

STAT646 - Homework 1

Daniel Osorio - dcosoriorh@tamu.edu
Department of Veterinary Integrative Biosciences
Texas A&M University

1. Consider the human gene with HGNC symbol SPRR4.

(a) Use the Biomart Ensembl database to obtain cDNA and peptide sequences for SPRR4. What are the two sequences?

```
> library(biomaRt)
> mart <- useMart(biomart = "ensembl",
+               dataset = "hsapiens_gene_ensembl")
> peptideS <- getSequence(id="SPRR4",
+               type="hgnc_symbol",
+               seqType="peptide",
+               mart = mart)[,1]
> peptideS
[1] "MSSQQQQRRQQQQCPPQRAQQQVKQPCQPPVVKCQETCAPKTKDPCAPQVKKQCPPKGTIIPAQQKCPSAQQAS"
> cdnaS <- getSequence(id="SPRR4",
+               type="hgnc_symbol",
+               seqType="cdna",
+               mart = mart)[,1]
> cdnaS
[1] "CTCTCCTGGGGTCCAGCTTGTCGCCTCTGGCTCACCTGTTCTAGAGCAATGTCTTCCCAGCAGCAGCAGCGGCGG"
```

(b) What is the Entrez ID for SPRR4?

```
> entrez <- getBM(attributes = c("hgnc_symbol", "entrezgene"),
+               mart = mart)
> entrezSPRR4 <- entrez[entrez[,1] %in% "SPRR4",2]
> entrezSPRR4
[1] 163778
```

(c) Retrieve GO information for SPRR4. What biological processes is the gene involved in? Where in the cell is the SPRR4 protein located?

```
> library(org.Hs.eg.db)
> library(GO.db)
> GO <- mget(x = as.character(entrezSPRR4),
+               envir = org.Hs.egGO)[[1]]
> GO <- data.frame(GO=unlist(lapply(GO, function(X){c(X[1])})),
+               ONTOLOGY =unlist(lapply(GO, function(X){c(X[3])})))
> as.vector(Definition(as.vector(GO[GO[,2] == "BP",1])))
[1] "The formation of a covalent cross-link between or within protein chains."
[2] "The process in which a relatively unspecialized cell acquires specialized"
[3] "The process in which the cytoplasm of the outermost cells of the vertebrate"
```

```
> as.vector(Term(as.vector(GO[GO[,2] == "CC",1])))
[1] "cornified envelope" "cytoplasm"
[3] "cell cortex"
```

2. Consider the human gene with HGNC symbol BRCA1.

- (a) Why is the BRCA1 gene relevant to breast cancer? *'BRCA' is an abbreviation for 'BRest CAncer gene.', it is a gene that encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. Mutations in this gene are responsible for approximately 40% of inherited breast cancers and more than 80% of inherited breast and ovarian cancers.*
- (b) Which probeset on the Affymetrix HGU133a Gene Chip microarray corresponds to BRCA1?

```
> library("hgu133a.db")
> affyIds <- select(hgu133a.db, keys=keys(hgu133a.db),
+                   columns = c("SYMBOL", "ENTREZID"))
> affyIds[affyIds[,2] %in% "BRCA1",]
```

| | PROBEID | SYMBOL | ENTREZID |
|-------|-------------|--------|----------|
| 4322 | 204531_s_at | BRCA1 | 672 |
| 12363 | 211851_x_at | BRCA1 | 672 |

- (c) According to the kegg Bioconductor package, what protein pathway is BRCA1 involved in? Note: This is not the only pathway BRCA1 is involved in. The kegg package is not complete here.

```
> library(KEGG.db)
> hsaPath <- unlist(mget(x = as.character(entrez[entrez[,1] %in% "BRCA1",2]),
+                       envir = KEGGEXTID2PATHID))
> KEGGPATHID2NAME[[gsub("hsa","",hsaPath)]]
[1] "Ubiquitin mediated proteolysis"
```

