Standard GCTA options

Option	Description
make-grm	Generates GRM from SNP data (BED, FAM, BIM files).
make-grm-bin	Generates binary GRM from SNP data (BED, FAM, BIM files).
bfile	Specifies stem name from BED file for analysis.
out	Specifies output stem name.
pheno	Specifies file containing phenotypic information.
mpheno	Gives the column number within phenotype file to use for analysis.
reml	Runs univariate restricted maximum likelihood (REML) analysis.
grm	Specifies GRM file for analysis.
grm-bin	Specifies binary GRM file for analysis.
reml-bivar	Runs bivariate restricted maximum likelihood (REML) analysis.
reml-maxit	Sets the maximum number of iterations to run (Default: 100).

There is a lot of overlap between PLINK options and GCTA options. For example, the **maf** option does the same thing in GCTA as it does in PLINK.

1.2: Basic GCTA syntax

Running GCTA is nice because it prints lots of useful information to the screen as it runs, and when it concludes. However, due to the computation required for creating a GRM, and the large size of the some of the input and output files, running it can take a bit of time. Depending on what you're doing for this lab, you can expect some processes to take up to 3 minutes, and of course, with bigger files, it would take more time. We will be placing all of our results in the directory called **results**.

The syntax used to run GCTA is similar to other pipelines we have used on the command line: GCTA is called and modified with option flags (see Section 1.1). For example, one of the first things you will usually need to do is take your SNP data (BED, FAM, BIM formats–just like in PLINK!) and make a GRM. **Try this with the following command and the "test" files in the data/test/ directory:**

```
./bin/gcta64 --bfile data/test/test --autosome --maf 0.01 --make-grm --out results/test
```

Problem 1
How many individuals are there in the test dataset?

1.3: Basic REML run

GCTA employs a restricted maximum likelihood (REML) method to estimate the proportion of phenotypic data explained by SNP data (Yang et al., 2010). **Run a basic REML analysis on the "test" GRM you created with the following command:**

```
./bin/gcta64 --grm results/test --pheno data/test/test.phen --reml --out results/test
```

Results can be found in the `results/test.hsq` file. **Open this file in your preferred text editor.** Recall from lecture that narrow-sense heritability is additive genetic variance over phenotypic variance.

Problem 2
Is this phenotypic trait *statistically significantly* heritable?

Problem 3
What is the heritability estimate? What is the standard error of this estimate?

Section 2: Manipulating GRMs for a Robust Assessment of Heritability

2.1: Effect of SNP Density on Heritability Estimation

The density at which SNPs are sampled throughout the genome can have an important effect on heritability estimates of complex traits. In order to illustrate this point, we have provided two large SNP datasets (250k and 500k markers) of 1,000 individuals and constructed a file with two phenotypes, named `two_phenotypes.txt`.

Problem 4
For phenotype 1, create a GRM using 250,000 whole genome SNPs (`shared_data/gcta_shared_files/250k`) to estimate heritability for the first phenotype.Note : *don't forget to specify you want column 1 in the phenotype file used for REML analysis.* (Hint : to save space use the `--make-grm-bin` function)

Problem 5
For phenotype 1, use the GRMs created by 500,000 whole genome SNPs (`shared_data/gcta_shared_files/500k`) to estimate the heritability for the first phenotype.

Problem 6
How did the estimate and/or SE change and why? What might this say about SNP density and estimating heritability more generally?

2.2: Effect of Subsetting a GRM to Causal Variants on Heritability Estimation

Some time-traveling scientists who have conducted GWAS of all humans on earth deliver you a dataset containing only causal variants (causal.bed, causal.fam, causal.bim). Let's explore how using this set of causal variants may change our estimation of heritability.

Problem 7
Create a GRM based only on the causal variant files (`shared_data/gcta_shared_files/causal`) and estimate the heritability.

Problem 8
How did the estimate and/or SE change compared to the REML analysis with the 500K GRM? Why might this be?

2.3: Estimating Genetic Correlation of Multiple Phenotypes

Problem 9
Using the GRM generated from 500k SNPs, estimate the genetic correlation between phenotype 1 and 2 using a bivariate, rather than a univariate, REML run. Genetic correlation between the two will be represented as C(G)_tr12 in the output file. Based on your notes, what is the interpretation of the "genetic correlation" (we did not get to this in class, but give it a shot)?

References

J. Yang, S.H. Lee, M.E. Goddard, P.M. Visscher (2011). GCTA: a tool for Genome-wide Complex Trait Analysis. *American Journal of Human Genetics*. 88(1): 76-82.

J. Yang, B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard, P.M Visscher (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42(7): 565-9.