MAGMA practical - answers

Questions step 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?*

There were 19,427 gene definitions in the NCBI37.3.gene.loc file; of these, 13,772 ended up in the .genes.annot 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.

Questions step 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?*

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.

With Linux commands:

```
#show significant genes and header (p-value is in column 9)
awk 'NR == 1 || $9 < 0.05 / 13772' step2.genes.out

#count number of genes with p < 0.05
awk '$9 < 0.05' step2.genes.out | wc -l
```

With R:

```
res = read.table("step2.genes.out", header=T, stringsAsFactors=F)
res[res$P < 0.05/nrow(res),]     #show significant genes
mean(res$P < 0.05)               #compute proportion with p < 0.05
```

NB: if in your terminal window R is using only part of the width of your screen and is cutting off lines and displaying them in part on later lines, you can use `options(width=200)` to adjust the display width setting (you may need to change the width value specified depending on the size and resolution of your screen).

Questions 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?

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.

With Linux commands:

```
#show significant sets and header (p-value is in column 7)
awk '$1 == "VARIABLE" || $7 < 0.05 / 1013' step3a.gsa.out

# extract gene information for NOTCH1 set
grep _SET1_ step3a.gsa.sets.genes.out > step3a.NOTCH1.genes.out

# show genome-wide significant genes in set
awk '$1 != "#" && $10 < 0.05/13772' step3a.NOTCH1.genes.out

# count number of genes with p < 0.05
awk '$1 != "#" && $10 < 0.05' step3a.NOTCH1.genes.out | wc -l
```

With R (create step3a.NOTCH.genes.out file as above first):

```
res = read.table("step3a.gsa.out", header=T, stringsAsFactors=F)
sum(res$P < 0.05/nrow(res))      #count number of significant sets
res[res$P < 0.05/nrow(res),]     #show significant sets
notch = read.table("step3a.NOTCH1.genes.out", header=T, strings=F)
notch[notch$P < 0.05/13772,]     #show genome-wide significant genes
mean(notch$P < 0.05)             #compute proportion with p < 0.05
```

Questions 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?*

For five of the other gene sets, their p-value when conditioning on CRITICAL_PATHWAY is no longer significant nor even below 0.05 anymore, while 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.

```
# for ease of comparison, show output file without CRITICAL_PATHWAY lines
awk '$1 != "CRITICAL_PATHWAY"' step3b.gsa.out
```

Questions 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?

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.

Questions 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?*

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).

Questions 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?*

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).

Questions 6: *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?*

Yes, the interaction term between BRAIN_EXPR and the I_LOVE_BRAINS pathway is highly significant when we correct for the 74 interaction terms tested (you can find this number in the screen output and .log file).

We only tested for positive interactions, so what this result suggests is that specifically the combination of the I_LOVE_BRAINS pathway with higher brain expression is relevant for the genetics of our phenotype. The I_LOVE_BRAINS set showed no signs of association in step 3a, which further suggests that it is probably only that combination with brain expression that is relevant, and there is very little effect in the I_LOVE_BRAINS genes that are more weakly brain-expressed (because otherwise we would have seen at least a somewhat lower p-value in step 3a).

Using Linux commands:

```
# show significant interaction terms
awk '($1 != "#" && $1 == "VARIABLE") || ($2 == "INTER-SC" && $9 < 0.05/74)' step6a.gsa.out

# show all output for the model with significant interaction
awk '($1 != "#" && $1 == "VARIABLE") || $3 == 74' step6a.gsa.out

# show original gene-set analysis result for corresponding gene set
awk '$1 == "VARIABLE" || $1 == "I_LOVE_BRAINS"' step3a.gsa.out
```

With R:

```
res = read.table("step6a.gsa.out", header=T, stringsAsFactors=F)

#show significant interaction, and output for rest of model
res[res$TYPE == "INTER-SC" & res$P < 0.05/74,]
res[res$MODEL == 74,]

# show original gene-set analysis result for corresponding set
sets = read.table("step3a.gsa.out", header=T, stringsAsFactors=F)
sets[sets$VARIABLE == "I_LOVE_BRAINS",]
```

Questions 6 (part 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?*

Consistent with what we concluded above, the analysis of the partitions shows us that there is quite a strong association hidden inside the top 25% brain expressed genes (Q4) in the I_LOVE_BRAINS gene set. There was absolutely no effect found in the other three quarters of the set, which also explains why it wasn't significant in 3a: the associations in Q4 simply got drowned out by the lack of association in the rest of the set.