Campus Estado de México

Análisis de biología computacional

Evidencia 1: Reporte escrito proyecto integrador

Adrián Landaverde Nava

Naomi Padilla Mora

Nadia Paola Ferro Gallegos

Daniela Jace Olguín Montiel

Liciel Reyes Ávila

27 de abril del 2021

Evidencia 1

1. Resumen

El cáncer colorrectal es una enfermedad que se puede presentar en tres formas, la familiar, hereditaria y la esporádica. Es una de las principales causas de muerte, principalmente porque este tipo de cáncer posee una tasa de mortalidad alta. A lo largo de este reporte se podrá apreciar el trabajo de investigación realizado correspondiente al cáncer de colon. Se dará una breve explicación sobre las alteraciones genéticas que tiene. Para este reporte se utilizó el lenguaje de programación R, para el análisis de datos. Se explicará qué es lo que se hizo en el código de programación para poder llegar a los resultados. Se dividirá en dos partes el análisis de estos datos. Se proporcionarán las tablas con los genes que tengan su función general.

Palabras clave: Cáncer de colon, análisis de datos.

2. Introducción

El cáncer de colón es una enfermedad de alta gravedad debido a que es una de las principales causas de muerte, principalmente porque este tipo de cáncer posee una tasa de mortalidad alta. Sin embargo, si se diagnostica en una etapa temprana, esto podría provocar que la persona enferma pueda curarse completamente. El principal síntoma que comienza a denotar la presencia del cáncer de colón es el pólipo, que consiste en el crecimiento del revestimiento interno del colon. "El cáncer colorrectal es un crecimiento incontrolado de las células del colon y/o del recto…" (Monge, 2019) [1]. Aunque no todos los pólipos se convierten en cáncer porque la mayoría suelen ser benignos, en algunos casos suelen volverse malignos. Entre otros síntomas, también podemos encontrar la sensibilidad y dolor en la parte inferior del abdomen, sangre en las heces, diarrea, estreñimiento y pérdida de peso inusual.

El riesgo de padecer esta enfermedad está relacionado con factores como el tipo de alimento que se ingiere, sobre todo si no se tiene una dieta alta en fibra, baja en grasas y con un bajo consumo de carnes rojas. Pero también existen otros factores como:

- Edad: El grado de riesgo presente varía dependiendo del rango de edad. El riesgo alto se presenta en personas entre 65 y 75 años, esto se debe a que la mayoría de los casos de cáncer de colón se han localizado en personas en estas edades. El riesgo medio está presente entre personas de 50 y 65 años. Y finalmente el riesgo bajo está en las edades inferiores a 50.
- Historial médico: Las personas que han padecido de la presencia de pólipos en el colon, cáncer de mama, útero u ovarios y colitis ulcerosa (la cual es una enfermedad inflamatoria intestinal).
- Estilo de vida: La aparición del cáncer de colon también puede deberse a factores que dependen de cómo una persona cuida su salud como el tabaquismo, alcoholismo, obesidad y sedentarismo.

Cabe mencionar que esta enfermedad también está ligada a la genética. "Algunas enfermedades hereditarias también aumentan el riesgo de padecer cáncer de colon. Una de las

más comunes se llama síndrome de Lynch…" (Medline, 2021) [2]. Hay factores que, si una persona los presenta, puede tener una probabilidad alta de contraer cáncer de colon. Entre los factores de riesgo está el tener una historia familiar de neoplasias colorrectales, facilidad para desarrollar pólipos, alta frecuencia de la presencia de enfermedades intestinales, obesidad, etc. Es por eso que se puede manifestar de tres formas como lo son la familiar, hereditaria y la esporádica.

El colon también llamado intestino grueso es fundamental para el tracto digestivo. Este tiene como funciones principales la absorción de alimentos, agua y minerales, también sirve como almacenamiento de residuos de excremento. Este posee un elevado recambio celular, además está expuesto a diversos agentes de orden biológico, químico y físico. Son estos motivos los que llevan al desarrollo de diversas patologías. La que más destaca es el cáncer de colon.

