Tecnológico de Monterrey

Evidencia 1 | Reporte Escrito Proyecto Integrador

Daniel Godínez Morado	A01752790
Germán Guzmán López	A01752165
Isabel Vieyra Enríquez	A01745860
Manuel Julio Romero Olvera	A01752662
Antonio Oviedo Paredes	A01752114

27 de abril del 2021

Análisis de biología computacional

Profesora Ana Laura Torres Huerta

Profesor Leonardo Mauricio Cañete Sifuentes

Tecnológico de Monterrey

Introducción

Siendo el colon parte del sistema digestivo, el cáncer de colon es un tipo de cáncer que comienza en el intestino grueso. Usualmente comienza como un cúmulo de células benignas llamadas pólipos que surgen en la parte interior del colon y, con el paso del tiempo, estos pólipos pueden convertirse en cancerosos.

Los pólipos pueden ser pequeños y producir pocos síntomas. Es por ello que los doctores recomiendan realizarse pruebas con regularidad desde los 50 años para aquellas personas que tienen un riesgo promedio o antes para personas que tengan un riesgo mayor para prevenir el cáncer de colon al identificar y remover los pólipos antes de que se conviertan en cáncer.

En caso de que se desarrolle el cáncer de colon, existen varios tratamientos disponibles para controlarlo, incluyendo cirugía, terapia de radiación, y tratamientos con fármacos tales como la quimioterapia, terapia dirigida e inmunoterapia.

Las señales y síntomas del cáncer de colon incluyen:

- Cambio persistente en la evacuación intestinal, incluyendo diarrea, estreñimiento o un cambio de consistencia en las heces.
- Sangrado rectal o heces con sangre.
- Malestares abdominales persistentes, tales como dolor abdominal o gases.
- Sensación de que el intestino no se vacía completamente.
- Debilidad o fatiguez.
- Pérdida de peso sin explicación.

Muchas personas con cáncer de colon no experimentan síntomas en los primeros estadíos de la enfermedad. Cuando surgen los síntomas, varían dependiendo del tamaño del cáncer y su ubicación en el intestino grueso.

De manera general, el cáncer de colon comienza cuando las células sanas del colon desarrollan mutaciones en su ADN y, como el ADN de la célula contiene un conjunto de instrucciones sobre lo que hace la célula, se sigue dividiendo incluso cuando no se necesitan nuevas células y conforme estas células se van acumulando, formando un tumor.

Con el tiempo, estas células cancerígenas pueden crecer, invadir y destruir tejido sano cercano o viajar a otras partes del cuerpo para formar depósitos (metástasis).

Entre los factores de riesgo, se pueden mencionar el tener condiciones inflamatorias del intestino de manera crónica, haber padecido cáncer de colon previamente, heredar ciertos síndromes genéticos, una dieta baja en fibra y alta en grasas, diabetes, obesidad, alcoholismo, y radioterapia [1].

Dado que el cáncer de colon ha ido aumentando en sus casos de incidencia hasta convertirse en aproximadamente el 10% de la mortalidad relacionada con cáncer en países occidentales [2], es importante que se conozca de manera concreta su mecanismo e identificar aquellos genes que están relacionados con esta enfermedad para poder ofrecer un mejor tratamiento.

Para llevar a cabo este análisis se hace uso de microarreglos ya que enfermedades como el cáncer de colon están relacionadas con la actividad de múltiples genes. De manera fundamental, se miden los niveles de ARN mensajero (mARN) que está presente en una célula sabiendo que si un gen se encuentra transcrito en el mARN entonces está siendo expresado.

El microarreglo consiste en muestras de todos los genes del genoma. Para el caso del cáncer de colon, se pueden comparar células de tejido sano y células tumorales por lo que son de estas células de las que se extrae el mARN y convertido en ADN complementario (cADN) que es marcado con tinte fluorescente. Esto permite que el cADN de las células tumorales se muestren de color rojo mientras que el cADN de las células de tejido sano se muestren verdes.

Como las muestras están mezcladas, se hace uso de un láser para medir su grado de fluorescencia y estos datos son convertidos en datos numéricos dependiendo de la intensidad de rojo, verde o amarillo que tenga la muestra y se comparan los valores para determinar si se muestra más el rojo (muestra de tejido de tumor) o el verde (muestra de tejido sano).

Para realizar este análisis en múltiples muestras es necesario hacer uso de herramientas de estadística que requieren la expresión logarítmica de los datos, se calcula la estadística descriptiva de los valores de expresión de los genes en el microarreglo, se normalizan los datos y se calcula el producto punto para cada par de genes de manera que un puntaje positivo indica un

comportamiento similar de los genes, un puntaje de 0 si los genes no tienen relación en su comportamiento o un puntaje negativo que muestra que se comportan de manera opuesta, es decir, si uno está inducido, el otro está suprimido. Conforme se van realizando estas operaciones, se agrupan aquellos genes que muestran un comportamiento más parecido y se vuelven a calcular los puntajes de similitudes y a agrupar hasta que se obtiene solo un par de genes. La graficación de estas agrupaciones forma un dendrograma [3].

