MAGMA practical

6th Mar. 2019

In this 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. Additionally, we will also perform a number of more advanced analyses using the generalized gene-set analysis framework that MAGMA provides, including conditional and joint gene-set analysis, gene-set interaction analysis, as well as analysis of tissue-specific gene expression levels.
Note: *All text in grey is optional to read: it contains details on how MAGMA works but is not essential for the current practical*

First extract magma_practial.zip and *cd* into the "magma_practial" folder. There are 2 folders named "files" and "output". The "files" folder contains input files as below. The "output" folder is empty.

- a gene-set file containing 1013 Reactome gene sets (reactome.sets)
- a gene covariate file containing tissue-specific gene expression levels per gene for 11 tissues, simulated based on real expression data (tissue_gex.cov)
- two additional auxiliary files (step3a.signif, step6a.partitioned.sets)

Some of the large files listed below are located under /data/magma/files. You do not need to copy them to the working directory, but you will specify the path to these files in your commands.

- a PLINK data set containing simulated GWAS data of a small but more or less realistic size (practical.bed, practical.bim, practical.fam)
- a file containing 10 PCs to correct for population stratification, to be included as covariates in the gene analysis (practical.cov)
- a gene definition file (NCBI37.3.gene.loc)

Note: all the input files except practical.bed are plain-text files so if you want to inspect their content you can do so either in a text editor or using a UNIX command (e.g. head, less or more). It is recommended to use the computers provided in the workshop as MAGMA v1.07 is already installed. If you use your own laptop, you need to install MAGMA from https://ctg.cncr.nl/software/magma (v1.07).

Step 1: annotation

We first need to tell MAGMA which SNPs belong to which genes, so step 1 is to annotate the SNPs in the provided data to genes. To do so, use the command:

```
magma --annotate window=1,0.5 \
      --snp-loc /data/magma/files/practical.bim \
      --gene-loc /data/magma/files/NCBI37.3.gene.loc \
      --out output/step1
```

A SNP is mapped to a gene if it is located either inside the transcription region of the gene, or in a window around it. In this case we specify the window to reach up to 1 kilobase upstream of the transcription start site, and 0.5 kilobases downstream of the transcription stop site.

The --snp-loc flag specifies which file to use to read the SNP locations from, and the --gene-loc flag specifies the file that defines the gene locations. The latter contains one row per gene, with the values: gene ID, chromosome, transcription start and stop site (in base-pair position), genomic strand (this relates to the direction in which the gene is transcribed: "front to back" or "back to front"; for genes on the negative strand, the transcription start site is the higher of the two base-pair values), and official gene symbol.

Running this command will create the file step1.genes.annot, containing the mapping of SNPs to genes. This will be used as an input file for the gene analysis. Each row in the file corresponds to a gene, containing: the gene ID, the mapping region (chromosome:start:stop), and then the list of SNP IDs mapped to that gene. A step1.log file will also be created, containing the output that was also printed to the screen. It provides you with useful information about the steps that were performed (such as the number of values read from input files or printed to output files), as well as any warnings and errors that occurred during execution.

Question 1: *how many gene definitions were in the gene location file and how many genes have ended up in the .genes.annot file? What caused this difference, and how do you think this could affect the gene-set analysis?*

Step 2: gene analysis

In this step we will run a gene analysis, performing a test of association for each gene and creating the input file needed for subsequent gene-set analyses. We will do so using the command:

```
magma --bfile /data/magma/files/practical \
      --covar file=/data/magma/files/practical.cov \
      --gene-annot output/step1.genes.annot \
      --out output/step2
```

The --bfile flag specifies the prefix of a binary PLINK file set (.bed, .bim and .fam). Because we are using raw genotype data as input, the default principal components linear regression model will be used for the analysis. It will use the phenotype embedded in the .fam file as the dependent variable, and will also include the variables in the practical.cov file specified using the --covar flag as additional covariates. For genes on the X chromosome, gender will also be included as a covariate. The --gene-annot flag tells MAGMA which SNP-to-gene mapping file to use to determine which genes to analyse and what SNPs they contain.

The **step2.genes.out** file is the main gene analysis output file, and contains the following information: the gene ID (GENE), gene mapping region (CHR, START and STOP), number of valid SNPs mapped to the gene (NSNPS), number of principal components extracted from those SNPs (NPARAM), sample size for that gene (N; in this case it is the same for all genes, but it can vary if there are missing values in the data), the test statistic and corresponding p-value (ZSTAT and P), and the R^2 and adjusted R^2 values (RSQ and RSQ_ADJ; these reflect the proportion of variance in the phenotype explained by the SNPs in that gene).