El cáncer de colon o colorrectal es una enfermedad que se puede presentar en tres formas, la familiar, hereditaria y la esporádica. Esta última es la más común. En esta enfermedad influyen los factores ambientales y hereditarios para su desarrollo. Para que aparezca un tumor o exista la posibilidad, debe de hacer una acumulación de mutaciones en genes supresores, reparadores de ADN y oncogenes. Las vías supresoras y mutadoras tienen como característica las alteraciones genéticas por cambios morfológicos en la secuencia adenoma/carcinoma. "Las vías alternas originadas por mutaciones en los genes BRAF y KRAS se relacionan con la progresión de pólipo a carcinoma." [3]

El síndrome de Lynch

Es un síndrome hereditario con la variante HNPCC. Esta patología es autosómica, dominante y hereditaria. Este síndrome con el 3% da lugar al cáncer colateral. "Los tumores surgen por inactivación somática del gen que 6 previamente estaba mutado en la línea germinal, ya sea por la pérdida de heterocigosidad o LOH siglas en inglés de loss of heterozygosity o por mutaciones somáticas o hipermetilación de promotores." [3]

Hay dos tipos de HNPCC, en uno los tumores se ubican en el endometrio, páncreas, ovarios, etc., fuera del colon y en el segundo los tumores se encuentran dentro del colon. Aquellas personas que presentaron HNPCC desarrollan adenomas, que muchas veces se vuelven malignos en un corto tiempo comparándolo con las personas con FAP. Algunas de las características de los tumores de HNPCC, presencia de mutaciones β-catenina, k-ras y/o APC. Sus particularidades son la coexistencia de adenomas y la infiltración linfocitaria.

Alteraciones genéticas

Se pueden categorizar en 3 fundamentales, "1) genes supresores de tumores o TSG por sus siglas en inglés, como lo son APC, DCC, TP53, SMAD2, SMAD4 y p16INK4a); 2) protooncogenes, como lo es Kras, N-ras; 3) genes reparadores del ADN, como los genes MMR y MUTYH). " [3]

El modelo de Fearon-Vogelstein ilustra la progresión tumoral por medio de los denominados “múltiples pasos”. “El modelo se completó en 2002 con el descubrimiento de la MAP quinasa (56,57), lo cual afirma que el patrón de mutaciones del gen MUTYH deriva en una deficiencia de la función proteica, que se caracteriza por un exceso de transversiones G→T en secuencias GAA, que son susceptibles de provocar la aparición de codones de parada (58). ” [3]

3. Descripción de los sets de datos

Para este análisis, se usó la plataforma *National Center for Biotechnology Information* (NCBI) para la búsqueda de un dataset adecuado. Este dataset lleva por título: *Gene Expression Profiles in Stage II and III Colon Cancer. Application of a 128-gene signature* [4], y tiene el código GSE31595. El objetivo de esta investigación fue predecir el resultado en pacientes con cánceres colorrectales en etapas II y III con el fin de conocer el riesgo de recaída, genes expresados y sobreexpresados junto con sus funciones. Dicho estudio fue realizado con el fin de conocer más acerca del cáncer de colon para poder tener un mayor acercamiento a un tratamiento efectivo en los pacientes dependiendo del estado en el que se encuentren. Asimismo, se usó la única plataforma que se tenía: GPL570. Este dataset contiene 37 muestras de personas danesas de mujeres y hombres entre 51 a 91 años cuya diferencia más significativa era que algunos tenían cáncer de colon en etapa II y otros en etapa III.

