Vignette 1 - Understanding the Results of MCIA

Max Mattessich Joaquin Reyna Anna Konstorum

9/1/2022

Part 2: Interpreting Global Loadings

Pathway analysis for the top factors using data from gene-centric omics blocks

First, we compute the first 10 global factors for the dataset:

```
data(NCI60) # this creates the dataset as `data_blocks`
mcia_results <- nipals_multiblock(data_blocks,
                                  preprocMethod='colprofile',
                                  plots = 'none',
                                  num_PCs = 10,
                                  tol=1e-12)
```

The NCI60 data set includes gene expression data which can be used with a gene set enrichment analysis (GSEA) but more interestingly, the global loading matrix is a gene by factor matrix and we can learn more about what pathway a given factor may be capturing by running each factor vector through GSEA. In the previous sections we saw how much each mRNA factor is contributing to the MCIA decomposition and so we will focus on factors X, Y, Z. We will start by running gsea_report() which reports on the most significant pathway for the given factor as well as the total number and finally a summary of all factors using the selectivity score as described by Cantini et al., 2021.

```
# extract mRNA global loadings
mrna_gfscores <- mcia_results$global_loadings
mrna_rows = str_detect(row.names(mrna_gfscores), '_mrna')
mrna_gfscores <- mrna_gfscores[mrna_rows,]
row.names(mrna_gfscores) <- str_remove(rownames(mrna_gfscores), "_[0-9]*_.*")

# load pathway data
path.database = '../data/c2.cp.reactome.v6.2.symbols.gmt'
pathways <- fgsea::gmtPathways(path.database)

# generate the GSEA report
geneset_report = gsea_report(mrna_gfscores, path.database,
                      factors = seq(1,10), pval.thr = 0.05)
```

```
## [1] "Running GSEA for Factor1"
## [1] "Running GSEA for Factor2"
## [1] "Running GSEA for Factor3"
## [1] "Running GSEA for Factor4"
## [1] "Running GSEA for Factor5"
## [1] "Running GSEA for Factor6"
## [1] "Running GSEA for Factor7"
## [1] "Running GSEA for Factor8"
## [1] "Running GSEA for Factor9"
## [1] "Running GSEA for Factor10"
```

GSEA Summary Table The report comes in the form of a list where the first element is a data frame with summary level of the GSEA analysis per factor. Ideally, each factor is capturing a very select number of pathways with a high significance. From this report (below) we can see that the most significant pathway is associated with Factor3 and see can also see that there is a large variation in the number of total (significant) pathways ranging from 7 (Factor 8) to 143 (Factor 4).

```
geneset_report[[1]]
```

```
##              min_pval total_pathways
## Factor1  2.461419e-09            119
## Factor2  2.724331e-08             95
## Factor3  1.716512e-20             88
## Factor4  5.158792e-14            138
## Factor5  6.303990e-06             26
## Factor6  1.143666e-06             72
## Factor7  9.685555e-07             13
## Factor8  3.360834e-09              8
## Factor9  8.556572e-08             18
## Factor10 8.727664e-05             20
```

Using the previous table we can see that something interesting is going on with Factor3 so we can re-run GSEA one more time for this factor in order to get a full list of top pathways:

```
factor3_paths = fgseaMultilevel(pathways, mrna_gfscores[,3], nPermSimple = 10000,minSize=15, maxSize=500

sig_path3 = factor3_paths[min(factor3_paths$pval) == factor3_paths$pval,]
```

The top scoring pathway is REACTOME_CELL_CYCLE which makes sense give this is cancer data set. This analysis can be repeated as necessary to make sense of other gene based factor loadings.

Selectivity Score The second element in the report is the selectivity score which is calculated as follows:

$$S = Selectivity\,Score = (N_c + N_f)/2L$$

where N_c is the total number of clinical annotations associated with at least a factor, N_f the total number of factors associated with at least a clinical annotation, and L the total number of associations between clinical annotations and factors. S has a maximum value of 1 when each factor is associated with one and only one clinical/biological annotation, and a minimum of 0 in the opposite case. An optimal method should thus maximize its number of factors associated with clinical/biological annotations without having a too low selectivity.

```
geneset_report[[2]]
```

```
## [1] 0.2160804
```

For the mRNA global loadings we can see that there is low selectivity for enriched pathways which suggests that there is some overlap between the signals capture by each factor.