- (d) What other genes are involved in the above protein pathway? Give their HGNC symbols

```
> pathGenes <- entrez[entrez[,2] %in% KEGGPATHID2EXTID[[hsaPath]],1]
> pathGenes
[1] "WWP2"      "PIAS2"      "PML"        "CDC26"      "FBXO4"
[6] "UBE2L3"    "CBLB"       "PPIL2"      "ANAPC13"    "SKP1"
[11] "UBE2Q2"    "SMURF2"     "UBE2W"      "NEDD4"      "SYVN1"
[16] "UBE2R2"    "RNF7"       "TRIM37"     "HERC2"      "HERC4"
[21] "ELOC"      "UBE2G1"     "ANAPC11"    "UBA6"       "CUL4B"
[26] "MGRN1"     "BTRC"       "MID1"       "DET1"       "DDB2"
[31] "RHOBTB2"   "UBE2B"      "WWP1"       "UBE3A"      "UBE2C"
[36] "TRAF6"     "FANCL"      "VHL"        "STUB1"      "FBXW7"
[41] "SIAH1"     "CUL4A"      "PIAS4"      "UBE2L6"     "UBE2H"
[46] "ELOB"      "ANAPC1"     "XIAP"       "FZR1"       "AIRE"
[51] "HUWE1"     "UBE20"      "UBE2U"      "CUL1"       "KLHL9"
[56] "BIRC6"     "UBE2E3"     "ITCH"       "UBA7"       "SMURF1"
[61] "UBE4A"     "UBR5"       "UBE2M"      "SOCS1"      "ERCC8"
[66] "DDB1"      "HERC1"      "CDC23"      "UBE2G2"     "UBE3C"
[71] "UBE2I"     "CUL2"       "NHLRC1"     "UBE2E2"     "CBLC"
```

```

[76] "ANAPC4" "UBE2E1" "UBA3" "CUL5" "PIAS3"
[81] "UBE2D1" "UBE2NL" "UBE2N" "SAE1" "NEDD4L"
[86] "ANAPC10" "CBL" "FBX02" "CDC27" "UBOX5"
[91] "UBE2K" "KLHL13" "CDC20" "FBXW11" "ANAPC2"
[96] "UBA1" "UBA2" "UBE2Z" "UBE2F" "UBE2S"
[101] "CUL3" "UBE2D3" "UBE2D2" "BIRC3" "TRIM32"
[106] "CUL7" "BIRC2" "SOCS3" "MDM2" "ANAPC7"
[111] "KEAP1" "UBE2D4" "RBX1" "RHOBTB1" "CDC34"
[116] "PRPF19" "CDC16" "PRKN" "UBE2J1" "TRIP12"
[121] "UBE2A" "UBE2QL1" "PIAS1" "MAP3K1" "BRCA1"
[126] "RCHY1" "UBE3B" "SKP2" "COP1" "FBXW8"
[131] "UBE2Q1" "UBE2J2" "HERC3" "ANAPC5" "UBE4B"

```

- (e) Use the topGO package to perform a GO enrichment analysis on the genes involved in the above KEGG pathway. In the runTest function, use the “classic” algorithm and the “fisher” test.

```

> library(topGO)
> allGenes <- rep(0,length(affyIds[,1]))
> names(allGenes) <- affyIds[,1]
> allGenes[affyIds[,2] %in% pathGenes] <- 1
> genSel <- function(X){return(X == 1)}
> GO_data <- new(Class = "topGOdata",
+               ontology = "BP",
+               allGenes = allGenes,
+               geneSel = genSel,
+               nodeSize = 10,
+               annot = annFUN.db,
+               affyLib = "hgu133a.db")
> enrichment <- runTest(GO_data, algorithm = "classic", statistic = "fisher")

```

- i. How many GO terms have p-values < 0.001?

```

> sum(score(enrichment) < 0.01)
[1] 944

```

- ii. What GO term has the smallest p-value (and is hence the “most enriched” in the pathway genes)? Does it describe a cellular location, a biological process, or a molecular function? How does the result compare to the KEGG pathway for BRCA1 that we found above?

```

> goTerm <- names(which.min(score(enrichment)))
> Definition(goTerm)

```

```

"A protein modification process in which one or more groups of a small prot
> Term(goTerm)

```

GO:0032446

"protein modification by small protein conjugation"