MAGMA practical - Boulder workshop 2015

In this short practical we will run through the three basic steps of performing a (MAGMA) gene-set analysis: annotation of SNPs to genes, gene analysis and subsequent gene-set analysis. We will also perform a conditional gene-set analysis to demonstrate the confounding effect of a gene set on the interpretation of our results. Input files are provided alongside this instruction, and consist of

- a PLINK data set containing simulated GWAS data of a more or less realistic size (boulder.bed, boulder.bim, boulder.fam)
- a covariate file containing 10 PCs to correct for population stratification (boulder_pca.cov)
- a gene-set file containing 1011 Reactome gene sets (reactome.sets)
- a gene definition file (NCBI37.3.gene.loc)

Step 1: annotation

In this step we will annotate the SNPs in the data to genes. To do so, use the command:

```
magma --annotate --snp-loc boulder.bim --gene-loc NCBI37.3.gene.loc
--out step1
```

In this case we are using the .bim file from the data itself to specify the chromosome and location for each SNP, though it is also possible to use an external file for this (for example when you are using an older data set and the SNP locations are on a different Human Genome build than your gene location file). The gene locations in the NCBI37.3.gene.loc file have been retrieved from the NCBI servers, and contain the Entrez ID, chromosome, and transcription start and stop site for each gene. MAGMA will only annotate SNPs to a gene if they fall between the provided start and stop site, unless a window around the gene is explicitly specified. The command will produce a log file and the file step1.genes.annot containing the gene annotations, with each line corresponding to a gene.

Question: how many gene definitions were in the NCBI file and how many genes have ended up in the .genes.annot file? What caused this difference, and how might this affect the gene-set analysis?

Step 2: gene analysis

Gene-set analysis in MAGMA has an explicit two-step structure, the first of which is to perform a gene analysis. In this case we will perform an analysis on raw GWAS data using the command:

```
magma --bfile boulder --covar file=boulder_pca.cov --gene-annot
step1.genes.annot --out step2
```

The data is stored in binary PLINK format in the three boulder.xxx files, and the 10 PCs are provided in the boulder_pca.cov file. By default all covariates in the file will be used, unless a subset is explicitly specified. Note that MAGMA will also by default include gender (as encoded in the .fam file) in the analysis for chromosome X genes, though for this simulated data all individuals have been set to female (you will find a warning in the log to this effect as well).

Aside from the log file, this command will produce two files: step2.genes.out and step2.genes.raw. The .genes.out file contains the results of the gene analysis formatted to be easy to read. Use the 'head' or 'more' command to look at the contents of the file and see if the content is clear to you. To study the results in more detail, you can load the file in R using the read.table() function. The .genes.raw file contains largely the same information as the .genes.out file, plus additional information on the correlation between genes. This file is intended as input for the gene-set analysis.

Question: how many genes are significant after Bonferroni correction, at an alpha of 5%? What percentage of the genes has a p-value below 0.05? Tip: in R you can take the mean of a boolean vector (a vector with TRUE and FALSE values) to obtain the percentage of TRUE values in that vecter, for example: mean(vec < 0.05)

Step 3: gene-set analysis

Having performed the gene analysis, we can now do the actual gene-set analysis:

```
magma --gene-results step2.genes.raw --set-annot reactome.sets --out
step3
```

The gene sets in the reactome.sets file are stored in a row-based format, with each row containing the name of the gene set followed by a list of gene names. In this case Entrez gene IDs are used, though as long as the names in the set file match the names in the gene definition file from step 1 the choice of nomenclature does not matter. Note that steps 2 and 3 can also be performed in one command, by adding the --set-annot flag to the command from step 2. This will still produce a .genes.raw file to serve as input for later analyses.

This step produces two files of interest, the step3.sets.out file and the step3.setgenes.out file. The first file contains the main gene-set analysis results, and can again be loaded into R using read.table(). For the sake of reference MAGMA does provide a self-contained p-value for each gene-set, though its use is generally discouraged and the competitive p-values should be used for determining the results of the analysis. For all competitively significant gene-sets (after Bonferroni correction, at an alpha of 5%) the gene analysis results for the genes in those sets are also provided in the .setgenes.out file. Each gene-set is tagged with a numbered set ID of the form _SET1_, _SET2_, etc. so you can use the grep command to extract the genes for a specific gene-set (eg. `grep _SET5_ step3.setgenes.out > set5.genes` to extract the genes for the fifth significant gene-set and store them in the file set5.genes). Files with gene results extracted this way can again be read into R using read.table(), since lines starting with a '#' are automatically skipped.

Question: how many genes are significant in the self-contained analysis (after Bonferroni correction, alpha of 5%), and how many in the competitive analysis? Are there any gene-sets significant in the competitive analysis but not in the self-contained analysis?

Extract the gene analysis results for one of the sets from the .setgenes.out file. Are any of the genes significant at the genome-wide level? What percentage of the genes has a p-value below 0.05? Is this higher than the percentage you find for the data set as a whole in step 2?

Step 4: conditional gene-set analysis

The reactome.sets file contains one strongly associated gene-set conveniently named CONFOUNDER, which acts as a confounder for some of the other strongly associated gene-sets in the analysis. In this final step we will perform a conditional gene-set analysis to demonstrate this effect:

```
magma --gene-results step2.genes.raw --set-annot reactome.sets
condition=CONFOUNDER --out step4
```

This step produces the same kind of output files as in step 3, except that the COMP_P column now contains the conditional competitive p-values and the CONFOUNDER gene-set itself is omitted from the output. Note that the self-contained analysis is not changed. Compare the output from the regular analysis in step 3 with those from the conditional analysis in this step by loading both .sets.out files in R and looking at the results for the gene-sets that were significant in step 3. Tip: all those significant gene-sets (plus one other) have a self-contained p-value lower than 1e-6, and since the order is the same in both files you can easily select and compare the relevant gene sets by extracting those with SELF_P < 1e-6.

Question: how does conditioning on the CONFOUNDER gene set affect the results, how many gene sets remain significant? Are there are any gene sets for which the association has become stronger by conditioning on CONFOUNDER?

Prologue

Congratulations, you have now performed your first (MAGMA) gene-set analysis! If you would like to know more: the accompanying paper will soon be published in PLOS Computational Biology, and the program, manual and related files will be made available on the MAGMA website at http://ctglab.nl/software/magma (note: manual and website are currently still undergoing maintenance). For further questions and suggestions you can email Christiaan de Leeuw at cdeleeuw@science.ru.nl.