# Análisis de cancer de colon en los adultos jovenes

Análisis de biología computacional, grupo 12 (BT1013.12) Equipo 1

Carlos Emiliano Brito Nieto A01632840 Jessica Nicole Copado Leal A01637876
Aldo Alejandro Degollado Padilla A01638391 Ulises Venegas Gómez A01637321

01 mayo 2020

### Inicializacion de informacion

```
#setwd("Lugar donde se encuentran todos los archivos")
setwd("C:/Users/G9-593/Documents/Universidad/Analisis Biologia/Reto")
library(ggplot2)
library("gplots")
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

### Introducción

El colon es la porción más larga del intestino grueso en el sistema digestivo, el cual está dividido en diferentes segmentos: ascendente, transverso, descendente y sigmoide (en donde se une con el recto, terminando en el ano). Dichos segmentos son provocados por filas de saculaciones, también llamadas haustras. Las funciones del colon son absorber agua y electrolitos, formar y almacenar las heces para poder ser expulsadas por el ano, y digerir químicamente ciertos nutrientes con la ayuda de bacterias intestinales. (Nigam, 2019).

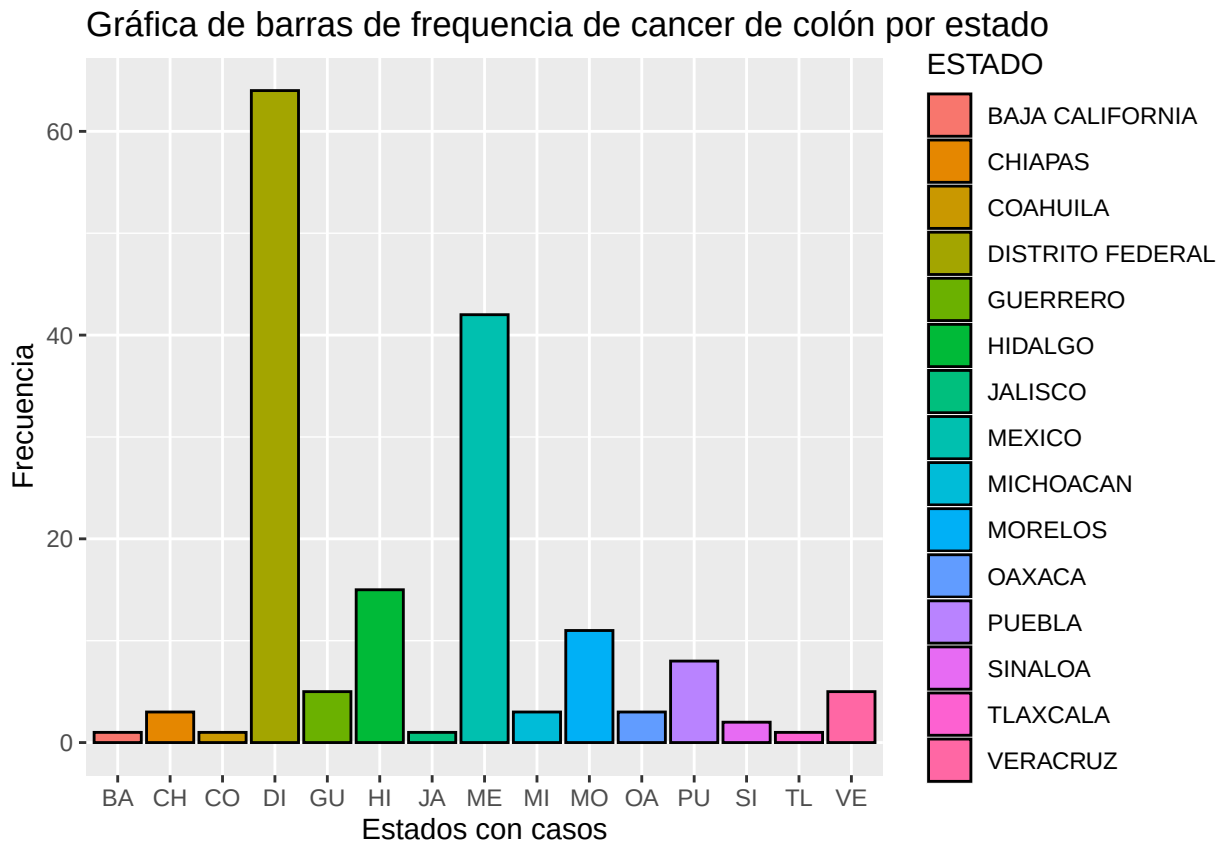
El cáncer es una enfermedad que consiste en la formación de un tumor, o una acumulación de células cancerígenas, en alguna parte del cuerpo. Este amontonamiento de células es causado por mutaciones en los genes del ADN de las células, convirtiéndolos en oncogenes. A diferencia de un gen normal, los oncogenes no detienen el crecimiento de las células y las células cancerígenas tienden a tener mutaciones constantes en su ADN. Los tumores malignos pueden extenderse a otras partes del cuerpo a través del flujo sanguíneo o del sistema linfático, mientras que los tumores benignos no se esparcen y una vez extraídos del cuerpo, generalmente, no vuelven a crecer. El cáncer, ya que se trata de una enfermedad genética, puede ser heredada o también puede que se desarrolle durante la vida de una persona por factores externos, tales como el tabaco (cáncer de pulmón) o la radiación. (National Cancer Institute, 2015).

Algunas riesgos que influyen en el desarrollo del cáncer de colon son: edad, historial familiar de cáncer colorrectal, historia personal, riesgo heredado, alcohol y raza; no obstante, las conductas de riesgo que más destacan son fumar cigarrillos, tener sobrepeso y no hacer suficiente ejercicio. De la misma manera, existen las conductas preventivas para evitar que ocurran este tipo de anomalías en las células. (National Cancer Institute, 2019)

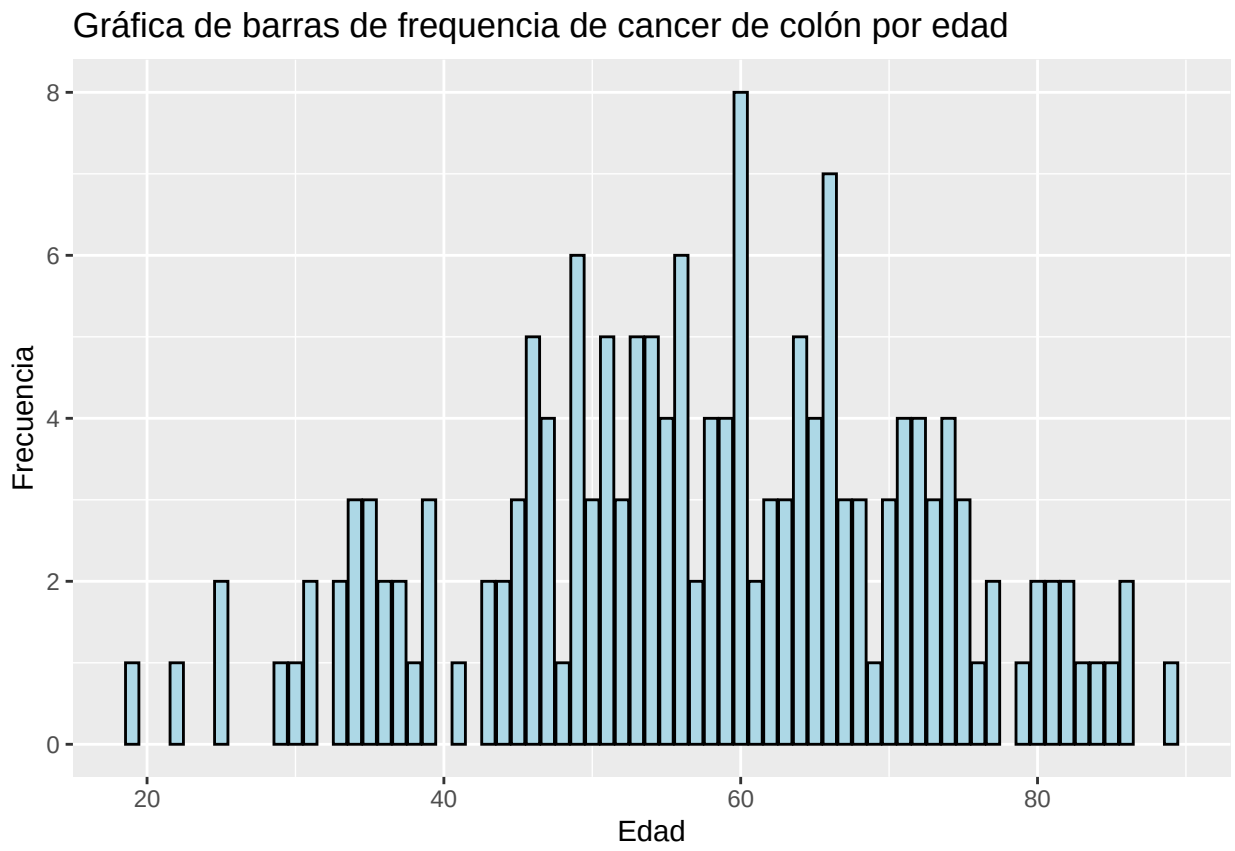
**Antecedentes**

