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 and its corresponding global loading matrix is a gene by factor matrix. We can learn more about the pathways a given factor may be capturing by running each factor vector (column of the global loadings matrix) through a gene set enrichment analysis (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 run gsea_report() which reports on the p-value of the most significant pathway as well as the total number of significant pathways for each factor. Finally the report will summarize all factors using the selectivity score as described by Cantini et al., 2021 (more details below).

```
# 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,]

# rename rows to contain HUGO based gene symbols
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,                                nproc=8)
```

Gather data and generate the report

```
## [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"
```

Investigating the 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 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
```

As just mentioned, Factor3 contains the most enrichment gene set so we can re-run GSEA for this factor in order to get a full list of enrichment scores across all gene sets:

```
# re-running GSEA
factor3_paths = fgseaMultilevel(pathways, mrna_gfscores[,3],
                                nPermSimple = 10000, minSize=15,                                nproc =
factor3_paths
```

```
##                                                                pathway
##   1:                       REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION
##   2: REACTOME_ABORTIVE_ELONGATION_OF_HIV1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT
##   3:                                    REACTOME_ACTIVATED_TLR4_SIGNALLING
##   4:          REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS
##   5:                        REACTOME_ACTIVATION_OF_CHAPERONE_GENES_BY_XBP1S
##  ---
## 336:                                  REACTOME_TRIF_MEDIATED_TLR3_SIGNALING
## 337:                                     REACTOME_TRIGLYCERIDE_BIOSYNTHESIS
## 338:                                          REACTOME_TRNA_AMINOACYLATION
## 339:                                     REACTOME_UNFOLDED_PROTEIN_RESPONSE
## 340:                          REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G
##              pval        padj    log2err         ES        NES size
##   1: 0.0577183480 0.156993907 0.07684109  0.2946342  1.3254666   98
##   2: 0.5741140882 0.684908035 0.02398695 -0.2871928 -0.9159034   18
##   3: 0.0505440505 0.145635400 0.08385351  0.3425159  1.3932001   58
##   4: 0.0006745322 0.004985673 0.47727082  0.5680802  1.9803912   29
##   5: 0.1203347659 0.262268080 0.06089253 -0.3440170 -1.2926426   34
##  ---
## 336: 0.1459671663 0.306721804 0.04835174  0.3188906  1.2452116   48
## 337: 0.0003457019 0.002938466 0.49849311 -0.6045312 -2.1234233   26
```

```
## 338: 0.0311582750 0.104889243 0.12218443 -0.3849306 -1.4784920   37
## 339: 0.1347765363 0.290025458 0.05822151 -0.2836415 -1.2228956   61
## 340: 0.7094339623 0.793445879 0.01786438  0.2318845  0.8441878   35
##                                          leadingEdge
##   1:        EIF3F,EIF2S2,RPS20,EIF4E,RPS25,RPS3A,...
##   2:              GTF2F1,POLR2H,COBRA1,POLR2C,POLR2E
##   3:     CREB1,HMGB1,TRAF3,MAPK14,MAP2K6,PPP2R5D,...
##   4:              MCM8,RPA1,MCM10,MCM3,MCM2,MCM6,...
##   5: C19orf10,ATP6V0D1,KDELR3,YIF1A,FKBP14,SYVN1,...
##  ---
## 336:      CREB1,HMGB1,MAPK14,MAP2K6,PPP2R5D,AGER,...
## 337:     AGPAT2,AGPAT3,FASN,SLC25A1,ELOVL1,GPD1L,...
## 338:            EPRS,AIMP2,VARS,AARS,DARS2,EARS2,...
## 339:  C19orf10,ATP6V0D1,HSP90B1,KDELR3,IL8,YIF1A,...
## 340:         PSME1,PSMA3,PSMF1,PSMB8,PSMB9,PSMB2,...
```

If we extract the most significant gene set (below) we can see that the REACTOME_CELL_CYCLE gene set comes up which makes sense given that the NCI60 data set is studying cancer. This analysis can be repeated as necessary to make sense of other gene based factor loadings.

