Whole genomes practical

Gibran Hemani

This practical shows you how to perform various whole genome analyses using GCTA on plink format data. To see scripts that will run on a linux server, visit the github page here:

https://github.com/explodecomputer/WholeGenomesPractical

Background

The use of very simple, single SNP approaches have been very successful in genetic studies. However, with the introduction of whole genome methods the scope of what we might be able to learn from genetic data has broadened significantly. Here we'll look at some of the fundamentals.

The purpose of GWAS is to identify particular SNPs that we are certain have an influence on a trait. In contrast, the purpose of a GCTA style 'GREML' (Genetic REML) or 'SNP heritability' analysis is to estimate how much of the variance of the phenotype can be explained by all the measured SNPs in our data.

The SNP heritability is estimated using a two step procedure. First a genetic covariance matrix, or genetic relationship matrix (GRM) is estimated. This is an n x n matrix where each element represents the genetic similarity of two individuals. The second step performs REML analysis to essentially estimate how much of the phenotypic covariance in the population is attributable to genetic covariance.

A note about software

The original implementation for large scale human data is GCTA. It is continually improving, and it has a huge number of features. We will use this to perform REML estimation of heritabilities. It also constructs genetic relationship matrices, which is something that we need, but we will use Plink2 to do this, as it does the same implementation but much faster.

> GREML analysis can be performed using GCTA. Earlier versions of GCTA were available for Mac, Windows and Linux, but for some time now the updated versions of GCTA have only been made for Linux. A major change is in the format - previously the GRM was stored in a text format, but now it is stored in a much more efficient binary format. If you are using Windows or Mac then you will have to use the text format, which can be quite slow. To compensate, we will only perform the analysis on the first 4000 individuals

Data

We have body mass index (BMI), C-reactive protein (CRP) levels, and hypertension case control status data on each of around 8000 individuals. This is located in `data/phen.txt` We also have covariates, including the first 10 genetic principal components, age, sex, and smoking status (`data/covs.txt`).

To see how this data was QC'd, take a look at the `scripts/qc_phen.R` script. The figures generated from this script are in the `images/` folder.

We also have SNP genotypes for these individuals. Approx 500,000 markers on 23 chromosomes.

Note: All the scripts and phenotype data used for this practical are in this repository. The genotype data can be made available upon request - just ask!

Using SNPs to estimate kinship

How far removed must two individuals be from one another before they are considered 'unrelated'? We can make estimates of the proportion of the genome that is shared identical by descent (IBD) between all pairs of seemingly unrelated individuals from the population by calculating the proportion of SNPs that are identical by state (IBS).

The result is a genetic relationship matrix (GRM, aka kinship matrix) of size n x n, diagonals are estimates of an individual's inbreeding and off-diagonals are an estimate of genomic similarity for pairs of individuals.

To do this using plink with the complete data, and creating the new GRM format, run the following command:

```
../../software/plink_mac --bfile ../../data/genetic/geno_qc --make-grm-bin --out ../../data/kinships/gen
```

To create the old format, and using only the first 4000 individuals, do:

```
../../software/plink_mac --bfile ../../data/genetic/geno_qc --make-grm-gz --out ../../data/kinships/gen
```

Relevant flags:

- `--make-grm-bin` signals that the new format (binary) genetic relationship matrix (GRM) should be constructed from the genetic data
- `--make-grm-gz` signals that the old format (gzipped text file) genetic relationship matrix (GRM) should be constructed from the genetic data
- `--keep` specifies a file where a list of IDs that should be retained is provided.

This creates three files:

- `geno_qc.grm.bin` is the n x n relationship matrix
- `geno_qc.grm.id` is the list of IDs that correspond to the matrix. Two columns - FID and IID, just like in plink.
- `geno_qc.grm.N.bin` is the n x n matrix that contains the numbers of SNPs that were used to calculate each relationship

Note that you can't actually see the `.bin` files - they are binary files that have machine stored numerical values. This makes it much faster to read the data, and makes for much smaller storage space than if they were human readable text files.

We can read this data into R though and explore what it looks like. Using R Studio open up the R script located in `software/analyse_grm.R`.

Using kinships to estimate heritability

See slides for more accurate treatment, but the intuition is as follows. Heritability is the measure of the proportion of variation that is due to genetic variation. If individuals who are more phenotypically similar also tend to be more genetically similar then this is evidence that heritability is non-zero. We can make estimates of heritability by comparing these similarities.

