Application of TPAC method to TCGA liver cancer RNA-seq data using MSigDB Hallmark collection

H. Robert Frost

1 Load and process TCGA liver cancer RNA-seq data

The following logic loads FPKM normalized counts for The Cancer Genome Atlas (TCGA) [1] liver cancer (LIHC) cohort. The TPAC function *tpacForCancer()* leverages Human Protein Atlas (HPA) normal tissue gene expression data ("HPA.normal.FPKM.GDCpipeline.csv") that was specially processed by the HPA group as FPKM values using a pipeline similar to that employed by GDC for the TCGA data (this data was generated for the "Human Pathology Atlas" paper [2]). For consistency with this HPA normal tissue data, the TCGA data is retrieved from the HPA provided TCGA gene expression data file rna_cancer_sample.tsv, which contains FPKM normalized counts and can be downloaded from https://www.proteinatlas.org/download/rna_cancer_sample.tsv.zip.
Generation of the LIHC-specific matrix from the rna_cancer_sample.tsv data was performed using the following R code (which is not executed here given the size of the data and processing time):

```
library(data.table)
library(reshape2)
tcga.file = "rna_cancer_sample.tsv"
data = read.table(file=tcga.file, header=T, sep="\t", stringsAsFactors=F,
                  # Columns: "Gene", "Sample", "Cancer", "FPKM"
                  colClasses=c("factor", "factor", "factor", "numeric"))
data = as.data.table(data)
setkey(data, Cancer, Gene)
lihc.data = data[.("LIHC")]
lihc.matrix = acast(lihc.data, Sample~Gene, value.var="FPKM")
saveRDS(lihc.matrix, file="lihc.rna.matrix.rds")
```

```
> # read in saved matrix
> liver.counts.fpkm = readRDS(file="lihc.rna.matrix.rds")
> # change to FPKM + 1
> liver.counts.fpkm = liver.counts.fpkm + 1
```

2 Load the MSigDB Hallmark collection

The following logic loads the MSigDB Hallmark collection using the msigdbr R package. The data frame returned by msigdbr is then converted into a list of gene ID vectors (each list element corresponds to a gene set and is a vector of Ensembl IDs). The *tpacForCancer()* function automatically transforms this into a list of vectors of gene indices using the *createGeneSetCollection()* helper function.

```
> # Load the MSigDB Hallmark collection using the msigdbr package
> hallmark.collection = msigdbr::msigdbr(category="H")
> # Create a gene.set.collection list of Ensembl IDs
```

```
> gene.set.names = unique(hallmark.collection$gs_name)
> num.sets = length(gene.set.names)
> message("Number of sets in MSigDB Hallmark collection: ", num.sets)
> gene.set.names[1:5]

[1] "HALLMARK_ADIPOGENESIS"        "HALLMARK_ALLOGRAFT_REJECTION"
[3] "HALLMARK_ANDROGEN_RESPONSE"   "HALLMARK_ANGIOGENESIS"
[5] "HALLMARK_APICAL_JUNCTION"

> gene.set.collection = list()
> for (i in 1:num.sets) {
+         gene.set.name = gene.set.names[i]
+         gene.set.rows = which(hallmark.collection$gs_name == gene.set.name)
+         gene.set.ensembl.ids = hallmark.collection$human_ensembl_gene[gene.set.rows]
+         gene.set.collection[[i]] = unique(gene.set.ensembl.ids)
+ }
> names(gene.set.collection) = gene.set.names
```

3 Execute TPAC method

Since we are processing TCGA RNA-seq liver cancer data, we can execute TPAC using the *tpacForCancer()* wrapper function. Note that the cancer types supported by *tpacForCancer()* can be accessed via the *getSupportedCancerTypes()* function.

```
> # Display the full list of cancer types supported by tpacForCancer()
> TPAC::getSupportedCancerTypes()

 [1] "urothelial cancer"    "breast cancer"        "cervical cancer"
 [4] "colorectal cancer"    "glioma"               "head and neck cancer"
 [7] "renal cancer"         "liver cancer"         "lung cancer"
[10] "ovarian cancer"       "pancreatic cancer"    "prostate cancer"
[13] "colorectal cancer"    "melanoma"             "stomach cancer"
[16] "testis cancer"        "thyroid cancer"       "endometrial cancer"

> # Get the normal tissue corresponding to liver cancer
> cancer.type = "liver cancer"
> # Execute TPAC
> tpac.out = TPAC::tpacForCancer(cancer.gene.expr=liver.counts.fpkm,
+                           cancer.type=cancer.type,
+                           gene.set.collection=gene.set.collection)
```