```
pacientesNuevoIngr <- read.csv("9_PACIENTES_DE_NUEVO_INGRESO.csv")

#Se arregla la base de datos para contener solo cancer
x<-grep("(TUMOR|OMA)",pacientesNuevoIngr$DESCRIPCION.DIAGNOSTICO)
y<-grep("PAPILOMA",pacientesNuevoIngr$DESCRIPCION.DIAGNOSTICO)
TodosCancer <- x[!x %in% y]
contadorCasosCancer <- length(TodosCancer)
#contadorCasosCancer
#Unicamente tumor de cancer
aux_colon_recto <- pacientesNuevoIngr[grep("COLON|RECTO", pacientesNuevoIngr$DESCRIPCION.DIAGNOSTICO),]
#aux_colon_recto
contadorCasosColonComp <- length(aux_colon_recto$FOLIO)
#contadorCasosColonComp
#Gráficas de cancer de colón por estado y por edades
#Cancer de colón por estado
plot_of_bars1 <- ggplot(data=aux_colon_recto, aes(x=strtrim(ESTADO, 2)))
plot_of_bars1 + geom_bar(color="black",aes(fill = ESTADO )) +
  xlab("Estados con casos") +  ylab("Frecuencia") + ggtitle("Gráfica de barras de frequencia de cancer
```




```
#Cancer de colon por edades
plot_of_bars2 <- ggplot(data=aux_colon_recto, aes(x=EDAD))
plot_of_bars2 + geom_bar(color="black",fill = "lightblue" ) +
  xlab("Edad") +  ylab("Frecuencia") + ggtitle("Gráfica de barras de frequencia de cancer de colón por
```