Para facilitar el cálculo de las operaciones y la graficación de los datos es en donde lenguajes de programación como R son sumamente útiles ya que nos permite hacerlo de manera mucho más rápida y precisa aun con computadoras personales e incluso con un mínimo de conocimiento en programación ya que en páginas como la de NCBI se han desarrollado herramientas web como GEO2R que identifican a aquellos genes que están diferencialmente expresados bajo ciertas condiciones dadas además de tener la opción de presentar resultados por medio de tablas y gráficos [4].

Estas ventajas son aprovechadas en una gran cantidad de estudios que tratan el cáncer de colon como el *Identification of hub genes and outcome in colon cancer based on bioinformatics analysis* [5] en donde exploran biomarcadores potenciales del cáncer de colon utilizando perfiles de expresión de genes en la base de datos Gene Expression Omnibus (GEO) que pueden ayudar a tener un diagnóstico más temprano o mejores tratamientos.

También en artículos como *Experimentally Derived Metastasis Gene Expression Profile Predicts Recurrence and Death in Colon Cancer Patients* [6] se hace uso de los perfiles de expresión de distintos genes tomados de células cancerígenas con alto índice de metástasis del modelo de un roedor para identificar a aquellos pacientes que padecieron cáncer de colon y el riesgo que tienen de recurrencia. El perfil de expresión del gen asociado a la metástasis fue refinado utilizando muestras de 55 pacientes como datos de entrenamiento y 177 pacientes como datos independientes. Esto resultó en un puntaje de metástasis asociado con un mayor riesgo de padecer metástasis o muerte debido al cáncer de colon.

Tecnológico de Monterrey

Descripción de los sets de datos

El set de datos se obtuvo de la página NCBI y la base de datos de GEO data sets filtrando con ayuda de conceptos clave como “Colon cancer”, “Homo sapiens” y “Expression profiling by array”.

Del dataset de 177 pacientes del Moffitt Cancer Center, con edades de 26 a 92 años, hombres y mujeres, clasificados en tres grupos, Well Differentiated (WD), Moderately Differentiated (MD) y Poorly Differentiated (PD) se analizaron 20 muestras de pacientes principalmente del grupo WD y MD. Cada muestra tiene aproximadamente 54675 renglones en donde se muestra el id y la intensidad de la señal, estos pacientes se dividieron en 4 etapas, nosotros analizamos la etapa 1 y 3 para que se notara la diferencia, dado que en la 3 es cuando un resultado más alto se relacionaba a la recurrencia del cáncer y se esperaba una mayor diferenciación al ir evolucionando el cáncer.

Desarrollo del código

Para el análisis de datos se empleó el lenguaje de programación R y las librerías de lima y GEOquery. Lo primero que se hizo fue generar un script de trabajo e importar las librerías anteriormente mencionadas.

El primer paso fue bajar el set de datos, con ayuda de la librería GEOquery, de la página NCBI.

```
1  library(GEOquery)
2  library(limma)
3
4  # 1 - Leer el conjunto de datos
5  gset <- getGEO("GSE17536", GSEMatrix =TRUE, AnnotGPL=TRUE)
6  if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else idx <- 1
7  gset <- gset[[idx]]
8
```

Con el set obtenido se acomodaron los valores de expresión de los pacientes en un data frame, donde las filas representan una sonda diferente del microarreglo y las columnas los pacientes. Como había sondas que no contaban con el símbolo del gen, primero se filtró el set de datos para eliminar estas sondas que podrían dificultar la interpretación de los resultados. Las columnas de este set se acomodaron en 2 grupos, de la 1 a la 10 los pacientes con cáncer en etapa 1 y de la 11 a la 20, los pacientes con cáncer en etapa 3.

```
# 2 - Seleccionar solo las sondas que cuentan con symbolo de gen
id=gset@featureData@data[["ID"]]
symbol=data.frame(gset@featureData@data[["Gene symbol"]])
row.names(symbol) = id
gene_id = id[symbol != ""]
gene_symbol = data.frame(symbol[symbol != ""])
row.names(gene_symbol) = gene_id

# 3 - Obtiene los valores de expresión
ex <- exprs(gset)

# 4 - Seleccionar 10 muestras de pacientes en etapa 1 y 10 en etapa 3
ex <- ex[,c(2,4,7,15,23,41,46,49,58,79,26,35,50,51,55,65,92,95,104,110)]
ex <- ex[row.names(gene_symbol),]
```

Para normalizar estos valores de expresión primero se calculó la media de cada columna y se dividió el valor de expresión entre el promedio correspondiente.

```
# 5 - Normalizacion
raw_means = apply(ex,2,mean,trim=0.02)
```

Con los valores normalizados se creó otro data frame con el cálculo del promedio de expresión por sonda y por grupo. Para esto se seleccionaron las columnas que corresponden a los pacientes con cáncer en etapa 1 y se calculó la media por cada fila. Después se realizó el mismo proceso pero con las columnas de cáncer en etapa 3.

```
# 6 - Medias
cancer_estado1_mean = rowMeans(microarray_norm[,1:10])
cancer_estado3_mean = rowMeans(microarray_norm[,11:20])

microarray_means = data.frame(cancer_estado1_mean, cancer_estado3_mean)
```

Para saber qué genes se sobreexpresan o subexpresan se calcularon las proporciones dividiendo los promedios de expresión de la etapa 3 entre la etapa 1, donde el valor de resultado representa cuántas veces más se expresa el gen en la etapa 3 con respecto a la 1.

```
# 7 - Proporciones
cancer_ratios = microarray_means$cancer_estado3_mean / microarray_means$cancer_estado1_mean

microarray_ratios = data.frame(cancer_ratios)

row.names(microarray_ratios) = row.names(ex)
```