Look at a subset of the TPAC scores in the generated S, S- and S+ matrices:

```
> tpac.out$S[1:5,1:5]

                 HALLMARK_ADIPOGENESIS HALLMARK_ALLOGRAFT_REJECTION
TCGA-2Y-A9GS-01A          3.064720e-03                 3.018793e-01
TCGA-2Y-A9GT-01A          7.926992e-14                 7.994741e-05
TCGA-2Y-A9GU-01A          9.918417e-01                 8.723984e-01
TCGA-2Y-A9GV-01A          0.000000e+00                 1.707913e-04
TCGA-2Y-A9GW-01A          5.827508e-01                 3.248046e-03
                 HALLMARK_ANDROGEN_RESPONSE HALLMARK_ANGIOGENESIS
```

```
TCGA-2Y-A9GS-01A               2.386972e-01          0.7538041194
TCGA-2Y-A9GT-01A               2.586985e-06          0.9999999828
TCGA-2Y-A9GU-01A               4.637972e-01          0.0003027669
TCGA-2Y-A9GV-01A               5.906668e-04          0.9997277821
TCGA-2Y-A9GW-01A               9.060539e-01          0.9999999999
                 HALLMARK_APICAL_JUNCTION
TCGA-2Y-A9GS-01A               0.24475713
TCGA-2Y-A9GT-01A               0.03069919
TCGA-2Y-A9GU-01A               0.14960836
TCGA-2Y-A9GV-01A               0.05386065
TCGA-2Y-A9GW-01A               0.41299687
```

> tpac.out$S.neg[1:5,1:5]

```
                 HALLMARK_ADIPOGENESIS HALLMARK_ALLOGRAFT_REJECTION
TCGA-2Y-A9GS-01A          1.642178e-03                 0.4488871715
TCGA-2Y-A9GT-01A          2.716716e-13                 0.0002430214
TCGA-2Y-A9GU-01A          9.848749e-01                 0.9616570059
TCGA-2Y-A9GV-01A          0.000000e+00                 0.0010408253
TCGA-2Y-A9GW-01A          6.410074e-01                 0.0089029673
                 HALLMARK_ANDROGEN_RESPONSE HALLMARK_ANGIOGENESIS
TCGA-2Y-A9GS-01A               2.649604e-01           0.852475176
TCGA-2Y-A9GT-01A               6.135243e-06           0.999999997
TCGA-2Y-A9GU-01A               5.329255e-01           0.001090343
TCGA-2Y-A9GV-01A               9.895708e-04           0.999862611
TCGA-2Y-A9GW-01A               9.256814e-01           1.000000000
                 HALLMARK_APICAL_JUNCTION
TCGA-2Y-A9GS-01A               0.05700461
TCGA-2Y-A9GT-01A               0.01287834
TCGA-2Y-A9GU-01A               0.56693484
TCGA-2Y-A9GV-01A               0.01254875
TCGA-2Y-A9GW-01A               0.16432979
```