```
contadorCasosJovenes = 0
for(i in 1:length(aux_colon_recto$EDAD)){
  if(aux_colon_recto$EDAD[i] < 50){
    contadorCasosJovenes = contadorCasosJovenes + 1
  }
}
#contadorCasosJovenes
porcentajeJovenes = contadorCasosJovenes/contadorCasosColonComp*100
#porcentajeJovenes
```

De acuerdo a las gráficas anteriores, hechas con la base de datos del Instituto Nacional de Cancerología del 2016. (INCAN, 2017), podemos ver que la cantidad de casos de cáncer de colon en la población mexicana es alrededor de 165 personas con cáncer de colon de un estudio de 2625 personas con cáncer (2811 personas totales), y de las cuales 48 son adultos menores a 50 años, lo que quiere decir que aproximadamente el 29.09% de todos los casos registrados de cáncer de colon en el 2016 fueron de adultos jóvenes. Esto quiere decir que de cada 10 casos de cáncer de colon, 3 de ellos pueden ser jóvenes menores a 50 años.

Otro dato que se tiene que tomar en cuenta es que se registraron la mayor cantidad de casos de cáncer de colon en el Distrito Federal, Ciudad de México, e Hidalgo, siendo estas ciudades zonas de gran concentración de gente y de problemas sociales, económicos y especialmente ambientales. La importancia de un kit de diagnóstico temprana no puede ser subestimada si existe una gran prevalencia de jóvenes con cáncer de colon comparativo a la población total del estudio.

**Justificación**

La realización de este proyecto sería de alta importancia en distintas dimensiones, a continuación se describirán con más detenimiento la importancia de este en cada uno de tres sectores de alto impacto en la vida diaria. La primera de esas tres dimensiones es la Social. El cáncer de colón genera un impacto social puesto que este es uno de los cánceres con más casos a nivel mundial y el número de muertes a causa de este también es alto, pero no solo esto sino que el grupo social que sigue siendo el más afectado por esta enfermedad es el de los adultos mayores o personas mayores de edad, aunque en los últimos años los casos en pacientes más jóvenes han ido en aumento y el cáncer llega a ser más agresivo (Amri, Bordeianous y Berger, 2015). Por lo que el desarrollo de este proyecto ayudaría a la generación de un análisis sobre esta enfermedad, el cual podría ser empleado para más investigación del funcionamiento de esta enfermedad para ayudar a salvar vidas de personas, es por eso que tendría un impacto social.

La siguiente dimensión sería la Económica. En cuanto a este aspecto este proyecto también sería de alta pertinencia puesto que lo relacionado a la investigación de la salud y la industria farmacéutica es una gran oportunidad de negocio puesto que estos generan grandes cantidades de dinero, pero las inversiones que se deben de realizar para hacer eso posible también son muy altas. Pero no solo es importante por lo que se acaba de mencionar si no que además el cáncer hoy en día representa un problema muy grande por las grandes cantidades de personas que se ven afectadas y que fallecen a nivel mundial por lo que de esta manera se sabe que el análisis e investigación relacionada a esta enfermedad si generaría una ganancia porque se puede decir que existe un mercado para esta puesto que va a llegar a ser utilizada.

En cuanto a la dimensión Ambiental, la importancia del proyecto en este dependerá de los recursos que se utilicen en la investigación puesto que por ejemplo en caso de que se use algún laboratorio para el análisis sería necesario usar ciertos materiales para poder hacer esto. No sólo se genera desperdicio por esos materiales que se usan, también algunos de los materiales que se emplean pueden ser tóxicos o contaminantes por lo que aumentan aún más el impacto que se generaría en el ambiente. En caso de que no se utilice algún tipo de experimentación en laboratorio no se tendría esa generación de desperdicio y solo se usarían computadoras para poder usar información y bases de datos que ya existen, de esta manera solo se estaría consumiendo energía (electricidad) para poder hacer el trabajo. Aun así en caso de no tener un uso eficiente de la energía y el tiempo que se usan estas se puede llegar a generar un alto gasto de energía de manera innecesaria, por lo que se podrían implementar planificaciones previas al trabajo para poder tener un uso más eficiente del tiempo con el que se trabaja en las computadoras y de esta manera disminuir el gasto de energía.

Según un artículo del IMSS que se publicó en 2019 en nuestro país se registran cerca de 15,000 casos nuevos de cáncer de colon cada año. Esta es una cifra muy elevada que nos demuestra la importancia del estudio del cáncer de colon en nuestro país. Hoy en día los avances tecnológicos son una herramienta de gran utilidad para esto, puesto que se tienen una gran cantidad bases de datos con información genética sobre muchos pacientes con esta enfermedad y de personas que no sufren de esta enfermedad. Esto nos sería de mucha utilidad puesto que así podemos ir comparando los genes usando herramientas como lo es RStudio. De esta manera podemos ir viendo los genes en los que se generan irregularidades o mutaciones en las personas con esta enfermedad. Esto podría ser especialmente útil para conocer mejor sobre qué tanto más propenso es de contraer esta enfermedad una persona con familiares que la padecen, así sabríamos sobre el riesgo genético que tendrían y de esta manera se podría tener un mayor monitoreo para en caso de que también llegue a tener la enfermedad se tenga una detección temprana. No solo nos serviría para saber si por herencia alguien corre riesgo pero también analizando los genes podríamos llegar a saber si gente con alguna cualidad genética específica corre también un mayor riesgo, de esta manera también se podría tener una detección más temprana de la enfermedad para así tener una mayor probabilidad de salvar su vida. Así que básicamente esto nos serviría para poder generar una base de datos con grupos de personas con ciertas características genéticas que son más propensos a tener cáncer de colon, para darles un monitoreo durante su vida y poder llegar a tener una detección temprana de esta enfermedad.

