Analysis of HPA datasets 22 and 23

Janine Yanes, Hiran Shyamsundar, Shreya Ghosh, and Ella Walmsley

Stat 486

5/5/23

Abstract

For our project, we will investigate protein creation (via gene expression) in cancerous and non-cancerous cell lines by combining two datasets (“RNA HPA cell line gene data” and “RNA HPA cell line cancer gene data”). Our project will aim to answer questions such as “Does higher gene expression always correlate with cancer? Are there some cancers where lower expression correlates with cancer? Are there even statistically significant differences? Do different types of cancer have different levels of expression?” and so on.

**Introduction**

Cancer is a disease characterized by uncontrollable cell growth. Normally, cells grow and differentiate at controlled and regulated levels. The causes of cancer are complex and multifaceted; they can be genetic, environmental, and lifestyle related.

We decided to look at genetic factors by utilizing two different datasets (“”) and (“”) from the Human Protein Atlas (HPA). We wanted to see if we could draw any conclusions about type of cancer and gene expression, namely nTPM, and to see if there was anything else that we could deduce. Each datasets had similar variables, as outlined below.

| Term | Definition |
| --- | --- |
| Gene Name | The name of the gene in shorthand, i.e. MATR3 is shorthand for “gene that encodes for Matrin 3” |
| Cancer (Cancer Cell Line Dataset Only) | The type of cancer that the cell line was taken from, |
| TPM | If you were to sequence one million full-length transcripts, a TPM value represents the number of transcripts you would have seen for a given gene; in other words, it represents the number of transcripts for a specific gene, normalized by the total number of transcripts in the sample, and then scaled to a million |
| pTPM | **transcripts per million protein coding genes** |
| nTPM | normalized transcript per million |

For our project, we chose to focus mostly on gene name (so specific types of genes), cancer (specific types of cancer, for example, bile cancer), and nTPM (normalized). By using the normalized nTPM, we sought to ensure that we were not incorrectly attributing higher rates or expression to naturally more abundant RNA .

**Preparing the data**

These data sets were quite large and somewhat unwieldy to work with. In order to get a better understanding of exactly what we wanted to do and how to work with them, we performed some basic visualizations first. In addition, we also broke the data down into smaller portions- groups of 5, 10, 20, or 100 genes, depending on what analysis we were using.

**Materials and Methods**

**Part 1:** Visualizing the Data

We decided to use scatter plots and boxplots to start our visualization. From our boxplots, we found that there were higher numbers of samples taken from cancer cells with upregulation rather than downregulation (see figure1).

*Figure 1: Code for and image of Boxplot of nTPM values for Cancer and Non-Cancer Cells (Outliers Removed)*

| just\_nTPM <- merged\_data |>  select(nTPM\_nocancer, nTPM\_cancer)  head(just\_nTPM)  summary(just\_nTPM$nTPM\_nocancer)  summary(just\_nTPM$nTPM\_cancer)  q1n <- quantile(just\_nTPM$nTPM\_nocancer, 0.25)  q3n <- quantile(just\_nTPM$nTPM\_nocancer, 0.75)  iqrn <- q3n - q1n  uppern <- q3n + 1.5 \* iqrn  lowern <- q1n - 1.5 \* iqrn  q1c <- quantile(just\_nTPM$nTPM\_cancer, 0.25)  q3c <- quantile(just\_nTPM$nTPM\_cancer, 0.75)  iqrc <- q3c - q1c  upperc <- q3c + 1.5 \* iqrc  lowerc <- q1c - 1.5 \* iqrc  box\_nocancer <- just\_nTPM$nTPM\_nocancer[just\_nTPM$nTPM\_nocancer >= lowern & just\_nTPM$nTPM\_nocancer <= uppern]  box\_cancer <- just\_nTPM$nTPM\_cancer[just\_nTPM$nTPM\_cancer >= lowerc & just\_nTPM$nTPM\_cancer <= upperc]  boxplot(box\_nocancer, box\_cancer,  names = c("Non-Cancer", "Cancer"),  xlab = "Cell Type",  ylab = "nTPM",  main = "Boxplot of nTPM values for Cancer and Non-Cancer Cells (Outliers Removed)") |
| --- |
|  |

This is consistent with current research as per the National Institute of Health. Upregulation refers to the gene being overexpressed (translated more than normal), and downregulation means that the gene was under-expressed (translated less than normal).

We also took the squared difference of the means between nTPMs for cancerous and noncancerous cells and found that some genes, like MATR3, had a much larger difference than anything else. By doing this, we focused on which genes showed a larger difference out of the twenty most reported genes. It should be noted that these differences are for all expressions for all cancer types reported.