Con estas proporciones es un poco difícil interpretar cuantas veces más se expresa el gen en una etapa con respecto a la otra. Por este motivo los valores se pasaron a escala de logaritmo base 2, que arrojaron un resultado más fácil de interpretar. Si el valor es positivo, significa que el gen se expresa 2 veces más en la etapa 3, y si el valor es negativo, significa que el gen en la etapa 3 se subexpresa 2 veces con respecto a la primera etapa.

```
# 8 - Cambio a log2
microarray_norm_log2 = log2(microarray_norm)
microarray_means_log2 = log2(microarray_means)
microarray_ratios_log2 = log2(microarray_ratios)

microarray_ratios_log2_v = unlist(microarray_ratios_log2)
names(microarray_ratios_log2_v) = row.names(microarray_ratios)
```

Para justificar que hay una diferencia estadística razonable entre las expresiones de los genes se calculó el p-value de cada sonda comparando, con ayuda de la función “t.test”, los valores normalizados de expresión de genes de los 2 grupos.

```
# 9 - t-test
get_pvalue <- function(values, idx1, idx2) {
  return(t.test(values[idx1], values[idx2])$p.value)
}

cancer_p = apply(microarray_norm, 1, get_pvalue, 1:10, 11:20)
```

Una vez obtenidos estos valores, se filtraron las sondas con un umbral de p-value<0.05 para obtener los genes que tengan una expresión estadísticamente diferente.

```
# 10 - Filtrar genes con valor de significancia p < 0.05
filtered_cancer_p = cancer_p[cancer_p < 0.05]
```

También se obtuvieron los genes que tenían un tamaño de efecto grande, para esto se filtraron las sondas que contaban con una proporción en base log2 mayor a 0.38 y menor a -0.38. Este valor se eligió para seleccionar los genes que se expresan un mínimo de 1.# veces con respecto a la muestra de cáncer en etapa 1. Como los valores se pasaron a logaritmo base 2 el valor que representa una expresión de 1.3 veces está dada por $log_2(1.3)$, que es aproximadamente igual a 0.38.

```
61
62  # 11 - Filtrar genes con tamano de efecto grande log2 > 0.38 y log2 < -0.38
63  filtered_cancer_log2 = c(microarray_ratios_log2_v[microarray_ratios_log2_v > 0.38],
64                           microarray_ratios_log2_v[microarray_ratios_log2_v < -0.38])
65
```

Con estas 2 listas de sondas filtradas se obtuvo una data frame que contiene el top 10 genes sobreexpresión-subexpresión, par esto se seleccionaron las 10 sondas con valores de log2 más altos y las 10 con valor de log2 más bajo que se encontraban dentro del filtro de p<0.01.

```
65
66  # 12 - Proporciones en log2 de sondas con p<0.05
67  sondas_log2_p = microarray_ratios_log2_v[names(filtered_cancer_p)]
68  slp = data.frame(sondas_log2_p)
69
70  # 13 - Top 10 genes que se sobreexpresan y subexpresan
71  sobreexp = sort(sondas_log2_p, decreasing = TRUE)[1:10]
72  subexp = sort(sondas_log2_p)[1:10]
73
74  genes_sobre_sub = data.frame("Genes sobre" = gene_symbol[names(sobreexp),],
75                               "valor exp sobre" = sobreexp, "Genes sub" = gene_symbol[names(subexp),],
76                               "valor exp sub" = subexp)
77  row.names(genes_sobre_sub) = c(1:10)
78
```

Otro data frame importante que se obtuvo contiene la lista de genes que consideramos como factor para el desarrollo de cáncer de colon. La lista se obtuvo comparando y filtrando las sondas con un p-value z 0.05 y una proporción en log2 > 0.38 y < -0.38. Esta lista representa los genes que tienen una diferencia estadística y un tamaño de efecto razonable entre cada etapa.

```
78
79  # 14 - Genes con valor de significancia p < 0.5 y tamaño de efecto grande.
80  genes_final_names = intersect(names(filtered_cancer_p), names(filtered_cancer_log2))
81  genes_final = data.frame("Gen" = gene_symbol[genes_final_names,],
82                           "Expresion en Log2" = microarray_ratios_log2_v[genes_final_names])
83
```

Para la generación de las gráficas primero se generaron 2 listas con las medias y la desviación estándar de la expresión de los genes candidatos. Con estos datos generó un nuevo data frame restando las medias individuales y el promedio general y dividiendo entre la desviación estándar. Con esta lista de datos se puede llamar a las funciones “hclust” y “heatmap” que generan las gráficas correspondientes.

```
# 15 - Generación de gráficas.
# Mapas de calor y dendogramas
microarray_selection = microarray_means[genes_final_names,]

medias = rowMeans(microarray_selection)
devs = apply(microarray_selection, 1, sd)

centered_microarray_selection = sweep(microarray_selection, 1, medias)
centered_microarray_selection = sweep(centered_microarray_selection, 1, devs, "/")

names(centered_microarray_selection) = c("cancer1", "cancer3")
hclustering = hclust(dist(centered_microarray_selection))

plot(hclustering)

names(microarray_selection) = c("E1", "E2")
heatmap(as.matrix(microarray_selection), Colv = NA,
        main = "Mapa de calor: Valores de expresión")
```