**Objetivo**

En los años recientes, el cáncer de colón en adultos menores a 50 años a ha ido en aumento a pesar de que los casos mundiales han ido disminuyendo recientemente. (Kasi y otros, 2018). El objetivo de este proyecto es hacer un análisis genético que permita concluir y reportar las características principales comunes que tienen los pacientes, así como resaltar la importancia del análisis estadístico en la medicina y biología.

**Métodos**

Lo primero que se debe de hacer al trabajar con una base datos es utilizar la función setwd() para cambiar la ubicación en donde se tiene guardado el archivo. Después mediante la función grep() buscamos muestras normales y muestras tumorales y las asignamos a variables distintas. Posteriormente, mediante la función apply() hacemos que las filas de la base datos sean evaluadas con la función x. Utilizamos la función mean() para calcular las medias tanto de las muestras normales como de las muestras tumorales y simplemente procedemos a restarlas. Se hace uso de la función abs() para asegurarnos de que el número resultante de la resta sea positivo, ya que el dato que buscamos es la diferencia. Guardamos los tres resultados obtenidos (media de muestras normales, media de muestras tumorales y diferencia de medias) en una nueva variable, lo cual nos generará una matriz de tres filas y n número de columnas, dependiendo de la cantidad de muestras normales y tumorales del archivo, en este caso fueron 15,970. Como nosotros creemos que es más fácil observar una matriz hacia abajo (con un número mayor de filas que de columnas) que hacía los lados (con un número mayor de columnas que de filas), decidimos utilizar la función t() para obtener la transpuesta de la matriz. A continuación, renombramos las tres columnas de nuestra matriz resultante (Tumor_Colorectal, Normal_Colorectal, Diff) utilizando la función colnames(). Para finalizar, simplemente utilizamos la función order(variable[3], decreasing=TRUE) en nuestra variable donde almacenamos las medias y las diferencias de medias, para ordenar las filas en base los datos de la tercera columna (diferencias de media) de manera descendente (de mayor a menor). Y de esta forma se obtienen genes con mayor diferencia de medias.

Como ya se dijo anteriormente, y debido a que vamos a trabajar con una base de datos, lo primero que se debe de hacer es utilizar la función setwd() para cambiar la ubicación por la dirección en la que se tiene guardado el archivo. Lo siguiente que debemos hacer es crear una matriz mediante la función matrix() que se llamará "matriz_ttest" el mismo número de filas que las que tiene la base de datos (que en este caso son 20,502) y con cuatro columnas. Dentro de esta función creamos una lista de dos dimensiones que contendrá los nombres de las filas de nuestra nueva matriz (nombres de las filas de la base de datos) y los nombres de las columnas ("Old", "Young", "P-Value", "FC" ). Esta lista se iguala al argumento dimnames de nuestra función matrix() para que de ella tome los nombres para nombrar tanto filas como columnas. Lo siguiente que hacemos es utilizar la función which() para encontrar los genes que pertenecen a adultos jóvenes y a adultos mayores. Para esto simplemente buscamos los datos de la columna de clases de la base de datos que contengan la palabra "YOUNG" o "OLD". Los datos que cumplan con la condición son almacenados en dos variables, "y_muestras" ó "o_muestras" dependiendo de si son pertenecientes al grupo de adultos jóvenes o al de adultos mayores. Después mediante la utilización de un for que irá de uno hasta el número de filas de la base de datos (20,502) iremos llenando la matriz creada anteriormente fila por fila. En este for lo que se hace es que por cada iteración (i) se hace uso de la función t.test(), ingresando en ella un el valor número i almacenado en la variable y_muestras y el valor de un número i almacenado en la variable o_muestras. A continuación, mediante el valor de salida estimate de la función t.test() obtendremos las medias del valor perteneciente a la variable y_muestras y del valor perteneciente a la variable o_muestras. Para terminar con esta for, se ingresan al renglón i de la matriz_ttest la media del valor perteneciente a la variable y_muestras, la media del valor perteneciente a la variable o_muestras, el valor de salida p.value de la función t.test() y la diferencias entre las medias antes mencionadas. Después ordenamos la matriz en base al p-value (de menor a mayor) y se obtienen los índices de los p-value (de menor a mayor), estos índices se asignan a una nueva variable. Posteriormente, se procede a observar si la diferencia de medias, perteneciente a los índices obtenido en el paso anterior, son mayores o menores a cero (positivos o negativos). Si son mayores a cero, quiere decir que el gen es más expresado en los adultos, y si son menores a ceros, quiere decir que el gen es más expresado en jóvenes. También se establece un filtro en el que sí el p-value no es menor a 0.01, ese índice no se toma en cuenta, ya que no es suficientemente seguro. Se crean dos variables, una con los datos

que su p-value es más pequeño que 0.01 y que su diferencia de medias es mayor a cero y la otra con los datos que su p-value es más pequeño que 0.01 y que su diferencia de medias es menores a cero. Gracias a esto ya podemos conocer los veinte datos que tienen más diferencia en sus medias y con ellos realizar un heatmap. Este tipo de gráfica la utilizaremos para tomar los dos conjuntos de datos que conseguimos antes (los genes de los jóvenes y los genes de los ancianos) y compararlos para observar qué genes, de los veinte con menor p-value, son los que se hacen más presentes tanto en jóvenes como en ancianos. Mientras más rojo esté un sector del heatmap, más se expresan esos genes en jóvenes en comparación a los ancianos y mientras más azul esté un sector, más se expresa esos genes en ancianos en comparación a jóvenes.

**Resultados y discusión**