Question 2: *how many genes are significant after Bonferroni correction (correcting for the total number of genes)? What percentage of the genes has a p-value below 0.05? How would you interpret that, does this indicate a lot of genetic signal in the data to you?*

Step 3a: basic competitive gene-set analysis

Having completed the gene analysis step, we will now perform a competitive gene-set analysis:

```
magma --gene-results output/step2.genes.raw \
      --set-annot files/reactome.sets \
      --out output/step3a
```

The --gene-results flag specifies which gene analysis .genes.raw output file to use as input for the gene-set analysis, and the --set-annot flag refers to the gene-set definition file. In this case we use the reactome.sets file, which contains 1013 gene sets. These are almost all real gene sets taken from various databases, reflecting known biological pathways. A few additional sets were added for the purpose of this practical. The gene sets are stored by row, with each row containing the name of the gene set followed by the list of gene IDs of genes that belong in that set.

With this command MAGMA will analyse each gene set in the reactome.sets file, one at a time. As you will see in the output log, a number of data-level properties of genes (eg. number of SNPs mapped to a gene) are automatically included as covariates in the analyses. In practice not all the genes mapped to a gene set in the reactome.sets will actually be included when analysing that set, because they are not present in the .genes.raw file. This could be because those genes were not included in the gene definition file during annotation or had no SNPs mapped to them; it could also be because all of the SNPs mapped to that gene were either missing from the genotype data, or were invalid (eg. because they had too many missing values).

This command will produce three output files: **step3a.gsa.out, step3a.gsa.genes.out** and **step3a.gsa.sets.genes.out**. The step3a.gsa.out contains the analysis results for all the gene sets, and has the following information: the name of the gene set (VARIABLE and FULL_NAME; the VARIABLE column is a truncated version of the full name, this is intended to make the file easier to read when there are very long variable names), the variable type (TYPE; in this case, all are gene set variables), the number of genes included in the gene set for the analysis (NGENES), and the linear regression parameters (BETA, BETA_STD, SE) and corresponding p-value (P). The BETA value is the actual model parameter as discussed in the lecture (with SE its standard error). BETA_STD is a standardized coefficient, dividing BETA by the standard deviation of the gene set (generally larger for larger gene sets). This can be useful for comparing the effect size of different gene sets.

The step3a.gsa.genes.out file contains information per gene for all the genes used in the analysis, and is very similar to the .genes.out file from step 2. The step3a.gsa.sets.genes.out file contains information per gene for significant gene sets (determined using Bonferroni correction for the total number of gene sets analysed). It contains mostly the same columns as the step2.genes.out file, in separate blocks for each of the significant gene sets. This is useful for better understanding the genes and associations of those genes in a significant set.

Question 3a: *how many gene sets are significant in the gene-set analysis (after Bonferroni correction for the total number of analysed sets)? How do you interpret a significant result for a gene set in a*

competitive analysis like this, what do you conclude from the fact that for example SIGNALING_BY_NOTCH1_T is significant?
Inspect the gene analysis results for the SIGNALING_BY_NOTCH1_T set in the .gsa.sets.genes.out file. Are any of the genes significant at the genome-wide level (ie. Bonferroni-corrected for the total number of genes in the data)? 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? Do you think the genes with p-value greater than 0.05 still contribute to the gene-set association?

Step 3b: conditional gene-set analysis

The reactome.sets file contains a very strongly associated gene-set helpfully labelled CRITICAL_PATHWAY. Gene sets often overlap with each other, and it is possible that some gene sets are significant simply because they overlap with this CRITICAL_PATHWAY. We will therefore run a conditional gene-set analysis to test whether this is the case here for any of the other significant gene sets. The command to do so is:

```
magma --gene-results output/step2.genes.raw \
      --set-annot files/reactome.sets \
      --model analyse=file,files/step3a.signif condition=CRITICAL_PATHWAY \
      --out output/step3b
```

(Note: this will likely print a "WARNING: analysis failed for 'CRITICAL_PATHWAY'" in the log and add two lines with NA p-values in the .gsa.out file; this is a minor bug and can be ignored, and the analysis will run normally)

The 'analyse' option of the --model flag tells MAGMA to only analyse a selection of gene sets, in this case all the gene sets listed in the step3a.signif file. This file lists all the significant gene sets from step 3a, for convenience this has already been created for you. With the 'condition' option we tell MAGMA that CRITICAL_PATHWAY should be included as an additional covariate in the gene-set analysis. As such, for each of the gene sets to be analysed (ie. those listed in step3a.signif), MAGMA will use a linear regression model containing two gene set variables: the gene set to be analysed, and the CRITICAL_PATHWAY gene set.