Para generar la gráfica de dispersión se llama a la función “plot” y dando como parámetros las medias de expresión en base log2 de cada etapa y el límite y nombre de los ejes. A su vez se graficó una línea roja que ayuda a interpretar la dispersión y expresión de los genes.

```
# Gráfica de dispersión
plot(microarray_means_log2$cancer_estado1_mean,
     microarray_means_log2$cancer_estado3_mean,
     xlim = c(0, 9), ylim = c(0, 9),
     xaxt = "n", yaxt = "n",
     main = "Gráfica de dispersión: Expresión en Etapa 3 vs Etapa 1",
     ylab = "Etapa3 (expresión en log2)",
     xlab = "Etapa1 (expresión en log2)")

axis(1, at = seq(0:9))
axis(2, at = seq(0:9))
abline(lm(microarray_means_log2$cancer_estado1_mean ~ microarray_means_log2$cancer_estado3_mean),
       col="red")
```

La gráfica R-I, al igual que la de dispersión, se obtiene mediante la función “plot”, pasando como parámetros la suma de los promedios de expresión log2, la diferencia entre los promedios de expresión en log2 de las dos etapas y los límites y títulos de los ejes.

```
# Gráfica R-I
plot(microarray_means_log2$cancer_estado1_mean + microarray_means_log2$cancer_estado3_mean,
     microarray_means_log2$cancer_estado1_mean - microarray_means_log2$cancer_estado3_mean,
     main = "Gráfica R-I: Expresión en Etapa 3 vs Etapa 1",
     xlab ="log2(Etapa 3 * Etapa 1)",
     ylab = "log2(Etapa 3 / Etapa 1)")
```

La última gráfica que se generó fue la de volcán, la cual, al igual que las últimas dos se genera mediante la función “plot” pasando como parámetros las proporciones de expresión en log2, los

p-values y los límites y nombres de los ejes. Para colorear los puntos que nos interesan se generó un vector que contiene el color del punto dependiendo si está entre los rangos definidos de los genes que se sobre expresan o subexpresan.

```
# Gráfica volcán
colores = rep(1, length(cancer_p))
colores[cancer_p < 0.05 & microarray_ratios_log2 < -0.38] = 2
colores[cancer_p < 0.05 & microarray_ratios_log2 > 0.38] = 3

plot(microarray_ratios_log2_v, cancer_p, col=colores, log = "y", ylim = rev(range(cancer_p)),
     main = "Gráfica de volcán: Sobre expresión y sub expresión",
     xlab = "Proporción de expresion en log2: Etapa 3 vs Etapa 1",
     ylab = "p-value")
```

Resultados

Un data frame que se obtiene es la lista de genes que tienen un valor de significancia $p<0.05$ y un tamaño de efecto grande. Estos son los genes que relacionamos con el desarrollo del cáncer, ya que hay tanto una diferencia estadística como un tamaño de efecto considerable.

	Gen	Expresion.en.Log2
1552767_a_at	HS6ST2	0.4169809
205625_s_at	CALB1	0.5543235
205626_s_at	CALB1	0.5314079
205825_at	PCSK1	-0.4817108
206032_at	DSC3	-0.3895832
207912_s_at	DAZ4///DAZ2///DAZ3///DAZ1	-0.5229974
230030_at	HS6ST2	0.4777648
235976_at	SLITRK6	-0.5821520
242601_at	HEPACAM2	-0.4766961

Otro resultado del programa de R es un data frame con los 10 genes que se sobreexpresan, los 10 que se subexpresan y sus valores de expresión en escala log2.

En algunos casos los nombres de los genes se repiten, esto se puede deber a las distintas variantes en la transcripción de un mismo gen, la misma razón por la cual hay más de 50,000 sondas cuando el ser humano cuenta con aproximadamente 20,000 genes.

	Genes.sobre	valor.exp.sobre	Genes.sub	valor.exp.sub
1	CALB1	0.5543235	SLITRK6	-0.5821520
2	CALB1	0.5314079	DAZ4///DAZ2///DAZ3///DAZ1	-0.5229974
3	HS6ST2	0.4777648	PCSK1	-0.4817108
4	HS6ST2	0.4169809	HEPACAM2	-0.4766961
5	SLCO1B3	0.3636220	DSC3	-0.3895832
6	SLC20A1	0.3620038	RETNLB	-0.3788246
7	HS6ST2	0.3535310	L1TD1	-0.3509375
8	DPP4	0.3447308	SLITRK6	-0.3495164
9	GPNMB	0.3392477	SLAIN1	-0.3479040
10	IBSP	0.3053662	HSPA2	-0.3461022

Al igual que en data frame pasado, hay algunos nombres repetidos, que se deben a la misma razón de las diferentes isoformas en las que un gen se puede expresar.

Después de recopilar los 10 genes que se sobreexpresan y subexpresan procedimos a dar un análisis más profundo a la función e importancia de dichos genes con las bases de datos brindadas por NCBI (National Center for Biotechnology Information) y para esto, seleccionamos 5 genes sobreexpresados para trabajar con ellos:

- CALB1
- HS6ST2
- SLCO 1B3
- SLC20A1
- DPP4

Empezamos por encontrar su localización en el ADN, después una vista general de su función,una breve explicación de la secuencia de identificación y de codón de inicio y de paro, el número de exones del gen, el porcentaje de A,T,C, y G, así como cuál es el tejido celular en el que más se expresa.