When genetic similarity is calculated by using SNPs then we are no longer estimating heritability per se, we are instead estimating how much of the phenotypic variance can be explained by all the SNPs in our model.

To calculate the SNP heritability of BMI we will use the genetic relationship matrix and the phenotype data. Let's do this both with and without covariates.

```
../../software/gcta_mac --grm ../../data/kinships/precomputed/geno_qc_4000 --pheno ../../data/genetic/p
```

Note that this may take a really long time. Press `ctrl+c` to cancel, the results are precomputed in the `precomputed/` folder

The flags are:

- `--grm` specifies the name of the old format of the GRM data. Here we are using a GRM that was computed using just the first 4000 samples
- `--pheno` specifies the location of the phenotype file
- `--mpheno 1` specifies which phenotype to use
- `--reml --reml-no-lrt` specifies to calculate the SNP heritability using REML, and to not calculate the likelihood ratio test.

If run, this should create a new file called `univariate_bmi_nocovar.hsq`. We can see what it looks like here:

```
Source  Variance        SE
V(G)    0.068817        0.002970
V(e)    0.004801        0.002452
Vp      0.073618        0.001191
V(G)/Vp 0.934781        0.033576
logL    6849.995
n       8227
```

This says the SNP heritability is 0.93 with a fairly low standard error. What happens if we recalculate but this time fitting the covariates?

```
../../software/gcta_mac --grm ../../data/kinships/precomputed/geno_qc_4000 --pheno ../../data/genetic/p
```

Note that this may take a really long time. Press `ctrl+c` to cancel, the results are precomputed in the `precomputed/` folder

This gives the following result:

```
Source  Variance        SE
V(G)    0.020764        0.002727
V(e)    0.040291        0.002662
Vp      0.061055        0.000960
V(G)/Vp 0.340092        0.043672
logL    7390.455
n       8227
```

This is a much lower estimate, suggesting that population structure was leading to a lot of inflation in the heritability estimates when we didn't adjust for principal components.

> Heritability studies from family data range from 0.4 to 0.8, this result is much lower. What is the explanation?

Bivariate heritability

We can estimate the genetic correlation between two traits. This represents the extent to which the genetic effects that influence trait A are correlated with the genetic effects that influence trait B.

In GCTA, if we want to estimate the genetic correlation between BMI and CRP we would use the following command:

```
../../software/gcta_mac --grm ../../data/kinships/precomputed/geno_qc_4000 --pheno ../../data/genetic/p
```

Note that this may take a really long time. Press `ctrl+c` to cancel, the results are precomputed in the `precomputed/` folder

The difference here is that the `--reml-bivar 1 2` flag replaces the `--mpheno` and `--reml` flags. We are now specifying that a bivariate heritability should be performed using the 1st and 5th phenotypes in the phen.txt file.

The results in `bivariate_bmi_crp.hsq` look like this:

```
Source  Variance        SE
V(G)_tr1        0.020853        0.002727
V(G)_tr2        0.501093        0.057464
C(G)_tr12       0.031048        0.009314
V(e)_tr1        0.040211        0.002661
V(e)_tr2        0.776698        0.055370
C(e)_tr12       0.061279        0.009039
Vp_tr1  0.061063        0.000961
Vp_tr2  1.277792        0.020138
V(G)/Vp_tr1     0.341491        0.043667
V(G)/Vp_tr2     0.392156        0.043663
rG      0.303739        0.078016
logL    2773.208
n       16464
```

The genetic correlation is denoted by `rG`, note that the standard error is higher than in the univariate analyses - the power of a bivariate analysis is lower than.

> What does is the biological interpretation of a genetic correlation existing between two traits?

Polygenic architecture

Under a polygenic architecture we hypothesise that the SNP heritability based on one chromosome should be related to the chromosome size - larger chromosomes have larger numbers of causal variants and so should explain more variance.

To test this, we will now construct the genetic relationship matrix using the QC'd genotype data **for only one chromosome**. Choose a chromosome to analyse, and then construct the GRM. Hint: Use the `--chr <x>` flag to specify which chromosome to analyse.

Once the GRM for the chromosome is constructed, calculate the SNP heritability for BMI using your newly created GRM.