```
#Cargar archivo
all_samples_diff <-load(file = "Multi_Cancer_Data.RData")
#Todas las muestras normales y de tumor de colon/recto separadas
tumor_colon <- grep("Tumor__Colorectal",colnames(multi_cancer_data))
normal_colon <- grep("Normal__Colon", colnames(multi_cancer_data))
#Se genera una matriz con la media normal, media del tumor y sus diferencias por gen para cancer de col
colon_samples_diff <- t(apply(multi_cancer_data, 1, function(x){
  m_tumor <- mean(x[tumor_colon], na.rm=T)
  m_normal <- mean(x[normal_colon], na.rm=T)
  diff_m <- abs(m_tumor - m_normal)
  c(m_tumor, m_normal, diff_m)
  }))
#Asignar nombres a las columnas
colnames(colon_samples_diff) <- c("Tumor_Colorectal", "Normal_Colorectal", "Diff")
#Se orden los genes de mayor a menor diferencias
colon_order_genes <- order(colon_samples_diff[,3], decreasing=T)

#Finalmente, se muestran los diez genes con mayor diferencia de expresión entre tejido normal y tumoros
knitr::include_graphics("Tabla Top 10 genes con más diferencia de medias (cáncer de colon).png")#Se uti
```

| | Tumor_Colorectal <dbl> | Normal_Colorectal <dbl> | Diff <dbl> |
|---|---|---|---|
| Selenium binding protein 1_RC_AA290679_at | 0.129818182 | 3.533909 | 3.404091 |
| DRA Down-regulated in adenoma_L02785_at | 0.392090909 | 3.630545 | 3.238455 |
| HBG2 Hemoglobin gamma-G _M10050_at | 0.723636364 | 3.812000 | 3.088364 |
| EST: zh89b04.s1 Soares fetal liver spleen 1NFLS S1 Homo sapiens cDNA clone 428431 3', mRNA sequence. (from Genbank)_RC_AA004415_at | -0.175090909 | 2.911545 | 3.086636 |
| Vasoactive Intestinal Peptide_HG2987-HT3136_s_at | -0.175636364 | 2.818182 | 2.993818 |
| IgG Fc binding protein_D84239_at | 0.005363636 | 2.910182 | 2.904818 |
| IgG Fc binding protein_D84239_at-2 | 0.005363636 | 2.910182 | 2.904818 |
| Carcinoembryonic antigen family member 2, CGM2_X98311_at | 0.234000000 | 3.046818 | 2.812818 |
| Alcohol dehydrogenase 3 (class I), gamma polypeptide_M12272_s_at | -0.333454545 | 2.416000 | 2.749455 |
| CA2 Carbonic anhydrase II_Y00339_s_at | -0.155818182 | 2.503636 | 2.659455 |

10 rows