Genes	Localización y Longitud	Función	Secuencia de identificación,codón de inicio y de paro	Exones	% A,T,C,G	Tejido más expresado
1) CALB1	8q21.3, (90058608..90082879), 24272 bp	Amortigua la entrada de calcio tras la estimulación de los receptores de glutamato.	Isoforma: NM_001366795.1 atg, tag	11	A: 32.56% T: 33.18% C: 17.27% G: 16.99%	Riñón
2) HS6ST2	Xq26.2 / 335356bp	Este gen codifica un miembro de la familia de genes heparán sulfato (HS) sulfotransferasa, que cataliza la transferencia de sulfato a HS	Isoforma: XM_005262491.3 ,atg,taa,	11	A: 28.14% T: 30.63% G: 20.90% C: 20.33%	Riñón
3) SLCO1B3	12p12.2 / (20810705..20916911) / 106208 bp	La proteína codificada es un receptor transmembrana que media la captación independiente de sodio de compuestos endógenos y xenobióticos y desempeña un papel fundamental en el transporte de ácidos biliares y bilirrubina.	Isoforma: NM_019844.4 atg, taa	17	A: 29.88 % T: 34.22 % G: 18.35 % C: 17.53 %	Hígado
4) SLC20A1	2q14.1, Cromosoma 2, (112645938-112663825),17887 bp	La proteína codificada por este gen es un sodio-fosfato que absorbe fluido para su uso en funciones celulares como metabolismo y síntesis de ácido nucleico y lípidos. También es un receptor retroviral que ocasiona que las células humanas sean susceptibles a infecciones.	atg,tga Isoforma: XM_017004771.2	13	A:25.59% T:32.12% C:20.04% G:22.23%	Colon
5) DPP4	2q24.2 (GCF_000001405.39) / 81971 bp	Es importante en el metabolismo de la glucosa.Degrada GLP-1. Además, parece funcionar de supresor en el desarrollo de algunos tumores.	Solo produce ARN.	28	A:29.60% T:30.85% C:19.40% G:20.16%	Intestino delgado

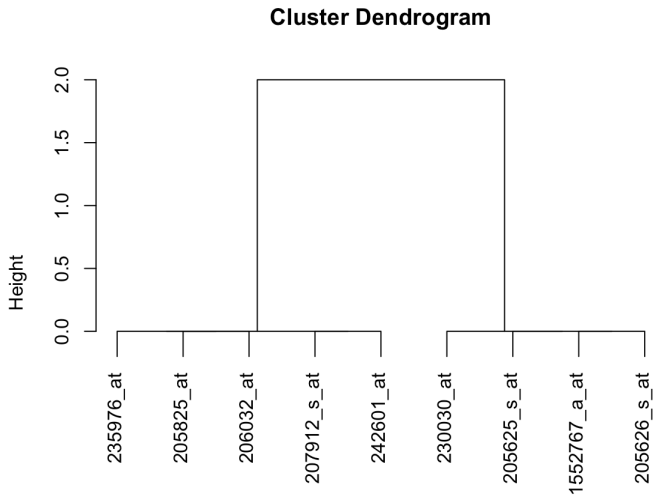
El siguiente paso fue buscar cuales son las mutaciones que podemos encontrar en estos genes. Utilizando las bases de datos de GnomeAD, Cosmic y Ensmbl ubicamos cuál es la cantidad de mutaciones encontradas al dia de hoy, cuales son los cambios estructurales en el gen y los cambios de aminoácidos dentro de este y si están en zonas codificantes. Conociendo esta información consideramos importante saber con qué frecuencia se pueden encontrar dichas mutaciones en diferentes poblaciones y a que edad se manifiestan.

Gen	Cambio de base (SNP)	Si están en zonas codificantes o no	Cambio de aminoácido resultante del SNP.	Cantidad de mutaciones	Edades	Población
CALB1	8-91072926-G-A	Exoma p.Asp201Asp	Variación de nucleótido	666	<30 años	Africana/Afroamericana
HS6ST2	X-132090867-C-G	Exoma p.Gly306Arg	AA mutation: p.A42T CDS mutation: c.124G>A	8401	30 - 35 años	Europea
SLCO 1B3	12-21030871-G-A	Exoma c.1135+1G>A	AA mutation: p.A42T CDS mutation: c.124G>A	3088	40 - 45 años	Europea
SLC20A1	c.276A>G (Substitution, position 276, A→G)	p.Ala351Thr	p.K92= (Substitution - coding silent)	316	<30 años	Europea
DPP4	2-162875726-C-T	Exoma, c.1298+7G>A	AA mutation: p.G51A CDS mutation: c.152G	1042	50-55 años	Africana