Por lo tanto, para este análisis se hicieron 4 grupos de personas. El primero con muestras aleatorias de personas entre 51 a 71 años (Adultos Jóvenes) con cáncer de colon en etapa II (GSM784850, GSM784884, GSM784860, GSM784871, GSM784858). El segundo con muestras aleatorias de personas entre 72 a 92 años (Adultos Grandes) con cáncer de colon en etapa III (GSM784890, GSM782674, GSM784852, GSM784867, GSM784887). El tercero, con muestras aleatorias de personas entre 51 a 71 años (Adultos Jóvenes) con cáncer de colon en etapa III (GSM784875, GSM784878, GSM782671, GSM784856, GSM784865). Y el cuarto con muestras aleatorias de personas entre 72 a 91 años (Adultos Grandes) con cáncer de colon en etapa III (GSM784881, GSM784873, GSM784886, GSM784889, GSM784851).

A partir del análisis de las muestras de estos 4 grupos, se pueden encontrar 2 listas principales de genes. Una de ellas donde a partir del valor de significancia de los valores de expresión se pueden encontrar una cantidad considerable de genes característicos del cáncer de colon, misma que puede servir para el uso de biomarcadores para la identificación del cáncer de colon. Además, una segunda lista puede ser extraída de estos valores de expresión, donde la diferencia de los valores de expresión entre las personas con cáncer de Colon en etapa II y en etapa III muestra un tamaño del efecto considerablemente grande, y a partir de estos genes, se puede analizar la función de estos y así poder entender cómo y por qué evoluciona de tal forma el cáncer de colon entre la etapa II y etapa III para poder generar tratamientos que detengan el crecimiento de esta enfermedad o incluso se pueda revertir.

4. Desarrollo del código

Lectura, obtención y filtrado de los datos. Para hacer este análisis, se usó el IDE R Studio para realizar un Notebook en R con todo el procedimiento. Por lo tanto, primero se descargó el dataset GSE31595, para posteriormente obtener los valores de expresión de los genes, de los cuales se obtuvo de un total de 54,675 sondas. Posteriormente, se seleccionaron los valores de expresión sólo para las 20 muestras mencionadas anteriormente. Finalmente, se obtuvo el nombre de los genes a los que pertenecían las sondas, eliminando los valores que no estuvieran asociados a un gen y los valores repetidos. Con este paso, se refiere a eliminar aquellos valores de expresión de las cadenas de nucleótidos que están presentes en el ADN de las personas, pero debido al proceso de transcripción del ADN, estas cadenas de nucleótidos no codifican para un gen que sea expresado. Por lo tanto, al final de este pre-procesamiento se obtuvieron los valores de expresión para las 20 muestras de 22,189 genes.

Procesamiento estadístico de los datos. Una vez obtenidos los datos de los genes, se realizaron diferentes operaciones estadísticas para analizarlos de una manera más certera. Primero se normalizaron los datos a través de la media de cada una de las columnas en el data frame, donde con trim=0.02 se eliminó una fracción de los menores y los más grandes valores, permitiendo eliminar el posible ruido presente en el data frame, para finalizar la normalización, cada una de las columnas se dividió por los promedios previamente obtenidos y se les multiplicó por 100. Posteriormente, se obtuvieron las medias de los valores de expresión de los genes para cada grupo de muestras. Luego, se calcularon las proporciones (o ratios) de las medias de los valores de expresión. Para este análisis se calcularon 2 proporciones, una de adultos jóvenes y otra para adultos grandes, donde el denominador fueron las medias de los valores de expresión de los genes en etapa II, y el dividendo las medias de los valores de expresión de los genes en etapa III. O en otras palabras, se calculó la proporción de expresión de los genes en Etapa II contra los genes en Etapa III para los adultos jóvenes y para los adultos grandes. Sin embargo, estos datos presentan entre sí diferencias grandes, por lo que para observar a una mejor escala los resultados y entender los resultados, se obtuvo el logaritmo base 2 de los datos normalizados de los datos.