```
#colon_samples_diff[colon_order_genes[1:10],]
```

El código muestra la tabla con el top 10 de los genes con mayor diferencia de medias entre las muestras de los pacientes sanos y los pacientes con tumor en el colon.

Se han visto reducciones en estudios previos de cáncer del primer gen, Selenium-binding protein 1 (SBP1). SBP1 es una proteína muy importante a nivel celular que une covalentemente el selenio. Sus funciones incluyen la diferenciación celular y la proteólisis (degradación de proteínas), y se podría considerar como un supresor de tumores por la reducción de esta proteína en pacientes con cáncer. Es por esto que existe una gran diferencia entre las medias de los pacientes sanos y enfermos. (Elhodaky y Diamond, 2018).

También se muestra una escasez del segundo gen mostrado, down-regulated in adenoma (DRA), en pacientes con cáncer de colon. Este gen produce una proteína encargada de transportar iones y se encuentra en la

mucosa del tracto gastrointestinal, y es particularmente importante en las etapas tempranas del desarrollo de un tumor porque se puede diferenciar entre la mucosa normal y el adenocarcinoma. (Antalis y otros, 1998).

El Vasoactive Intestinal Peptide (VIP) es el encargado de varias funciones en el sistema inmunológico, como controlar la homeostasis mediante receptores diferentes. Tiene un efecto general desinflamatorio y ha sido propuesto para tratar enfermedades autoinmunes como el lupus o la esclerosis múltiple. La ausencia de este neuropéptido se puede reconocer en pacientes con cáncer, ya que, la quimioterapia puede causar que este se debilite. Por lo tanto, la media de los pacientes sanos y enfermos va a ser claramente diferente. (González y otros, 2007)

Podemos decir que los genes mostrados tienen la mayor diferencia de medias porque los pacientes con cáncer sufren una reducción de estas proteínas que son de gran importancia en las células, las cuales en un paciente sano los niveles que muestra son óptimos. Estos datos son de gran ayuda para la continuidad de los estudios sobre el cáncer colorrectal.

```
load(file = "TCGA_COADREAD_comp_data.RData")
#Se cargan y separan las muestras de jovenes y adultos, para hacer el t-test que separara los 20 genes
y_muestras = which(tcga_coadread_class == "Young")
o_muestras = which(tcga_coadread_class == "Old")
#Se genera una matriz para poner los datos de medias, el Fold change y el P-value del t-test que se va
matriz_ttest <- matrix(NA, nrow =nrow(tcga_coadread),ncol = 4,
                       dimnames = list(rownames(tcga_coadread),c("Old", "Young", "P-Value","FC")))
#Se rellena la matriz con un for
for(i in 1:nrow(tcga_coadread)){
  aux_ttest <- t.test(tcga_coadread[i,y_muestras],tcga_coadread[i,o_muestras])
  aux_ttest_y <- aux_ttest$estimate[1]
  aux_ttest_o <- aux_ttest$estimate[2]

  matriz_ttest[i,] <- c(aux_ttest_y,aux_ttest_o,aux_ttest$p.value, (aux_ttest_y-aux_ttest_o))
}
#Se toman los 20 primeros genes con el menor P-test, es decir, con la mayor diferencia de expresión de
#matriz_ttest[order(matriz_ttest[,"P-Value"])[1:20],]
#Se diferencias los primeros 20 genes de si son expresados más en jovenes o adultos
index_order_pvals <- order(abs(matriz_ttest[,"P-Value"]))


#Se toman los primeros 20 genes menos expresados
print("Genes por P-value para Heatmap")
```

```
## [1] "Genes por P-value para Heatmap"
```

```
matriz_ttest_pval <- matriz_ttest[index_order_pvals[1:20],]
data.frame(matriz_ttest[index_order_pvals[1:20],])
```

```
##                      Old      Young      P.Value          FC
## MTERF         8.39193346  7.6258149 8.054799e-07  0.76611852
## PRND          4.16365896  2.9448635 4.748019e-06  1.21879543
## FZD9          4.51446092  3.3978873 1.000739e-05  1.11657367
## MLF1          8.38633153  7.4555142 1.630553e-05  0.93081730
## TBC1D3P2      0.01050615  0.0882879 6.441823e-05 -0.07778175
## PCSK1N        5.16845587  3.5492721 7.661858e-05  1.61918372
## LOC100009676  7.22734164  7.0113538 9.199702e-05  0.21598788
## PIWIL1        4.27838391  5.7870158 9.945316e-05 -1.50863187
```

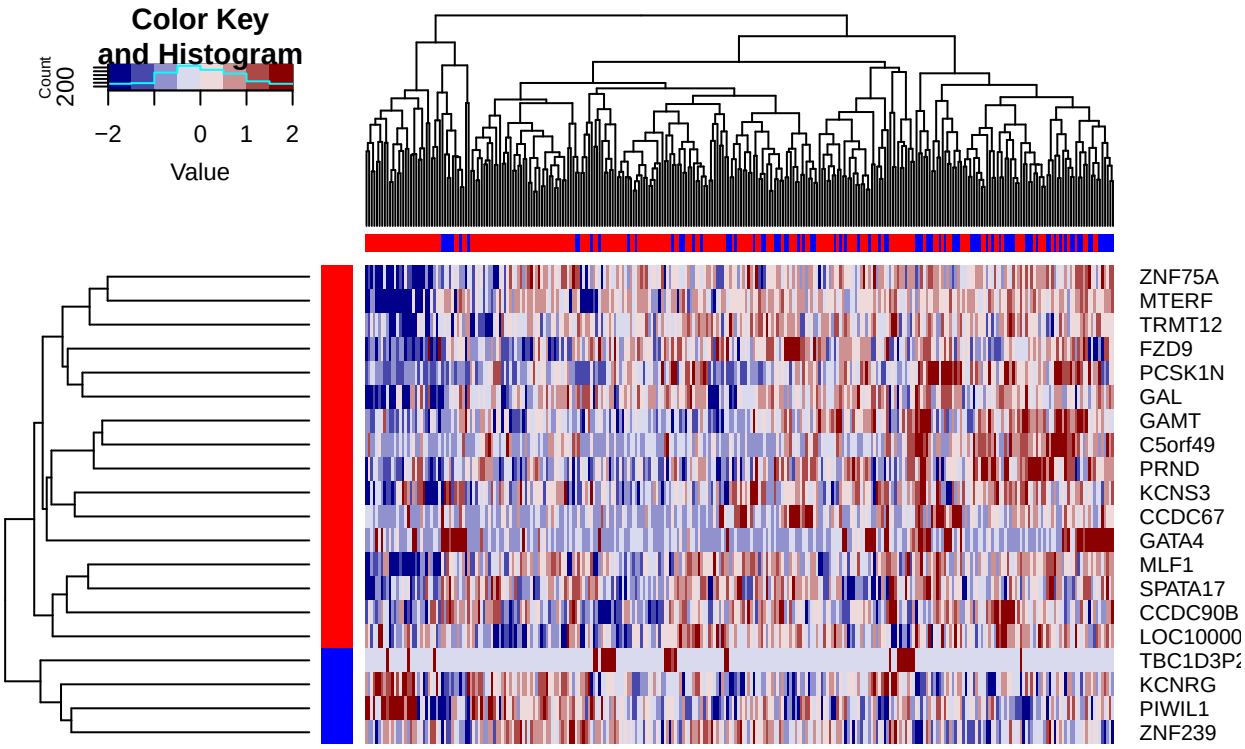
```
## KCNRG         4.13745416  4.6079456 1.726251e-04 -0.47049146
## ZNF239        7.58108540  8.0794047 1.856189e-04 -0.49831925
## KCNS3         8.48503259  7.8863608 1.893069e-04  0.59867177
## C5orf49       1.52838523  0.9461492 2.356647e-04  0.58223598
## CCDC90B      10.35628610 10.1679943 2.933591e-04  0.18829181
## TRMT12        9.10249051  8.7368068 3.015905e-04  0.36568375
## GAL           7.78198621  6.6496453 3.363220e-04  1.13234089
## SPATA17       4.79342438  4.2025535 3.523368e-04  0.59087087
## GATA4         3.31701100  1.5999120 3.551113e-04  1.71709902
## ZNF75A        8.06511522  7.6697710 4.194081e-04  0.39534427
## GAMT          6.18290849  5.5904200 4.682134e-04  0.59248846
## CCDC67        1.57932997  0.9437192 5.721168e-04  0.63561072
```