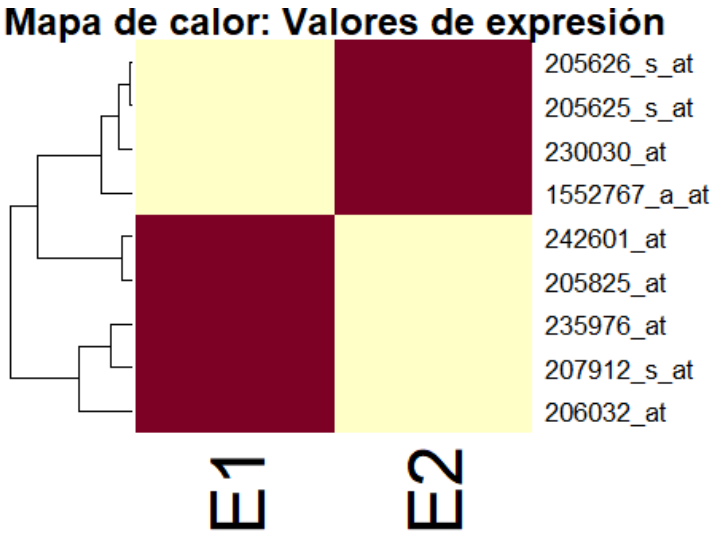
El paso siguiente fue hacer un concentrado de la información obtenida para poder apreciarla mejor y entender un poco mejor la relación entre lo investigado.

Gen	Localización	Función	Tejido más expresado	Cantidad de mutaciones	Edades	Población	Cambios estructurales y de aminoácido
CALB1	8q21.3	Amortigua la entrada de calcio tras la estimulación de los receptores de glutamato	Riñón	666	<30 años	Africana/Afroamericana	p.Asp201Asp
HS6ST2	Xq26.2	Este gen codifica un miembro de la familia de genes heparán sulfato (HS) sulfotransferasa, que cataliza la transferencia de sulfato a HS	Riñón	8401	30 - 35 años	Europea	AA mutation: p.P541L CDS mutation: c.1622C>T
SLCO 1B3	12p12.2	Media la captación independiente de sodio de compuestos endógenos y xenobióticos y desempeña un papel fundamental en el transporte de ácidos biliares y bilirrubina.	Hígado	3088	40 - 45 años	Europea	AA mutation: p.A42T CDS mutation: c.124G>A
SLC20A1	2q14.1	La proteína codificada por este gen es un sodio-fosfato que absorbe fluido para su uso en funciones celulares. También es un receptor retroviral..	Colón	316	<30 años	Europea	AA Mutation: p.K92 CDS Mutation: c.276A>G
DPP4	2q24.2	Es importante en el metabolismo de la glucosa.Degrada GLP-1. Además, parece funcionar de supresor en el desarrollo de algunos tumores.	Intestino delgado	1042	50-55 años	Africana	AA mutation: p.G51A CDS mutation: c.152G

El primer gráfico que se obtiene es un dendograma que en este caso agrupa jerárquicamente a los genes obtenidos por su similitud de comportamiento en su expresión. Mientras más abajo esté localizado el grupo, significa que los genes son más parecidos. La razón por la que solo hay dos grupos localizados hasta abajo se puede deber a que desde un inicio nuestros datos no presentaron diferencias tan grandes en la expresión, por lo que los grupos son muy similares.

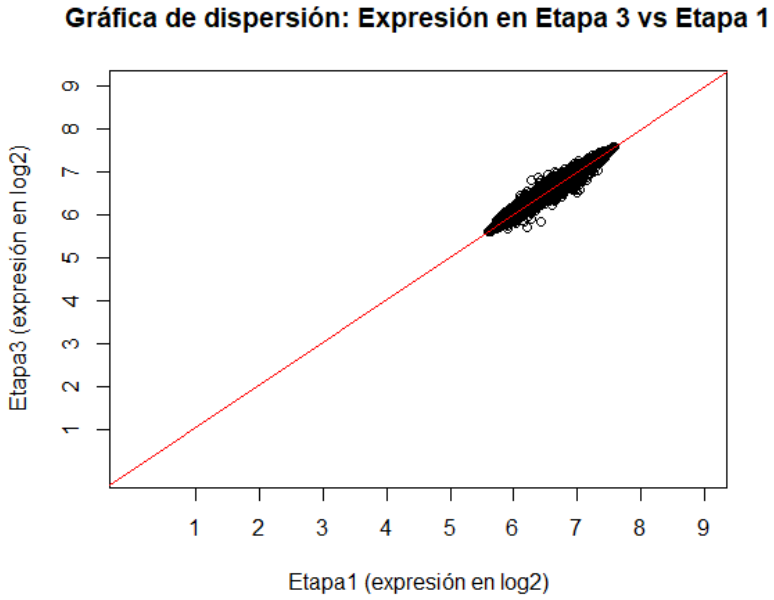


El mapa de calor que se obtuvo describe de una forma gráfica los valores de expresión de cada gen por grupo. Como las diferencias de expresión eran muy pequeñas, no se alcanza a percibir una gama amplia de colores, pero la diferencia abrupta entre el color crema y el vino confirma que hay un cambio en la expresión de ese gen entre las 2 etapas.