```
# extracting to most significant gene set
sig_path3 = factor3_paths[min(factor3_paths$pval) == factor3_paths$pval,][1,]

as.list(sig_path3)
```

```
## $pathway
## [1] "REACTOME_CELL_CYCLE"
##
## $pval
## [1] 3.830388e-23
##
## $padj
## [1] 1.302332e-20
##
## $log2err
## [1] 1.246233
##
## $ES
## [1] 0.4993744
##
## $NES
## [1] 2.62987
##
## $size
## [1] 301
##
## $leadingEdge
## $leadingEdge[[1]]
##   [1] "E2F1"      "HIST1H2BM" "PPP2R5C"   "CASC5"     "NINL"      "CCNE2"
##   [7] "POLE"      "RAD21"     "HIST1H4L"  "SYNE2"     "HIST1H4A"  "NEDD1"
##  [13] "RBL1"      "TP53"      "SKP2"      "PCNA"      "TINF2"     "STAG1"
##  [19] "MLF1IP"    "PCM1"      "PLK4"      "CDKN2C"    "AURKB"     "PSME1"
##  [25] "ACTR1A"    "DNA2"      "CENPK"     "MCM8"      "TFDP1"     "H2AFZ"
##  [31] "ALMS1"     "SMC3"      "HSP90AA1"  "H2AFX"     "CENPA"     "ANAPC10"
##  [37] "CEP135"    "PSMA3"     "HIST1H4K"  "TUBGCP3"   "MIS12"     "CCND3"
```

```
##  [43] "LMNB1"     "PPP2R5D"   "CEP63"     "CDC25B"    "SGOL1"     "CENPO"
##  [49] "PCNT"      "SPC25"     "ZWINT"     "TERF1"     "MNAT1"     "POLD3"
##  [55] "CDK5RAP2"  "CEP70"     "HIST1H2BH" "CEP57"     "RPA1"      "MAD2L1"
##  [61] "PSMF1"     "HIST1H2BI" "DCTN1"     "CEP250"    "DYNC1I2"   "TUBGCP5"
##  [67] "DIDO1"     "STAG2"     "CENPI"     "RRM2"      "LIG1"      "KIF18A"
##  [73] "PPP2R5E"   "SMARCA5"   "PSMB8"     "MAPRE1"    "DYRK1A"    "HIST1H4C"
##  [79] "ATM"       "PAFAH1B1"  "FGFR1OP"   "SMC1A"     "HIST1H2AB" "MCM10"
##  [85] "SGOL2"     "MCM3"      "MCM2"      "PSMB9"     "ANAPC4"    "CDK1"
##  [91] "PMF1"      "POLA2"     "CENPN"     "MCM6"      "POLE2"     "MAX"
##  [97] "NUMA1"     "MCM7"      "PSMB2"     "CDC7"      "NDEL1"     "CLASP1"
## [103] "MCM5"      "PSME2"     "HJURP"     "E2F2"      "KIF20A"    "LIN52"
## [109] "RPS27"     "BTRC"      "RAD9A"     "HIST1H2AJ" "DCTN3"     "POLA1"
## [115] "KNTC1"     "HIST1H4B"  "DYNLL1"    "PSMB10"    "RAD1"      "PSMB7"
## [121] "WEE1"      "CEP192"    "BUB1"      "KIF2C"     "RFC3"      "TYMS"
## [127] "RBL2"      "CHEK1"     "CEP76"     "KIF23"     "CCDC99"    "FBXO5"
## [133] "AKAP9"     "ORC2"      "CENPJ"     "TUBGCP6"   "CENPQ"     "TAOK1"
## [139] "E2F3"      "CEP290"    "RFWD2"     "RAD17"     "CCNB2"     "AHCTF1"
## [145] "GINS1"     "HDAC1"     "CCNA2"     "WRAP53"    "GINS4"     "GINS2"
## [151] "ORC6"      "AZI1"      "CSNK1D"    "ANAPC1"    "DBF4"      "TERT"
## [157] "CDC25A"    "NUP85"     "HAUS2"     "TERF2IP"   "ITGB3BP"   "CEP72"
```

Investigating the 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 one factor, N_f the total number of factors associated with at least one 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 which suggests that there is some overlap between the signals capture by each factor.