The output files from this step are of the same kind as in step3a, the only difference is that now the .gsa.out file contains an additional MODEL column. Each row still corresponds to the results of a single gene set, so this MODEL column tells you which rows belong together in the same regression model. The parameter estimates and p-value therefore reflect the strength of the gene set effect when the other gene sets in the same model are taken into account. So for example, in this case model 1 will contain both CRITICAL_PATHWAY and SIGNALING_BY_NOTCH1_T, and the results for the SIGNALING_BY_NOTCH1_T reflect its effect conditional on CRITICAL_PATHWAY. Keep in mind that these multi-variable models are symmetrical: the CRITICAL_PATHWAY result for model 1 thus reflects the effect of CRITICAL_PATHWAY conditional on SIGNALING_BY_NOTCH1_T.

When interpreting results from a conditional analysis, it is always useful to compare the conditional association of a gene set with its marginal association (ie. the association that the variable had before conditioning on the other gene set). This tells you how much of that marginal association could be

explained by the other gene set. To do so we could go back to the step 3a results file, but we can also just rerun that analysis with only the gene sets of interest included:

```
magma --gene-results output/step2.genes.raw \
      --set-annot files/reactome.sets \
      --model analyse=file,files/step3a.signif \
      --out output/step3c
```

Question 3b: *how does conditioning on the CRITICAL_PATHWAY gene set affect the associations of the other gene sets? How many of those gene sets remain significant (at the original Bonferroni-corrected threshold) when the CRITICAL_PATHWAY effect is taken into account? What would you conclude about the gene sets that are no longer significant? Does the CRITICAL_PATHWAY remain significant in all cases? How would you interpret the results from models in which it does not, if any (especially when P>0.05)?*

Step 4a: basic tissue expression analysis

As mentioned briefly in the lecture, continuous gene properties can be analyzed in much the same way as gene sets. Here we will analyse gene expression values (on a log2(RPKM) scale, higher values mean stronger expression) for different tissue types, which can provide insight into the tissue-specificity of our genetic associations. This analysis is run as follows:

```
magma --gene-results output/step2.genes.raw \
      --gene-covar files/tissue_gex.cov \
      --model direction-covar=positive \
      --out output/step4a
```

The --gene-covar flag is used to specify a file containing continuous gene properties, in this case the tissue_gex.cov file containing gene expression values. Each row in the file corresponds to a gene, with the gene ID listed in the first column followed by the all the gene expression variables in subsequent columns. The file contains expression variables for eleven different tissues, as well as a twelfth variable containing the mean expression across all the tissues.

As when analysing the gene sets, this command will analyse each of the expression variables one at a time. The 'direction-covar' option sets the direction of the test that is performed. In this case we are testing whether the effect of the expression variable is positive.

The command will generate a **step4a.gsa.out** output file, which has all the same columns as the .gsa.out file from step 3a. Because the variables are continuous, the NGENES column is set to the total number of genes in the analysis. You will see that this is actually a few hundred genes less than before, this is because for some of the genes no gene expression data was available (this is quite common with such data). Those genes were therefore discarded from the analysis.

Question 4a: *how do you interpret a significant result for a continuous gene property in an analysis like this, what do you conclude from the fact that for example BRAIN_EXPR is significant? We performed a*

one-sided test for positive association, do you think testing for negative associations would also be useful? How would you interpret a significant negative association for one of these tissue expression variables?

How many tissue expression levels are significantly (positively) associated with the genetic associations (after Bonferroni correction for all tissue variables)? Taking all the results together, do you think they are very informative about the phenotype?

Step 4b: conditional tissue expression analysis

As with the gene-set analysis, conditional analysis can again be used to obtain more specificity in our results. In this case, the significance of AVERAGE_EXPR in the previous step show us that in general, more strongly expressed genes also tend to be more strongly associated with our phenotype. This means that associations found for the specific tissue expression levels may simply reflect this general relation, rather than saying anything specifically about the expression in that tissue type. In this step we will therefore condition on the average gene expression level per gene to obtain associations that are specific to individual tissue expression levels:

```
magma --gene-results output/step2.genes.raw \
      --gene-covar files/tissue_gex.cov \
      --model direction-covar=positive condition-hide=AVERAGE_EXPR \
      --out output/step4b
```

In this case we are using the 'condition-hide' option rather than 'condition', this suppresses output for AVERAGE_EXPR in the step4b.gsa.out file. This does not otherwise affect the results of the analysis, but since we are not very interested here in the output for AVERAGE_EXPR itself in these models for now it is helpful for making the output file easier to read.

Question 4b: *how many tissue expression levels remain significant (at the original threshold) now that we have accounted for the overall average effect of gene expression? Taken together, what do you conclude from the results of step 4a and 4b?*

Step 5: joint analysis of gene sets and tissue expression levels

Confounding and overlap of association signals can of course also happen between gene sets and continuous gene properties, and with gene expression variables, it is not uncommon for this to happen. For example, for psychiatric phenotypes like schizophrenia there is often a strong association with brain-specific expression levels. Any gene set that happens to contain many strongly brain-expressed genes is therefore more likely to be significant as well, even if the underlying pathway or biological process has nothing to do with schizophrenia. We can again use conditional analysis to account for this.

