MAGMA practical - answers

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, as well as analysis of tissue-specific gene expression levels. Input files are provided alongside this instruction, and consist of:

* the MAGMA v1.10 executable (magma)
* the MAGMA manual (manual_v1.10.pdf)
* a PLINK data set containing simulated GWAS data of a small but more or less realistic size
  (magma_practical.bed, magma_practical.bim, magma_practical.fam)
* a file containing 10 PCs to correct for population stratification, to be included as covariates in
  the gene analysis (magma_practical.cov)
* 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)
* a gene definition file (NCBI37.3.gene.loc)
* two additional auxiliary files (step3a.signif, step6a.partitioned.sets)

Notes
* To determine significance in this practical, we will use the traditional significance threshold of 0.05. We will use Bonferroni correction to account for multiple testing. As a reminder, a test is significant after Bonferroni correction when its p-value is smaller than 0.05/K, where K is the number of tests you want to correct for.

Step 0: set-up
Log in to the remote computer. Create a new folder for today's tutorial on gene- and pathway analysis called thursday_magma and copy the following files into that new folder. Use the following commands:

```
mkdir thursday_magma
cd thursday_magma
cp /home/douglasw/Boulder2025/magma_session.zip .
unzip magma_session.zip
```

The PLINK data used in the practical is located in the /usr/local/data/ folder. We will reference this directly, rather than copy it. For readability, we will store the name of this folder in a variable DATA, as follows:

```
DATA=/usr/local/data/
```

Step 1: annotation
In this step we will annotate the SNPs in the data to genes. To do so, use the command:

```
./magma --annotate window=1,0.5 --snp-loc $DATA/magma_practical.bim --gene-loc input/NCBI37.3.gene.loc --out output/step1
```

This tells MAGMA to perform annotation, mapping SNPs to genes based on the transcription region of each gene. 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.

* Questions:

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?

There were 19,427 gene definitions in the NCBI37.3.gene.loc file; of these, 13,772 ended up in the .genes.annot.txt file (these numbers can be obtained from the output log and .log file for the annotation). The reason the other 5,655 genes were not present in the output is that for those genes none of the SNPs in the data fell inside their transcription region or the +1kb/+0.5kb window around it. This is largely caused by the relatively small number of SNPs; in modern GWAS data,

to which genotype imputation is also always applied, the number of dropped genes is usually only a few hundred.

Since in a gene-set analysis the genes are the data points, this means that effectively the 'sample size' for the gene-set analysis and therefore the power to detect effects will be lowered, and for some gene sets it may not be possible to perform an analysis because all or most of the genes they contain are among those 5,655.

Step 2: gene analysis

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

```
./magma --bfile $DATA/magma_practical --covar file=input/magma_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) that will be used for the analysis. 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.

This command will create two output files: step2.genes.out and step2.genes.raw. The .genes.raw file contains all the information MAGMA needs to perform the analyses in the next steps. It is a plain- text file and you can inspect the content if you want, but it is not really designed to be read by people and all the relevant output is also in the .genes.out file.

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-squared and adjusted r-squared values (RSQ and RSQ_ADJ; these reflect the proportion of variance in the phenotype explained by the SNPs in that gene).

* Questions:

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?

Note: To answer this, you can either load the output file into R and manipulate the resulting dataframe, or you can use Linux commands instead. Remember that google is your friend here. Most coding problems you encounter have been solved by many others before. Dedicated forums usually contain all the answers to your questions.

Here are two stack overflow pages that can help you with this question:

- print values below a threshold: https://stackoverflow.com/questions/6848606/print-only-values-smaller-than-certain-threshold-in-bash
- count lines in output: https://stackoverflow.com/questions/12457457/count-number-of-lines-in-terminal-output

There are two genes significant at the Bonferroni-corrected threshold of $\alpha = 0.05/13{,}772 = 3.63 \times 10^{-6}$. There are 857 genes with a p-value smaller than 0.05, which i s 6.22% of the total. Since by chance we would expect about 5% if there was no genetic signal in the data at all, this suggests a modest amount of genetic signal in the data. Given that the sample size of the GWAS data is only 2,500 individual however, we would generally not expect it to be much higher than this due to lack of statistical power.

```
# show significant genes, with header
awk 'NR == 1 || $9 < 0.05/13772' output/step2.genes.out

# count number of nominally significant genes
awk 'NR > 1 && $9 < 0.05' output/step2.genes.out | wc -l
```

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 input/reactome.sets --out output/step3a
```

The --gene-results flag specifies which gene analysis .genes.raw output file to use to perform the analysis, and the --set-annot flag which 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, using the linear regression framework explained in the lecture. 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. You won't need it for this practical. 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 analyzed). 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.

* Questions:

how many gene sets are significant in the gene-set analysis (after Bonferroni correction for the total number of analyzed 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.txt file.

Are any of the genes significant at the genome-wide level (i.e. 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?