```r
#Se diferencian entre jovenes y adultos
index_de_high <- which(matriz_ttest_pval[,"P-Value"] < 0.01 & matriz_ttest_pval[,"FC"] > 0)
de_genes_high <- rownames(matriz_ttest_pval)[index_de_high]

index_de_low <- which(matriz_ttest_pval[,"P-Value"] < 0.01 & matriz_ttest_pval[,"FC"] < 0)
de_genes_low <- rownames(matriz_ttest_pval)[index_de_low]

#Finalmente se hace el heatmap
index_order_class <- order(tcga_coadread_class)
row_colors <- c(rep("red",length(de_genes_high)), rep("blue",length(de_genes_low)))
col_colors <- ifelse(tcga_coadread_class[index_order_class] == "Young","blue","red")
aux_mat <- tcga_coadread[c(de_genes_high, de_genes_low), index_order_class]

#Se agregan detalles
aux_mat <- t(apply(aux_mat, 1, scale))
colnames(aux_mat) <- colnames(tcga_coadread)
colors_h <- colorRampPalette(c("darkblue","white","darkred"))(8)
h_breaks <- seq(from=-2, to=2, length=9)
heatmap.2(aux_mat, col=colors_h, trace="none", breaks=h_breaks, labCol="",
ColSideColors = col_colors, RowSideColors = row_colors)
```




De los genes obtenidos a través del heatmap, podemos ver que solamente 4 de ellos son más expresados en adultos y 16 de ellos son expresados mayormente en jóvenes. En base a los mismos datos podemos decir que estos 20 genes tienen una expresión significativa para el estudio en la diferenciación en una base de datos efectiva, entre jóvenes y adultos. Los datos recolectados en el heatmap nos expresan también un posible interés adicional en explorar estos genes a más detalle, como el MTERF o el KCNRG, para encontrar posibles relaciones entre los genes y enfermedades como el cáncer y las razones por las cuales los adultos mayores son más propensos a enfermedades.

El análisis estadístico es una ciencia frecuentada debido a que utiliza datos reales para obtener porcentajes, promedios, desviaciones, etc. y estas aproximaciones son realmente útiles y confiables para utilizarlas con otros datos de la misma especie. En este curso nos enseñaron que una de la forma de evaluar la confiabilidad de nuestros datos mediante el análisis estadístico es el t test, ya que gracias al p-value que obtenemos al aplicarlo podemos saber que tan probable es que haya un error en el experimento. Si el p-value es mayor a 0.5, se suele decir que ese experimento no es confiable ya que hay bastante riesgo de un resultado no relacionado o que no demuestre nada relevante.

Los resultados del análisis de datos nos dan a entender que las expresiones de ciertos genes comparados tanto en jóvenes como adultos pueden modificar la manera en la cual se puede expresar una enfermedad. Algunos genes existen para proteger al cuerpo de células cancerígenas, pero las cuales son sujetas a mutaciones conforme uno crece. El funcionamiento de una célula normal se debe a que los genes más expresados en jóvenes, usualmente demuestran habilidades para contrarrestar la creación y proliferación de células cancerígenas. El hecho de que cada 3 de 10 personas con cáncer tengan cáncer de colon nos dice que hay un problema en las células de los adultos jóvenes y que las mutaciones en ciertos genes se están volviendo más frecuentes, con resultados catastróficos (La proliferación de cáncer).

## Conclusión Grupal

En México la cifra de casos al año de esta enfermedad es alta por lo que la creación de un Kit de diagnóstico para poder detectar esta enfermedad es importante puesto que este sería muy útil para poder tener una detección temprana de esta enfermedad para poder tener una mayor probabilidad de que la persona pueda ser atendida de forma adecuada y se pueda tratar la enfermedad. Las tecnologías computacionales son una gran herramienta para poder hacer esto puesto que con estas se es capaz de analizar mucha información, la cual se puede obtener de las distintas bases de datos existentes, de manera más ágil y efectiva. Pero esto no solo es útil en este caso, el uso de la tecnología se puede emplear en la investigación médica y biológica en general y hoy en día esta está teniendo un mayor crecimiento por todo el interés que rodea el tema de Big Data. También se genera un impacto a nivel social, económico y ambiental. En lo económico por que hoy en día ya existe un mercado para todo esto por lo que se puede tener un impacto en el mismo. En el social porque la tecnología ya es parte de nuestras vidas diarias y ha tenido un impacto en la sociedad y en la manera de pensar de esta en distintos aspectos. Y por último en el ambiental porque cualquier desperdicio y gasto de energía tienen un impacto en nuestro medio ambiente pero el que tan grande sea el impacto en esta depende directamente de qué tan grande sea el gasto o desperdicio.

## Conclusiones Individuales

### *Carlos Emiliano Brito Nieto*

La actividad 0 me dio a entender la legalidad y funcionalidad de las bases de datos para investigación y análisis, y la manera correcta en hacerlo. Las actividades Notebook 1-4 me ayudaron a tener un entendimiento general de cómo funciona Rstudio y el lenguaje R, guiandome en las siguientes actividades haciendo for's y estableciendo variables y el uso de métodos como length() y View(). La actividad 1 me dio a entender todo lo relacionado con genes, desde ácidos nucleicos, proteínas, genes, funciones, diferencia en expresiones, etc. Y me permitió desarrollar una idea más clara del proyecto. La actividad 2 y 3 me permitieron lidiar con el análisis de genes detallado, la importancia de las bases datos y la interpretación de datos adecuada para sacar conclusiones correctas en información analizada adecuadamente. La actividad 4 hizo caso lo mismo que la actividad 2 y 3, pero me permitió darle más importancia al análisis de genes específicos en base a Student's Test, dándole una mayor importancia todavía a ciertos genes con P-Values menores a 0.05. Finalmente, la actividad 5 me dio la habilidad de buscar más a detalle información de ciertos genes, comparándolos con mutaciones y generando una idea de cómo pueden afectar al ser humano. Todas las actividades me permitieron obtener información y patrones a base de habilidades fundamentales matemáticas (Como T-test) y computacionales (El uso de Rstudio). Entre las actividades 2 y 3, también entendí la diferencia y el valor que tienen los diferentes métodos de comparación estadística, siendo diferentes en valor dependiendo de la información utilizada.

### *Jessica Nicole Copado Leal*

A lo largo del período aprendimos información de valor y obtuvimos las herramientas necesarias para la resolución de la situación problema sobre el cáncer de colon en los jóvenes. Es por esto, que empezamos con la Actividad 0 en la cual leímos un artículo sobre Big Data y realizamos una breve reflexión de por qué nos gustaría trabajar en cierta empresa y que es lo que podríamos aportar con el uso de herramientas computacionales. También, aprendimos la estructura Project-Based Learning (PBL) la cual nos ayudó en recabar información con datos precisos sobre los genes, respondiendo a las preguntas del problema planteado.

En las siguientes actividades, lo que obtuvimos de aprendizaje fue el conocimiento de funciones fundamentales en R para poder hacer una investigación. Dentro de este aprendizaje se incluyen los loops que nos ayudan a la optimización y a que corra más rápido el código. Este control de flujo se utiliza, por ejemplo, cuando se quiere comparar el p.value() de cada gen y se tienen que revisar las medias de cada uno. También aprendimos funciones para hacer operaciones básicas como abs(), exp(), y summary(), el cual sirve para visualizar leer el resumen de un dataset. Lo último que aprendimos en esta actividad fue la creación de gráficas como de pastel, de barras, scatterplot, entre otras. Las gráficas son de suma importancia ya que, visualmente, nos proveen de información muy útil sobre los datos calculados. Por ejemplo, nos ayudó a ver la información de los genes con más diferencia de expresión en adultos jóvenes y adultos mayores.

Con la Actividad 2, pudimos analizar información real de pacientes con cáncer. Tuvimos que hacer unos cuantos ajustes a esta información para acotar los resultados y tener mejores resultados al momento de hacer el siguiente ejercicio, el cual se trataba de comparar los genes de órganos diferentes en pacientes sanos y en pacientes con cáncer. En esta actividad aprendimos a manejar de manera más eficaz la información que obtenemos al buscarla en bases de datos como NCBI o pubMED.

