AD PROJECT SUMMARY

Ian Brettell

2 March 2017

Sample and data

1. Our sample comprises **193 individuals** from the AIBL Study (Australian Imaging, Biomarkers & Lifestyle study of ageing).
2. For that sample, we have the following data:
    a. **Aβ-pos or –neg status** (binary);
    b. Metadata including:
        i. **Age** at the time of data collection (numeric);
        ii. **Sex** (binary);
        iii. The presence of at least one **apoe-e4** allele (binary);
    c. **Gene expression** data for **~22,000 genes** (numeric) based on *blood samples*; and
    d. **Genotype** data for **~2,000 SNPs** known from GWAS to be associated with AD (factor with three levels: 0/0, 0/1, 1/1).

Structure of analysis

3. We carried out the project in three parts:
    a. Analyse the gene expression data for associations with Aβ status;
    b. Analyse the genotype data for associations with Aβ status;
    c. Find eQTLs using both genotype and gene expression data;
    d. Carry out enriched network decomposition (**END**) analysis using gene network information and the gene expression data.
4. Note: references below to "sections" are references to the section numbers in the 'summary_to_date' html file I created with R Markdown, so that you can see the integrated code and outputs.

Analyse gene expression data – section 1

5. First we created PCA plots to visualize the expression data, and filtered the sample by removing three outliers (section 1.2).
6. We then annotated the gene data with their gene symbols, entrez IDs, and ensemble IDs (section 1.4).

7. Finally, we ran limma over the expression data to determine which genes were differentially expressed (**DE**) between the Aβ-pos and –neg groups (section 1.6). That revealed **865 genes** (probes) that were DE between the two groups at the nominal p-value (but not at the FDR-adjusted value), **530** of which we could annotate with gene names.

Analyse genotype data – section 2.1 – 2.8

8. First we annotated our ~2,000 SNPs and their ~18,000 "proxy" SNPs (SNPs they are in LD with), with their locations and the names of the genes they reside in (section 2.3 – 2.5).
9. We then used PLINK to test whether there were associations between genotype and Aβ status, using chi square (section 2.7) and logistic regression including covariates (section 2.8). That analysis revealed a number of SNPs that were significant at the nominal p-value, but not at the FDR-adjusted value: see the summary tables at section 2.11.

Find eQTLs – section 2.9 – 2.10

10. Using the MatrixEQTL package in R, we analysed eQTLs in:
    a. The whole sample (section 2.9);
    b. The Aβ-pos group alone (section 2.10); and
    c. The Aβ-neg group alone (section 2.10).
11. That analysis revealed a number of SNPs that affected the differential expression of a number of genes: see the summary tables at section 2.11.

END analysis – section 3

12. We took the 530 annotated genes that were found to be DE in the original limma, and queried the KEGG database to find the networks they sat within.
13. Of those 530 genes, 168 of them were found to sit within 199 different networks. We then assembled a list of those original DE genes that sat within each of those networks.
14. Next, we applied James Doecke's END method to each network as follows:
    a. import the expression data for each gene in the network;
    b. run an SVD over that data;
    c. pull out the first eigenvalue for each individual; and

d. run limma over those eigenvalues, comparing the Aβ-pos and –neg groups.

15. When run over the entire list of networks, the END produces a table of p-values for the networks for their differential expression between those groups.
16. We found that by using the END method on the network, we improved our ability to distinguish between Aβ-pos and –neg groups (see the final plot at the end of section 4.3.4, which is annotated with the network names).

Intersection between genes found through the above analyses

17. No DE genes from the original limma intersected with any genes found in the SNP analyses, but six of those genes were also found to be differentially expressed in the eQTL analysis: *ATR*, *FSCN1, PPP1R12A, CHUK, MAPK9,* and *CRLS1*. They were all associated with a single SNP: rs3752472. We will investigate this further.

Validation using ROSMAP data

18. We are in the process of obtaining gene expression data from brain tissue of a separate cohort of around 400 individuals from the US. We seek to validate our findings using that data. Any genes that are found to be significant in both data sets may have a central role in the disease's pathology, and its differential expression patterns in the blood could be used as a biomarker.