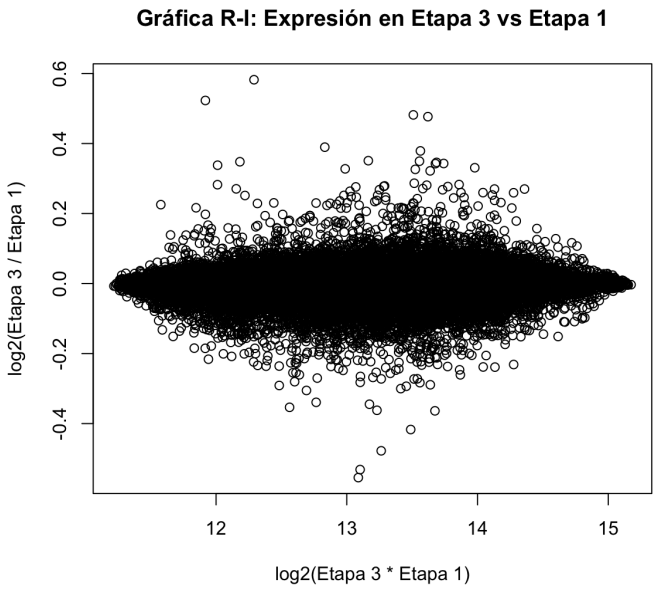


La gráfica de dispersión representa de forma más sencilla la tendencia que tienen los genes de sobre expresarse o sub expresarse. En el caso de esta gráfica, los genes que se encuentran cerca de la línea roja tienen una variación de expresión muy pequeña, mientras que los puntos que se encuentran más alejados son los que tienden a expresarse más en un grupo u otro. Si los puntos se alejan de la línea por debajo significa que se expresan más en la etapa 1, mientras que el

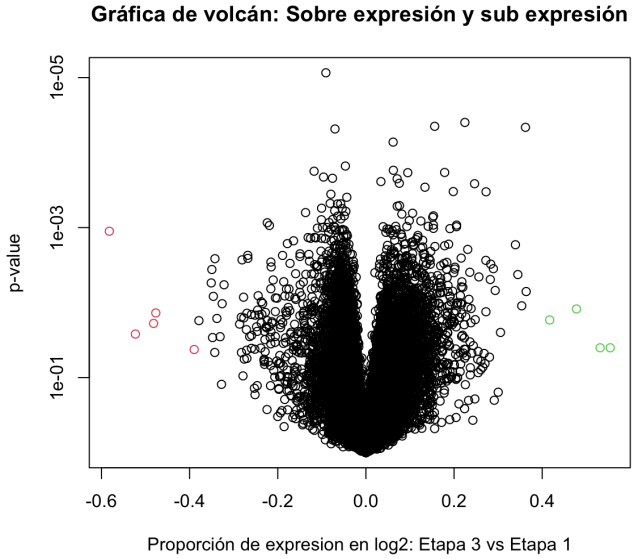
alejarse por encima representa una expresión mayor en la etapa 3. A pesar de la mínima variación, se pueden observar algunos genes que salen de la nube central, éstos tienen una probabilidad alta de ser parte de la lista final, ya que hay una diferencia considerable en la expresión.



La gráfica R-I es similar a la de dispersión, solo que en esta ocasión la nube de puntos está rotada para quedar horizontal. La gráfica tiene un significado similar a la pasada, los puntos que están alejados de la línea central son los genes que tienen una diferencia en su expresión pero la diferencia radica en la dirección en la que se alejan los puntos. Si los puntos se separan hacia arriba, se expresa más en la etapa 1 y si se separan hacia abajo, se expresa más en la etapa 3.



Por último la gráfica de volcán ayuda a visualizar los genes que se sobre expresan, se sub expresan y que entran en los requerimientos para categorizarlos como genes candidatos para el desarrollo del cáncer de colon. La nube de puntos se grafica con el p-value de cada gen en el eje "y" y la proporción de expresión en log2. Con esta visualización y delimitando los valores del p-value y la proporción de expresión se pueden colorear los puntos que entran en los rangos, donde los puntos en color rojo representan los genes de la etapa 3 que se subexpresan y los verdes los que se sobre expresan con respecto a la etapa 1.



Tecnológico de Monterrey

Análisis de resultados

En los valores de expresión finales se puede observar que no existe tanta diferencia entre los resultados obtenidos y el programa arrojó valores que consideramos demasiado bajos. Puede que estos resultados se deban a que no hay una diferencia estadística o un tamaño de efecto tan grande entre las diferentes etapas del cáncer. Debido a esto, se planeó realizar un incremento en el umbral propuesto, esto para poder realizar un análisis sin alterar nuestro procedimiento y desarrollo previo.

Al igual que con el p-value, nuestro umbral de tamaño de efecto para considerar un gen candidato se modificó. En principio el valor era de 1, que representaba una expresión del doble, pero debido a los resultados del análisis estadístico, este valor se tuvo que reducir a 0.38, lo que representa una expresión de 1.38.

En la parte del análisis biológico, pudimos identificar que los genes obtenidos tienen una amplia gama de repercusiones en condiciones y padecimientos relacionados con deficiencias de nutrientes, proteínas y en algunos casos de células estructurales. Encontramos varios artículos que indican que algunos genes tiene que ver, en su mayoría con leucemia por ejemplo, caída del cabello o calvicie y el más llamativo fue un artículo que explicaba su relación con el cancer de mama que fue el único caso en el que no se registraba una deficiencia tal cual, también se encontró relación directa con el cáncer colorrectal, en el caso del gen HS6ST2. Otro factor importante fue que al momento de investigar las características de los genes en las páginas como Cosmic o genomAD, pudimos identificar que el gen se expresa siempre en el colon, pero también en tejidos relacionados con el colon o cercanos a este como el riñón.