En la Actividad 3, seguimos trabajando con bases de datos y keywords que nos ayudan al momento de realizar una búsqueda. También lo trabajamos desde RStudio, en donde ubicamos los mismos datos pero en un código. Esta actividad, de igual manera, nos ayudó a que la búsqueda y recopilación de datos fuera óptima.

En la última actividad, la 4, utilizamos el conocimiento que obtuvimos las primeras semanas para hacer una búsqueda de genes y programar el script ideal para la obtención de sus medias. Aprendimos la importancia del t.test() y el p.value(), el cual usamos en varios puntos de este trabajo para poder hacer la comparación de sus medias. En mi opinión, esta actividad fue la más fructífera ya que también se usaron funciones estadísticas.

Todas estas actividades y el proyecto nos ayudaron a aplicar el conocimiento que obtuvimos estas cinco semanas sobre la biología y más específicamente, el cáncer de colon. Creo que las competencias se demostraron eficientemente al haber obtenido una exploración más completa de la biología usando herramientas computacionales como RStudio.

***Aldo Alejandro Degollado Padilla***

Durante estas cinco semanas aprendí un poco de biología y 'un mucho' del lenguaje R. Todos estos aprendizajes me otorgaron el conocimiento necesario para resolver satisfactoriamente el reto que se nos presentó al inicio de la materia. En la actividad 0 aprendí lo que se necesita para que nuestras investigaciones estén bajo el margen de la ley y dentro de las leyes morales. En esta actividad recopilamos algunas fuentes de las que sería seguro investigar y que nos ayudaron durante todo el tiempo que duró la materia. Esta actividad representó también el primer contacto con mi equipo de trabajo. Así que se podría decir que fue el pilar base para el resto de actividades.

En la actividad 1 fue en la que podría asegurar que más aprendí de biología. Fue aquí donde se nos introdujo el National Center for Biotechnology Information (NCBI) en donde nosotros podíamos consultar casi cualquier dato relacionado con el gen que quisiéramos. También aprendí que si buscamos bien hay herramientas muy útiles en internet que nos ayudan a obtener resultados rápidamente que de otra forma requerirían mucho tiempo. Un ejemplo es el software que nos ayudó a detectar repeticiones en la serie de nucleótidos de nuestros genes.

Una cosa que me ayudó mucho y no fue una actividad fue observar la explicación de la Dra. Yocan de cómo funcionaba el lenguaje R. Esto fue clave para que la realización a las posteriores actividades no fueran tan complicadas. Los Notebooks que se nos proporcionaron me ayudaron a entender la sintaxis (chunks, el uso de '<-' para asignar datos), conceptos (operadores matemáticos) y funciones básicas (abs(), summary(), length())de este lenguaje de programación.

En la actividad 2 comenzamos con la utilización de RStudio y aplicamos por primera vez los aprendizajes de este lenguaje. Esta ocasión fue la primera ocasión en la que trabajamos con una base de datos y la primera vez que instalamos una librería (ggplot), con ella elaboramos nuestras primeras gráficas en R. También aplicamos, y yo en lo personal comprendí mejor, algunas funciones como data.frame(), which(), grep(), names(). Algo que me pareció interesante fue que trabajamos con datos reales, eso me hizo sentir algo de lastima ya que cada número de esa base de datos que presentaba cáncer no era solo un número, era o fue una persona de verdad.

En la actividad 3 aprendimos continuamos con la utilización de funciones que ya conocíamos y otras nuevas. Yo creo que lo más relevante de esta actividad fue que aprendimos a buscar dentro de R con palabras clave, esto facilitaba el trabajo. También aprendimos a eliminar caracteres y palabras específicas con funciones como gsub().

En la actividad 4 aprendimos un método de análisis estadístico llamado t test que nos ayudó a concretar la actividad. También aprendimos que al aplicar este método se obtiene un dato llamado p-value que nos

ayuda a conocer la fiabilidad de nuestro experimento/cálculo. Esta actividad también fue nuestro primer contacto con los heatmaps.

Todos estos aprendizajes me ayudaron para ser un poco más culto en el mundo de la biología y a manejar otro lenguaje de programación que puede servirme mucho en el futuro cuando necesite analizar datos.

***Ulises Venegas Gómez***

La primera actividad que realizamos la cual fue la Actividad 0 me sirvió de base para algunos aspectos de la materia. A lo que me refiero es que por ejemplo con esta obtuvimos algo de conocimiento sobre el aspecto legal del uso de las bases de datos y también de cómo el Big Data está siendo usado para la salud, más en específico con algunas compañías de Silicon Valley. También nos hizo reflexionar sobre los conocimientos que ya teníamos y lo que pensabamos que ibamos a necesitar aprender. Con la siguiente actividad (Actividad 1) ya entramos más con el tema de los genes y utilizamos NCBI, por lo que me fui familiarizando con distintos aspectos de los genes puesto que tuve que investigar algunos de un gen en específico. También para poder hacer esto tuve que navegar por NCBI y ver donde se encontraban las distintas cosas que estaba buscando. Para la Actividad 2 ya usamos RStudio por lo que tuvimos que empezar a usar los distintos conocimientos básicos de este que obtuvimos durante la clase, por lo que aquí pusimos en práctica los conceptos básicos de RStudio como la lectura de bases de datos y la creación de gráficas. Con la siguiente usamos PUBMED con el cual hicimos una búsqueda de manera manual y también esa misma búsqueda la hicimos por medio de R para ver cómo se obtienen los mismo resultados y se podían buscar palabras clave en los artículos que obtenemos de la búsqueda. Con la Actividad 4 usamos el t-test para poder obtener el p. value, además de que comparamos los genes en pacientes con cáncer contra los de pacientes sanos. Y la última actividad sirvió para hacer comparaciones de distintos aspectos de los genes con un cáncer dentro de una misma base de datos. Todas estas me sirvieron para aprender a usar R el cual es muy bueno para lo relacionado a bases de datos y el análisis de las mismas, y para un mejor uso de R tuvimos que saber los estándares de este para poder hacer un buen código.

**Referencias**


Kasi, P., Shahjehan, F., Cochuyt, J., Li, Z., Colibaseanu, D. y Merchea, A. (2018). Rising Proportion of Young Individuals With Rectal and Colon Cancer. *Clinical Colorectal Cancer and other Gastrointestinal Malignancies*. Recuperado de https://doi.org/10.1016/j.clcc.2018.10.002

Gastrointestinal tract 5: the anatomy and functions of the large intestine. (2019). *Nursing Times*. Recuperado de https://www.nursingtimes.net/clinical-archive/gastroenterology/gastrointestinal-tract-5-anatomy-functions-large-intestine-23-09-2019/

What Is Cancer? (2015). *National Cancer Institute*. Recuperado de https://www.cancer.gov/about-cancer/understanding/what-is-cancer

Colorectal Cancer Prevention (PDQ)-Patient Version. (2019). *National Cancer Institute*. Recuperado de https://www.cancer.gov/types/colorectal/patient/colorectal-prevention-pdq

Pacientes de Nuevo Ingreso. (2017). *INCAN*. Recuperado de https://datos.gob.mx/busca/dataset/pacientes-de-nuevo-ingreso

Elhodaky, M y Diamond, A. (2018). Selenium-Binding Protein 1 in Human Health and Disease. *International Journal of Molecular Sciences*. Recuperado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6274749/

Antalis, T., Reeder, J., Gotley, D., Byeon, M. Walsh, M., Henderson, K, Papas, T. y Schweinfest, C. (1998). Down-regulation of the down-regulated in adenoma (DRA) gene correlates with colon tumor progression. *Clinical Cancer Research*. Recuperado de https://www.ncbi.nlm.nih.gov/pubmed/9717812

Amri, R., Bordeianou, L., y Berger D. (2015). The conundrum of the young colon cancer patient. *Surgery*. https://www.ncbi.nlm.nih.gov/pubmed/26298030

En México cada Año se Diagnostican Cerca de 15 Mil Nuevos Casos de Personas con Cáncer de Colon. (2019). IMSS. Recuperado de http://www.imss.gob.mx/prensa/archivo/201903/070