2020-12-23 09:56:06

PM_enrichment

This is our main result, running the model `geneExpression ~ DX + nuiscancePCs` in the post mortem (PM) data. It gives a list of genes with associated p-value and statistics, and I ran that for ACC and Caudate separately. In most papers they do a FDR cut-off in that list and run with the selected genes. Nothing survives if we do that, so we explore these results in different ways.

- ORA (over-representation analysis): given two lists of genes A and B, selected from a universe U of genes, calculate whether the overlap between A and B is significant.
- GSEA (gene set enrichment analysis): given a **ranked** list of genes A and a gene set B, gives the probability that the genes in B are ranked higher than genes not in B. That's a simplistic explanation, and the different algorithms (WebGestalt (WG), camera, GAGE, fgsea) perform different math to calculate that probability. Another variable here is how to calculate the rank for each gene in A. For these results, I used `-log(P)*sign(t)` as the rank, and WG to calculate GSEA.

So, in this folder you'll see GSEA results for pmACC and pmCaudate, using GeneOntology sets and also gene sets we devised ourselves.

go_pics

I then selected 2 gene sets from Gene Ontology and made boxplots of the leading genes in those sets. I made them for the raw `geneExpression ~ DX` but also for `resid(geneExpression ~ nuiscancePCs) ~ DX`, but the results are similar. I think it's easier to visualize them in the raw though. As usual, in these boxplots the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The whiskers extend from the hinge to the largest/smallest value no further than 1.5*IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). Data beyond the end of the whiskers are called "outlying" points and are plotted individually. The notches extend `1.58*IQR / sqrt(n)`. This gives a roughly 95% confidence interval for comparing medians. If notches don't overlap, it suggests that the medians are significantly different.

PM_overlap

I used different nominal p-value thresholds to select individual genes from pmACC and pmCaudate results, and evaluated whether the ACC and Caudate overlap was significant using ORA. I asked Gauri to check the 10 genes in the top result (.png), and she compiled the list in the .xlsx.

PRS

I used the genotypes in the PM dataset and the ADHD GWAS to calculate PRS. The first result is PRS_models.txt, which only looks at `DX ~ PRS` to establish whether PRS is related to ADHD at all in this cohort, regardless of gene expression. We normally use that to select a PRS threshold based on R^2. In this case, we'd select PRS0.5.

We can then replace DX by PRS to run `geneExpression ~ PRS + nuiscancePCs` in the PM cohort, again split between ACC and caudate. That also results in a list of genes with p-values and statistics. I thresholded the lists using different nominal p-values, and did an ORA analysis against the genes in the `geneExpression ~ DX + nuiscancePCs` result, and that's listed in the two .csv files.

However, that result doesn't take into consideration whether genes are over or under expressed. I then created *UpDown*.csv results, which calculates the overlap only within the list of genes up or down regulated, based on their tstats. The issue then is how to calculate the gene universe for the hypergeometric test. We select a certain number of PRsgenes and PMgenes, and check if their overlap is significant. The bigger the universe of possible selections, the more significant the overlap. When not splitting between up/down, the universe is straight-forward: it's simply all genes we're working with. But when we split it, it could be argued that the universe is only the genes up (or down), or it's still the entire gene universe (as being up or down could be seen as another form of selection). I provided p-values for both (pvalWhole is the latter scenario).

Even though the up/down splits weren't veyr significant, we can still visualize them by splitting the more significant results that used just the pvalues into up and down-regulated, and displaing them in Venn diagrams (venn_pics.zip).

ABCD

Here we use the ABCD genotypes to impute the gene expression in the brain. That gives two big tables (ACC and Caudate), listing the imputed gene expression for each subject. We then use the brain scans of each subject in the model `impGeneExpression ~ brain + age + sex + fsqc_qu_motion + fsqc_qu_pialover + fsqc_qu_wmunder + fsqc_qu_inhomogeneity + fsqc_qu_artifact`. As usual, that generates a list of genes with associated stats, which we can compare to the PM gene lists.

Similarly to what I did for the PRS analysis, I ran ORA for different nominal p-value cut-offs (imp*.pnc). I also listed the 27 genes at the most significant cut-off for ACC (acc_top_genes.png). And I also ran the ORA splitting between up and down-regulated genes (all*upDown*csv).

I also ran GSEA for the ABCd results (enrichment*txt) using the GO sets and our own sets, but there was nothing significant.

SPredXcan

This analysis does not use any genotypes or brain data (either ours or ABCD). Instead, it goes directly from the ADHD GWAS results, using the MetaXcan databases, to a list of expressed genes in either ACC or Caudate, with stats on how each imputed gene differs by the two groups in the GWAS.

There was no significant overlap between imputed trnscripts and the RNA pmACC results (acc_overlaps.txt). That result is formatted a bit differently (older scripts), and you'll see a matrix for the intersection numbers and one for pvalues, where rows and columns correspond to imputation and pmACC nominal thresholds. Note that here the same question about the gene count for the universe applies. The pvalues I used in the .txt files is for all genes, which would give the best possible (lowest) p-value. Still, not much there, so I didn't compute the per-direction p-value.

Another way to check the overlap between imputed and PM result is to calculate Spearman correlation between the two ranked list of genes. Here, we can rank the lists based on p-value, or the same rank used for GSEA (`-log(P) * sign(t)`). Those results are in correlations.txt. It worked for the absolute values of ACC, but that was it.

imputation_enrichment

These results are somewhat similar (analytically) to the ABCd results. The main difference is that here we use our own data, and I use the MetaXcan tools not only to run the imputation (from our genotypes to imputed genes in the two brain regions), but I also use their tool for association between brain and imputed genes. Specifically, I first calculated `resid(brain ~ sex + age + 3QCmetrics)` running an LME with all our WNH subjects, and then used that adjusted phenotype in the PrediXcanAssociation tool to calculate `impGenes ~ adjBrainPhenotype`. But looking at their code, all they do is a OLS in python. Their tool is extremely finicky too, so that's why when I ran the ABCD dataset I decided to do it myself in R. I also prefer the model that takes into account the covariates in the same regression.

There was no significant overlap between these results and the PM brain. The enrichment results here are the ones we've been looking at for a while. The enrichment to Gene Ontology had nothing earth-shattering either.

WGCNA_WG_pmACC

This was my first stab at network analysis with the PM data. I only ran it for the ACC, and it gave 3 clusters nominally associated with Diagnosis (p < .05), here referenced as black, yellow, and royalblue. The enrichment results for those clusters somewhat mirror what we saw when enriching the actual PM results. WCGNA might be more powerful if they survive some FDR threshold after removing non-robust clusters. Not sure if they will (in fact, I can't even tell whether those 3 clusters are robust). That's the next step.

TODO

- run DX*brain for PM brain
- beef up WGCNA results
  - check ACC clusters robustness
  - make pictures of the genes there
  - repeat everything for Caudate
- check PRS to ABCD imputation overlap (regardless of circularity)
- check how diagnosis for each brain bank was conducted
- run more traditional DTE and DTU pipelines while paper is being written
- run FUSION