For our example data we will do so in two steps, first correcting for the effect of average expression, then additionally correcting for the effect of brain-specific expression as well.

```
magma --gene-results output/step2.genes.raw \
      --set-annot files/reactome.sets \
      --gene-covar files/tissue_gex.cov \
      --model analyse=file,files/step3a.signif \
      condition-hide=AVERAGE_EXPR \
      --out output/step5a
```

```
magma --gene-results output/step2.genes.raw \
      --set-annot files/reactome.sets \
      --gene-covar files/tissue_gex.cov \
      --model analyse=file,files/step3a.signif \
      condition-hide=AVERAGE_EXPR,BRAIN_EXPR \
      --out output/step5b
```

We are only re-analysing the gene sets that were previously significant, as the aim is only to investigate whether those significant associations may have been the result of confounding caused by gene expression effects. We are again using the 'condition-hide' option to make the output files a bit tidier, since we are only interested now in what happens to the associations of the gene sets, not those of the expression variables.

Question 5: *how strongly are the gene-set p-values affected by conditioning on the general gene expression levels? And when you also condition on the brain-specific expression? How would you interpret these results?*

Step 6: tissue by gene set interaction analysis

In a normal gene set analysis we implicitly assume that most or all genes in a particular gene set are relevant to the phenotype we are investigating. In practice however, this may not be the case. For example, the gene sets we use typically reflect general biological functions and processes whereas most phenotypes exhibit tissue-specific involvement of genes. Even if a particular biological function is relevant to that phenotype, this may be restricted to genes that are more strongly expressed in a specific tissue. Statistically, this would likely manifest as an interaction between expression in that tissue and the gene set.

In our example, brain expression was significantly associated, so we will test for interactions between brain expression and all large gene sets (interactions between continuous variables and gene sets are often unstable for small gene sets; we will leave the cut-off at the default value of at least 100 genes in the gene set).

```
magma --gene-results output/step2.genes.raw \
      --set-annot files/reactome.sets \
      --gene-covar files/tissue_gex.cov \
      --model analyse=sets direction-interaction=positive \
      interaction-each=BRAIN_EXPR \
      --out output/step6a
```

With the 'analyse=sets' option we can specify that we are only interested in analysing the gene sets, and not in any of the other tissue expression variables in the tissue_gex.cov file. We are also setting the testing direction to test for positive interaction effects. With the 'interaction-each' option set, MAGMA will now perform an interaction analysis with BRAIN_EXPR for each of the gene sets (provided they contain at least 100 genes). Each interaction model will be a linear regression containing three variables: the BRAIN_EXPR tissue expression variable, a gene set variable, and the interaction of BRAIN_EXPR with that gene set.

A **step6a.gsa.out** file will be created containing the analysis results, with the same structure as in step 3b. The file again contains a MODEL column which tells you which rows go together in the same model; for this analysis it will be three rows per model, two for the main effects and one for the interaction. In this case we are only interested in the p-values for the interaction terms, denoted by INTER-SC in the TYPE column.

Question 6.1: *is there a significant interactions in the output? If there is, how do you interpret the significant result, what do you conclude from it about the effect gene set involved in that interaction? Was this gene set significant in the earlier gene-set analysis?*

With a significant tissue by gene set interaction, it can be useful to perform a follow-up analysis of the gene set partitioned by expression in the tissue. To do so, the genes in the gene set are first sorted from lowest to highest expression. The set is then subdivided into smaller sets based on this ordering. This can make the results of the analysis somewhat easier to interpret, and can help protect against possible outlier effects (which the interaction analysis can be sensitive to).

In this case a step6a.partitioned.sets has already been provided for this for the significant interaction from step 6a. It contains four subsets corresponding to the partitioned gene set, from lowest expression in Q1 to highest expression in Q4. The full gene set is also included in this file, for comparison. We will analyse these partitions as normal gene sets, conditioning on the brain expression.

```
magma --gene-results output/step2.genes.raw \
      --set-annot files/step6a.partitioned.sets \
      --gene-covar files/tissue_gex.cov \
      --model analyse=sets condition-hide=BRAIN_EXPR \
      --out output/step6b
```

Question 6.2: *how do you interpret the results of this additional analysis? Does it further clarify the significant interaction and explain why the gene set wasn't significant on its own?*

Prologue

The MAGMA software and auxiliary files, as well as the manual and reference to the accompanying papers, can be found on the MAGMA website at http://ctglab.nl/software/magma. For further questions and suggestions you can also email Christiaan de Leeuw at c.a.de.leeuw@vu.nl.