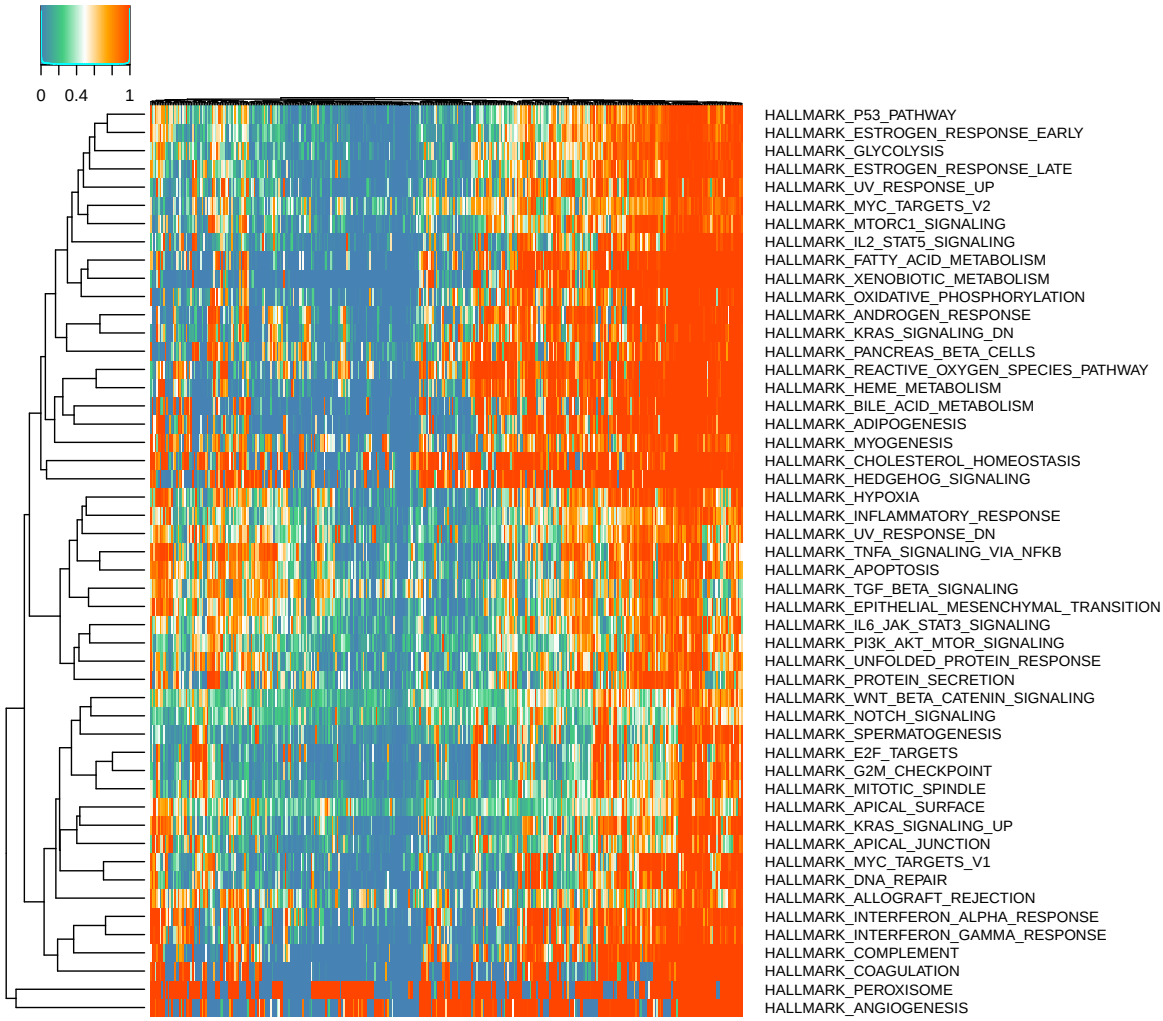
> tpac.out$S.pos[1:5,1:5]

```
                 HALLMARK_ADIPOGENESIS HALLMARK_ALLOGRAFT_REJECTION
TCGA-2Y-A9GS-01A            0.85091309                  0.232426214
TCGA-2Y-A9GT-01A            0.08466006                  0.190018804
TCGA-2Y-A9GU-01A            0.96036440                  0.001615544
TCGA-2Y-A9GV-01A            0.36058713                  0.072145624
TCGA-2Y-A9GW-01A            0.16157345                  0.193789454
                 HALLMARK_ANDROGEN_RESPONSE HALLMARK_ANGIOGENESIS
TCGA-2Y-A9GS-01A                 0.37298345             0.1748558
TCGA-2Y-A9GT-01A                 0.03279483             0.6134479
TCGA-2Y-A9GU-01A                 0.02420909             0.3539980
TCGA-2Y-A9GV-01A                 0.11124720             0.8844703
TCGA-2Y-A9GW-01A                 0.20945836             0.3179549
                 HALLMARK_APICAL_JUNCTION
TCGA-2Y-A9GS-01A               0.47087802
TCGA-2Y-A9GT-01A               0.10049191
TCGA-2Y-A9GU-01A               0.09223503
```

```
TCGA-2Y-A9GV-01A               0.17151841
TCGA-2Y-A9GW-01A               0.61519815
```

Visualize the TPAC scores in the **S** matrix as a heatmap (this is generated using similar logic as the heatmaps included in the TPAC paper [3]).

```
> library(gplots)
> my_palette = colorRampPalette(c("steelblue", "seagreen3",
+                                 "white", "orange", "orangered"))(n = 299)
> breaks = 300
> heatmap.2(t(tpac.out$S),
+           col = my_palette, dendrogram="both", na.rm=T,
+           symm=F, scale = "none", trace = "none",
+           xlab=NA, ylab=NA, labCol=NA, sepcolor="white",
+           sepwidth=c(0, .2), symkey=F,
+           Rowv=T, Colv=T,
+           breaks=breaks, margins=c(2,27),
+           key.title=NA, key.ylab=NA, key.xlab=NA,
+           key.ytickfun=function() {
+             return(list(labels=FALSE, tick=FALSE))
+           },
+           lwid=c(.5,4), lhei=c(.5,4), main = NA)
```



References

[1] Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nat Genet **45**(10), 1113–20 (2013). doi:10.1038/ng.2764

[2] Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J.M., Brunnström, H., Glimelius, B., Sjöblom, T., Edqvist, P.-H., Djureinovic, D., Micke, P., Lindskog, C., Mardinoglu, A., Ponten, F.: A pathology atlas of the human cancer transcriptome. Science **357**(6352) (2017). doi:10.1126/science.aan2507

[3] Frost, H.R.: Tissue-adjusted pathway analysis of cancer (tpac). bioRxiv (2022). doi:10.1101/2022.03.17.484779. https://www.biorxiv.org/content/early/2022/03/19/2022.03.17.484779.full.pdf