*Figure 1: Code for and image of 100 most common gene names and differences in mean squared for them*

| **#finding the 100 most common gene names in rnac** #initializing gnvec  gnvec = NULL  #stores gene name column of rnac in rnac.gn  #note: rnac and rnac.cancer have the same gene names, although not the same amount of data  rnac.gn = rnac$Gene.name  #stores rnac's 100 most common gene names  for (i in 1:100) {  #stress the most common value in rnac.gn in gn  gn = names(which.max(table(rnac.gn)))  #adds gn to gnvec  gnvec = c(gnvec, gn)  #removes all data with gene name gn from rnac.gn  rnac.gn = subset(rnac.gn, rnac.gn!=gn)  }  #initializing cancerv and rnacv  cancerv = NULL  rnacv = NULL  #finds the average nTPM for each gene name in gnvec  for (i in 1:100) {  #mean nTPM in rnac.cancer for gene name gnvec[i] is stored in cancerv  cancerv = c(cancerv, mean(subset(rnac.cancer, Gene.name==(gnvec[i]))$nTPM))  #mean nTPM in rnac for gene name gnvec[i] is stored in rnacv  rnacv = c(rnacv, mean(subset(rnac, Gene.name==(gnvec[i]))$nTPM))  }  #initializing Diff and Diff.sq  Diff = c(1:100)  Diff.sq = c(1:100)  #creates a data frame showing the difference in nTPM between normal and cancer cells  comparison = data.frame(  Gene.name = gnvec, #lists gene name  Normal.nTPM = rnacv, #lists mean nTPM in normal cells for each gene name  Cancer.nTPM = cancerv, #lists mean nTPM in cancer cells for each gene name  Diff = (cancerv - rnacv), #lists the difference between cancer and normal nTPM for each gene name  Diff.sq = (cancerv - rnacv)^2 #squares Diff column to make all values positive  )  comparison #prints comparison dataframe |
| --- |
|  |

|  |
| --- |
|  |

From these results, we can see that there are significant differences between the rates of normal nTPM and cancer nTPM. Again, though- this isn’t for each unique cancer. Were there significant differences between those? First, we created a table to determine if there were any significant differences at all and only included those that were significant (as per a t-test). Once we saw that there were, we took counts of each and determined that for some genes, high differences squared were associated with cancer nearly 100% of the time. We then manually checked with the National Institute of Health to confirm this.

*Figure 3: Code for and image of tibble with t test for difference squared of nTPM expression means*

| gene\_diff <- merged\_data |>  group\_by(`Gene name`, Cancer) |>  summarise(  diff = (mean(nTPM\_cancer) - mean(nTPM\_nocancer))^2,  p\_value = t.test(nTPM\_cancer, nTPM\_nocancer)$p.value  ) |>  filter(p\_value < 0.05) |>  arrange(desc(diff))  gene\_tibble <- gene\_diff |>  select(`Gene name`, Cancer\_Type = Cancer, Cancer, squared\_diff = diff, p\_value)  gene\_tibble |
| --- |
|  |

*Figure 1: Code for and image values of genes with highest numbers of difference squared*

| value\_counts <- table(gene\_tibble$`Gene name`)  sorted\_counts <- sort(value\_counts, decreasing = TRUE)  print(sorted\_counts) |
| --- |
|  |
|  |

We also wanted to see if any genes had a statistically significant difference between the cancerous nTPM and non cancerous nTPM, so we created a table (see figure 5)

*Figure 5- Genes That Had Significant Differences Nearly all or All Times They were Reported. Descriptions via National Institute of Health*

| Gene Name | Gene Full Name | Gene Codes for | Cancer Association |
| --- | --- | --- | --- |
| ADGRA2 | Adhesion G protein-coupled receptor A2 | Adhesion G protein-coupled receptor A2 protein | Upregulation of ADGRA2 is associated with breast, prostate, and lung cancer |
| AHRR | Aryl-hydrocarbon receptor repressor | Aryl-hydrocarbon receptor repressor protein | Downregulation of AHRR may be associated with lung cancer |
| ANKRD44 | Ankyrin repeat domain 44 | Ankyrin repeat domain-containing protein 44 | Downregulation of ANKRD44 is associated with colorectal cancer |
| ARHGEF5 | Rho guanine nucleotide exchange factor 5 | Rho guanine nucleotide exchange factor 5 protein | Associated with some cancers |
| BTK | Bruton tyrosine kinase | Bruton tyrosine kinase protein | Upregulation of BTK is associated with leukemia and lymphoma |
| | C8B | | --- | | Complement C8 beta chain | Complement C8 beta chain protein | Associated with some cancers |
| CBLN4 | Cerebellin 4 | Cerebellin-4 protein | Associated with some cancers |
| CD38 | CD38 molecule | ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase | Upregulation of CD38 is associated with multiple myeloma and chronic lymphocytic leukemia |
| CD74 | CD74 molecule, major histocompatibility complex, class II invariant chain | Major histocompatibility complex, class II invariant chain protein | Upregulation of CD74 is associated with various types of cancer, including lymphoma and leukemia |
| CD79B | CD79B molecule | B-cell antigen receptor complex-associated protein beta chain | Associated with some cancers |
| | CD84 | | --- | | CD84 molecule | CD84 cell surface antigen | Upregulation of CD84 is associated with various types of cancer, including lymphoma and leukemia |
| CEACAM21 | Carcinoembryonic antigen-related cell adhesion molecule 21 | Carcinoembryonic antigen-related cell adhesion molecule 21 protein | Associated with some cancers |
| CNTN1 | Contactin 1 | Contactin 1 protein | Associated with some cancers |
| COPZ2 | COPI coat complex subunit zeta 2 | Coatomer subunit zeta-2 protein | Not associated with any cancers |
| CREBBP | CREB binding protein | CREB-binding protein | Upregulation of CREBBP is associated with various types of cancer, including leukemia and lymphoma |
| CUL7 | Cullin 7 | Cullin-7 protein | Associated with some cancers |
| CYP3A43 | Cytochrome P450 family 3 subfamily A member 43 | Cytochrome P450 3A43 enzyme | Associated with some cancers |
| DCN | Decorin | Decorin protein | Downregulation of DCN is associated with various types of cancer, including breast and lung cancer |
| DPM1 | Dolichol-phosphate mannosyltransferase subunit 1 | Dolichol-phosphate mannosyltransferase subunit 1 protein | Associated with some cancers |

