TFG\_Carmen Martinez

This repository contains all the scripts and data that were used during the development of a Weighted Gene Co-expression Network Analysis on brain-expressed genes. This study was the main topic of a Bachelor’s thesis (TFG) in Biotechnology by the student Carmen Martínez.

The analysis was supported by reference materials on WGCNA created by Zhang, Langfelder and Horvath.

DATA Clean data (allData and datExpr).RData: data file containing gene expression data (datExpr), already pre-processed, and extra gene data (gene symbol, ILMN probe ID, GeneBank Accession Number, GO terms, etc) phenoData\_binarized.RData: contains data frames of binarized traits for posterior analysis, both annotated with text labels (binarized\_phenoData1) and without text (binarized\_phenoData)

SCRIPTS

1. Data pre-processing and binarization of trait data frame.Rmd: pre-processing of data set. This included removal of non-identified genes (no Gene ID assigned to the ILM probe); removal of duplicated genes or samples (no duplications found), and of genes whose expression values showed no variance across samples. Binarization of phenotypical traits (autistic/control and brain region) allowed for posterior correlation analysis. Condition status was assigned values 1 for autism and 0 for controls. In the case of brain regions, comparisons in the form of XvsAll were performed, where X (region of interest in that comparison) was assigned 1, while the remaining regions were assigned 0. 1b. phenoData\_binariez contains the matrices derived from this piece of code (one of the data frames includes text labels for each sample; the other doesn’t, and is therefore used for thecorrelation analysis that will follow)
2. DEG Screening.Rmd: Differentially Expressed Genes screening using limma package. DE criteria were set at |Fold Chage|>1.3 and FDR <0.05. IDs of DEGs were saved for later analysis, together with their FC and FDR. 2A. DEG analysis in Cortex and Cerebellum: We performed a first analysis grouping together all cortex samples, so the DEG analysis was Autistic Cortex vs Control Cortex + Autistic Cerebellum vs Control Cerebellum (no distintction betwen Frontal and Temporal cortex). The Cortex DEG screening analysis was not included in the final report since it was not considered informative enough for the aim of the study, as we aimed to differentiate among Temporal and Frontal cortex expression patterns. 2B. DEG analysis in Frontal and Temporal Cortex: Differentially Expressed Genes screening in Temporal cortex, Frontal cortex and Cerebelum (in all cases, the comparison was Autistic vs. Control). DEGs were plotted in a volcano plot using the ggplot2 package. 2C. Graphic representation of DEGs in volcano plots.
3. WGCNA.Rmd: step-by-step module indentification, characterization and visulization

* 3.1. Sample clustering 3.2. Load external data (condition status and brain region) into sample dendrogram 3.3. Soft thresholding parameter selection accordin to connectivity and R2 scale-free topology fit. 3.4. Module construction 3.4.1. Adjacency, TOM and dissTOM calculation from datExpr. 3.4.2. Module definition by gene hierarchical clustering + Dynamic tree cut and close module merging according to Module Eigengene dissimilarity 3.4.3. Plot dissTOM (computationally heavy) 3.5. Module representation in Multi-Didmensional Scaling plot (MDS) 3.6. Module-trait relationship and plotting in labelled heatmap 3.7. Module significance to autism and plotting in barplot. Module significance is calculated as average gene significance of all genes in the module. 3.8. Selection of data for Functional enrichment analysis: relevant modules and DEGs 3.8.a. Selection of gene IDs of relevant modules to be uploaded to DAVID database 3.8.b. Plot GO enrichment for relevant modules - Orange - Lightcyan - Darkturquoise - Darkgrey - Darkgreen - Black 3.8.c. Plot GO enrichment for DEGs. 3.9. Network visulization in VisANT for Orange and Lightcyan. Top 20 connected nodes were selected, with a Topological Overlap Threshold of 0.1