Así como este caso pudimos apreciar varios en el que existe una correlación entre órganos y como en efecto de cascada una pequeña variación genética o estructural puede tener repercusiones en un sistema complejo al grado de ocasionar enfermedades graves y no tan graves que pueden cambiar la vida de las personas.

También pudimos obtener un mejor entendimiento de la evolución del cáncer colorrectal en este caso, cuales son sus causas probables, sus síntomas, efectos secundarios y que tan grave puede llegar a ser.

Tecnológico de Monterrey

Reflexión individual

A modo de conclusión se puede decir que se logró el objetivo planteado al inicio del proyecto.Se logró elaborar de manera satisfactoria un análisis de los datasets seleccionados de NCBI y gracias al lenguaje de programación R, se elaboró un analisis computacional en la muestra de expresiones genéticas de dos grupos seleccionados. Así pudiendo identificar aquellos genes que resaltan más dentro de una muestra de pacientes con cáncer de colón.

Antes de este proyecto jamás había elaborado algún análisis computacional de esta manera. El buscar bases de datos y datasets fue una experiencia nueva para mí, aprendí mucho sobre el análisis y los parámetros que se deben considerar al momento de elaborar una investigación profunda sobre expresiones genéticas. Los criterios y parámetros que considero más importantes dentro de nuestro proyecto para tomar decisiones con el análisis estadístico que se llevó a cabo en R son: El grupo de etapas de los pacientes y el pvalue para filtrar nuestros genes finales. La etapa del paciente seleccionado es un factor que afecta mucho en los genes que tendremos como resultado al final del análisis, esto es porque si se escoge un paciente de etapa 1 y otro de etapa 2, no resaltan mucho las expresiones genéticas de los pacientes con cáncer. En cuanto al p value, podemos decir que de igual manera es un parámetro muy importante, siendo este el valor que se le da al umbral para filtrar cuánta diferencia hay entre nuestras expresiones genéticas de los pacientes en etapa 1 y 3. Este umbral es el valor que nos arroja los 10 genes más expresados en pacientes con cáncer dentro de nuestra muestra, por lo que el modificarlo aunque sea una décima, cambiaría todo nuestro resultado.

Como ya se mencionó anteriormente, estos dos parámetros, cambiarían la cantidad de resultados y la diferencia que hay entre ellos. El cambiar la etapa de nuestros paciente muestra, haría que los genes arrojados al final de nuestro programa no sean los correctos o los genes clave para identificar un paciente con cáncer a una edad temprana. Mientras que el pvalue si tiene un valor diferente, podría resultar en no mostrar genes diferenciales lo suficientemente importantes para el diagnóstico de esta enfermedad.

Este es un claro ejemplo de la ciencia de datos aplicada en la resolución de la salud publica, como se puede apreciar, sin necesidad de ser profesionales pudimos elaborar un

proyecto que considero es un gran paso para la detección temprana de cáncer de colon. El uso de la ciencia de datos en la salud va desde el beneficio de tener un sin fin de información al alcance de la mano de cualquier persona que quiera investigar un poco sobre enfermedades y genes. También otro beneficio son las bases de datos y los programas que pueden ser manipulados fácilmente con muchos lenguajes de programación. Es decir que simplemente se necesitan unos cuantos conocimientos para poder comenzar con tus propias investigaciones sobre las expresiones genéticas y la salud.

Tecnológico de Monterrey

Citas en Formato APA

1. Mayo Clinic. (s. f.). Colon cancer - Symptoms and causes. Recuperado 26 de abril de 2021, de https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669
2. Kuipers, E., Grady, W., Lieberman, D. *et al.* Colorectal cancer. *Nat Rev Dis Primers* 1, 15065 (2015). https://doi.org/10.1038/nrdp.2015.65
3. How to Analyze DNA Microarray Data. (2002, 28 febrero). HHMI BioInteractive. https://www.biointeractive.org/classroom-resources/how-analyze-dna-microarray-data
4. About GEO2R - GEO - NCBI. (s. f.). National Center For Biotechnology Information. Recuperado 26 de abril de 2021, de https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html
5. Yang, W., Ma, J., Zhou, W., Li, Z., Zhou, X., Cao, B., Zhang, Y., Liu, J., Yang, Z., Zhang, H., Zhao, Q., Hong, L., & Fan, D. (2018). Identification of hub genes and outcome in colon cancer based on bioinformatics analysis. *Cancer management and research*, 11, 323–338. https://doi.org/10.2147/CMAR.S173240
6. Smith, J. J., Deane, N. G., Wu, F., Merchant, N. B., Zhang, B., Jiang, A., Lu, P., Johnson, J. C., Schmidt, C., Bailey, C. E., Eschrich, S., Kis, C., Levy, S., Washington, M. K., Heslin, M. J., Coffey, R. J., Yeatman, T. J., Shyr, Y., & Beauchamp, R. D. (2010). Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*, 138(3), 958–968. https://doi.org/10.1053/j.gastro.2009.11.005
7. Hatabe, S., Kimura, H., Arao, T., Kato, H., Hayashi, H., Nagai, T., Matsumoto, K., DE Velasco, M., Fujita, Y., Yamanouchi, G., Fukushima, M., Yamada, Y., Ito, A., Okuno, K., & Nishio, K. (2013). Overexpression of heparan sulfate 6-O-sulfotransferase-2 in colorectal cancer. Molecular and clinical oncology, 1(5), 845–850. https://doi.org/10.3892/mco.2013.151