There clearly was a difference (see figure 6) for many of the cancers, but was that just a false discovery? We moved on to the methods to find out.

**Part 2:** Benjamini Hochberg Analysis

We elected to go with a Benjamini Hochberg Analysis to see if our results were statistically significant and to mitigate the false positive discovery rate. See the code and results below (see figure 6)

*Figure 6- Benjamini Hochberg Analysis*

| **#t-test and FDR for 100 most common genes** pvec = NULL  gene = comparison$Gene.name  for (i in 1:length(gene)) {  a = subset(rnac.cancer, Gene.name==gene[i])  b = subset(rnac, Gene.name==gene[i])  for (j in nrow(a)) {  a$nTPM[j] = a$nTPM[j] - mean(b$nTPM)  }  t.test(a$nTPM, mu = 0)  pvec = c(pvec, t.test(a$nTPM, mu = 0)$p.value)  }  #taken from fungraphs.pck  fdr <- function(v1,Q, ind=F) {  #v1 is the vector of p values  #v1 is ordered  o1<-order(v1)  #sorted v1 is stored in pvec  pvec<-v1[o1]  #number of tests (length of v1) is stored in m  m<-length(v1)  #qline is set according to independence  qline<-Q\*c(1:m)/m #if independent  #if not independent, Q\*c(1:m)/(m\*(sum(1/c(1:m)))  if(!ind){  c1<-sum(1/(c(1:m)))  qline<-Q\*c(1:m)/(m\*c1)  }  #creates plot of points  plot(c(c(1:m),c(1:m)),c(qline,pvec),type="n",xlab="ordering",ylab="pvalue")  lines(c(1:m))  lines(qline, col = "red") #qline is drawn in red  points(c(1:m),pvec)  #calculates Pstar (pmax) and identifies all P<= pmax  dv<-pvec-qline  I1<-(dv<0) #I1 is made up of all pvalues below the line (less than qline)  I0<-I1 #I0 set equal to I1  if(sum(I0)>.5){ #tested at the .05 level  pmax<-max(pvec[I1]) #pmax = the largest pvalue below (less than) the qline  #interesting p-values (p-values less than pmax) are stored in I2  I2<-pvec<=pmax  #points with interesting p-values are plotted in cyan  points(c(1:m)[I2],pvec[I2],col="cyan")  #lists interesting tests  out<-list(interesting=o1[I2], ind=ind)  }  else{ #if no interesting p-values  #calculates posterior distribution  vec<-qbeta(c(.5,.95,.99,.999),1,length(v1)+1)  #lists .5 percentile, .95 percentile, and .999 percentile  out<-list(q.5=vec[1],q.95=vec[2],q.99=vec[3],q.999=vec[4])  }  out  } **fdr(pvec, .05, T)****fdr(pvec, .05, F)** |
| --- |
| **fdr(pvec, .05, T)**    **fdr(pvec, .05, F)** |

From these graphs, we can deduce that the pattern suggests that more significant results (lower p-values) are associated with a higher proportion of false discoveries (larger q-values); therefore, a slanting up q-line in the Benjamini-Hochberg graphs suggests that the statistical procedure is appropriately adjusting for multiple comparisons and controlling the FDR in both one tailed and two tailed comparisons. In other words, since the vast majority of out data points are below the red line, this dataset has relative low false discovery rates since the points with a low false discovery rate are below the qline, or desired false discovery rate

**Conclusion**

Cancer is an extremely complex disease, and our dataset represents that. We found that genes could be upregulated or downregulated, and that some genes that had significantly different nTPM values for cancer and non cancer were always associated with cancer, and some weren’t. Our Benjamini-Hochberg graphs suggests that the statistical procedure is appropriately adjusting for multiple comparisons and controlling the FDR in both one tailed and two tailed comparisons and that we most likely had a low false discovery rate.Genes are complicated entities- they can affect each other in unexpected ways. Our report and analysis most likely fails to do this justice, but it was still interesting to see what genes had a stronger relationship with cancer.

**Acknowledgements**

Thank you very much to Dr. Luvalle and Hirofumi for all of their support and assistance this semester.

**Dataset Source**

Dataset: <https://www.proteinatlas.org/about/download>

* See datasets 22 and 23