There are ten gene sets significant at the Bonferroni-corrected threshold of $\alpha = 0.05/1{,}013 = 4.94 \times 10^{-5}$. A significant result in a competitive gene-set analysis means that the mean genetic association of genes in the gene set is higher than the mean genetic association among all the other genes in the data (probably; it could of course still be a type 1 error). We would conclude from this that (in this example) there is evidence that the Notch1 signaling pathway plays a role in the genetics of our phenotype.

None of the genes in the gene set are significant, the lowest p-value among them is 3.48×10^{-4}. However, 28.3% (15 out of 53) of genes in the set has a p-value below 0.05, much

more than the 6.22% found in the data as a whole. This shows that the level of association is indeed much higher inside the gene set, than in the rest of the genes. The mean gene p-value in the data is 0.49. The mean p-value among the genes in the set, even when looking at only those with p-values greater than 0.05, is still lower than this, at 0.41. This suggests that at least some of the genes with p-values greater than 0.05 are still positively contributing to the gene-set association.

```
# show significant gene sets, with header
awk '$1 != "#" && ($7 == "P" || $7 < 0.05/1013)' output/step3a.gsa.out

# count number of genes in NOTCH1 set with p-value < 0.05
awk '$1 == "_SET1_" && $10 != "P" && $10 < 0.05' output/step3a.gsa.sets.genes.out | wc -l
```

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 input/reactome.sets --model analyse=file,aux/step3a.signif condition=CRITICAL_PATHWAY --out output/step3b
```

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 beexplained 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 input/reactome.sets --model analyse=file,aux/step3a.signif --out output/step3c

* Questions:

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 do you interpret the results from models in which it does not?

For six of the other gene sets, their p-value when conditioning on CRITICAL_PATHWAY is no longer significant nor even below 0.05 anymore. For 5 of those 6 the p-value for CRITICAL_PATHWAY conditioned on those sets is still significant. This suggests that for those five gene sets, their original competitive p-value is actually the result of confounding: the associations for these gene sets found in step 3a is most likely entirely caused by the fact that they overlap to a considerable degree with CRITICAL_PATHWAY, which has a strong association; the sets do not have a genuine, biologically relevant association.

For the model with ANOTHER_CRITICAL_PATHWAY, the associations of this set and CRITICAL_PATHWAY both disappear entirely. This can happen if there is strong overlap between gene sets, and suggests that the two gene sets are tapping into the same association signal and the model cannot determine which of the two is the more likely source. The stronger the overlap, the greater the change in p-value; in this case the two gene sets almost completely overlap, which explains why in this analysis their originally very low p-values have disappeared entirely. The conclusion we would draw here is that there is a single strong association signal, but we cannot determine which of the two sets is most likely to have the true association. We therefore keep them both, and only interpret them as a pair.

For three of the gene sets the p-value doesn't really change when conditioning on CRITICAL_PATHWAY. For these, we can conclude that their associations are independent of the CRITICAL_PATHWAY association.

Step 4a: basic tissue expression analysis

Continuous gene properties can be analyzed in much the same way as gene sets. In this tutorial 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 input/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.txt 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.

* Questions:
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?

BRAIN_EXPR reflects gene expression measured in the brain, with higher scores denoting stronger expression. We would therefore interpret the positive association for BRAIN_EXPR as suggesting that genes tend to have stronger genetic associations with the phenotype the more strongly they are expressed in the brain. Had the association been negative (and significant), we would have interpreted this as suggesting that genes tend to have weaker genetic associations the more strongly brain-expressed they are (or equivalently, that they tend to have stronger genetic associations the more weakly brain expressed they are). From a biological perspective such a negative effect for an expression variable doesn't seem very plausible, and the one-sided positive test is probably preferable to improve power to detect positive effects.

All the tissues except SKIN_EXPR are significant at the threshold of $\alpha = 0.05/12 = 0.00417$, and so is the overall expression effect AVERAGE_EXPR. Although it is clear from this that gene expression plays a role, because almost everything is significant we cannot draw any more specific conclusions than that.

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 input/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.

* Questions:
  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?

Once we condition on AVERAGE_EXPR, only BRAIN_EXPR remains significant; effects for the other tissues have entirely disappeared, with the lowest p-value among them still well above 0.05. The most likely interpretation of the results from 4a and 4b is that there are two real effects: an overall positive effect of gene expression in AVERAGE_EXPR, and an additional positive brain-specific expression effect in BRAIN_EXPR. Associations found in 4a for the other tissues are the result of confounding. These were not actually specific to those tissues, but simply reflected the overall gene expression effect (as the tissue-specific expression variables are all strongly correlated with AVERAGE_EXPR, and can therefore easily pick up on that overall effect).

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 input/reactome.sets --gene-covar input/tissue_gex.cov --model analyse=file,aux/step3a.signif condition-hide=AVERAGE_EXPR --out output/step5a

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

We are only reanalyzing 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.

* Questions: 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?

In this case the gene-set p-values barely change at all compared to step 3a/3c, in either of the two conditional analyses. This indicates that the associations in these gene sets are independent of the gene expression effects (ie. here we have essentially ruled out confounding by gene expression).