Whole genome methods practical

Gibran Hemani

08 June, 2018

This practical shows you how to perform various whole genome analyses using GCTA on plink format data. It will be run on a linux server. All the scripts are available for you to download 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.

Logging in to the server

Log into bluecrystal using PuTTY. Run the following command to access a compute node:

```
qsub -I -q teaching -l nodes=1:ppn=1,walltime=02:00:00
```

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.

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.

Exercises

0. First we need to setup the scripts and programmes to run on our server. We will need Plink2, GCTA and git:

   ```
   module add tools/git-2.18.0
   module add apps/plink-1.90b3v
   module add apps/gcta-1.24.3
   module add languages/R-3.2.4-ATLAS
   ```

   Check that you can run the programs, e.g. try typing `git`, `plink` and `gcta` and make sure that they can run.

1. Now we download the scripts by using git to clone the repository:

   ```
   git clone https://github.com/explodecomputer/WholeGenomesPractical.git
   ```

   This will take a few moments to download. Once it's finished you can see that there is a new directory called `WholeGenomesPractical` by typing `ls -l`

2. We will now construct the genetic relationship matrix using the QC'd genotype data. Choose a chromosome to analyse, and then using a text editor such as `nano` modify the following file to analyse the chromosome that you are interested in:

   ```
   cd WholeGenomesPractical/scripts
   nano construct_grm_chr.sh
   ```

   Change the line `CHR=""` to have your chromosome number e.g. `CHR="5"` for chromosome 5. Save and exit (`ctrl+x`, and then type `y` to save). Now execute the script:

   ```
   ./construct_grm_chr.sh
   ```

   - What is the algorithm doing?

    We are estimating the genetic similarity between each pair of individuals. This is based on average numbers of shared alleles for every SNP on the chromosome being analysed.

3. We have now calculated a genetic relationship value for every pair of individuals. If the sample comprises only 'unrelated' individuals then each pair of individuals should have a genetic relationship less than 0.05 (and a relationship with themselves of approximately 1). Use the `analyse_grm.R` script to read in the GRM files into R and plot the distribution of relationships:

   ```
   R --no-save < analyse_grm.R
   ```

   - Why is it important to make sure that related individuals are not included in this analysis?

     We are estimating SNP heritability. The genetic similarity of related individuals includes all variants not just common SNPs captured on the SNP chip. It also correlates with common environment, so the estimate will be biased towards full heritability + common environment effects.

   - How might these graphs look different if you used the entire genome instead of just one chromosome to calculate relationships?

     The variance would be lower because we are obtaining a more precise estimate of kinship by incorporating information from many more SNPs

4. Calculate SNP heritabilities with and without covariates. What are the SNP heritabilities for each chromosome for BMI? Use the commands in `estimate_heritability.sh` to do this and plot them on the board.

   - How does SNP heritability relate to chromosome size? Why?

     Larger chromosomes have higher heritability because under the infinitestimal model they have more causal variants

   - What is the danger of calculating SNP heritability **without fitting covariates**?

     Confounding from population stratification will inflate the extent to which genetic similarity associates with phenotypic similarity

   - What would happen if you calculated SNP heritability in a case control study, **where the cases and controls were genotyped in separate batches**?

     The technical batch effects would make cases look much more similar to cases than to controls. As a consequence the heritability would be inflated because individuals with the same phenotype would be estimated to be more genetically similar

5. In addition to estimating the SNP heritability of each trait, we can calculate how similar the genetic effects are for a pair of traits. This is also known as the genetic correlation. Perform bivariate GREML analysis to calculate genetic correlations between each pair of traits. Use the commands in `estimate_bivariate.sh` to do this.