Prueba t-test. Para encontrar aquellos genes que realmente presentaran un valor de expresión significativamente diferente comparado con los demás, se realizó una prueba t-test para encontrar el p-value de los datos, mismos que estaban separados en 2 grupos (Adultos jóvenes con cáncer de colon en etapa II y Adultos Grandes con cáncer de colon en etapa III). Debido a que la hipótesis alternativa espera que los genes de expresión en Adultos jóvenes sean diferentes a los genes de expresión en Adultos grandes, la hipótesis nula indica que estos genes de expresión deben de ser iguales. Y dado que todas estas muestras son de cáncer cuyo principal diferenciador es la etapa en que se encuentran el cáncer, se espera que no haya muchos datos diferentes entre sí, por lo tanto se eligió un p-value no tan estricto de 0.05 para negar o rechazar a la hipótesis nula y a su vez, aceptando como verdadera la hipótesis alternativa. A partir de esto, se encontraron 783 genes en los adultos jóvenes y 936 genes para los adultos grandes cuyo p-value fue menor a 0.05. Finalmente, con el fin de acortar la lista de genes y que esta fuera más precisa, se seleccionaron aquellos genes que estuvieran en ambos grupos, mismos que fueron un total de 46 genes (ANKRD46, BZW2, CCDC43, CDK5R1, COPS5, COQ8A, CTSV, EIF3H, ETNK1, FAM107B, FOXD4, GABARAP, HOXD8, HTRA2, IL2RA, ITPKA,

KNOP1, LILRB5, LOC100508408/SNORD14B/RPS13. LONRF3, MCU, MKRN3, MTIF2, OTUD6B, RASSF10, RNF43, RPL14, RPL27, RPL30, RPL7, RPL8, SH3D21, SMCHD1, SNORA5B/TBRG4, SNORD54/RPS20, SPPL2A, STK17A, TCFL5, TMEM140, TOMM6/PRICKLE4, UBE2V2, UQCC1, UTP23, ZDHHC23, ZFAND1, ZNF7).

Agrupación de los genes. Para observar cómo se relacionaban estos 46 genes entre sí se realizó un dendrograma (figura 1), mismo que agrupa en conjuntos (o clústers) a los genes dependiendo de la cómo se relacionan los valores de expresión de los mismos. Además de esto se realizó un mapa de calor (figura 2) para observar cómo se expresan estos genes agrupados en los 4 grupos.

Correlación de la expresión de los genes. Para poder comparar y contrastar los valores de expresión de los genes en etapa II y en etapa III se hicieron varios gráficos. En primer lugar, se realizaron 2 gráficos de dispersión, uno para los adultos jóvenes (figura 3), y otro para los adultos grandes (figura 4). En estos gráficos se puede observar la tendencia que siguen estos genes al comparar sus valores de expresión, donde se puede ver que no presentan una gran variabilidad, y que son pocos los datos que se alejan del conjunto de datos. Además, también se pudo observar la correlación de estos datos mediante los gráficos R-I, que de igual forma se tiene uno para los adultos jóvenes (figura 5) y otro para los adultos grandes (figura 6), mismos en los que se puede observar que están altamente correlacionados y que son muy pocos los datos alejados de esta tendencia, cerca del 0.5, que después de obtener su exponente 2, se observa que estos datos más alejados se expresan aproximadamente 1.4 más veces en una etapa que en otra.

Selección de los datos por p-value y tamaño del efecto. Al observar el comportamiento de los genes con base en su p-value y el tamaño del efecto se pueden encontrar los genes que se expresan más en una etapa que en otra, y que además no están correlacionados con la otra. Para esto, se realizaron 2 gráficos de tipo volcán, uno para adultos jóvenes (figura 7) y otro para los adultos grandes (figura 8). Dado que las muestras no presentan mucha diferencia entre sí, se eligieron los genes que tuvieran un p-value menor a 0.05 y donde la proporción de los valores de expresión (o tamaño del efecto) fuera mayor o menor al logaritmo base 2 de 1.4 (0.485 aproximadamente). Los valores dentro de estos intervalos se dibujaron de diferentes colores, donde en color verde se muestran aquellos con tamaño del efecto mayor a 0.485 y los rojos con un tamaño del efecto menor a 0.485. Esto se refiere a que los genes en color verde se sobreexpresan en las muestras en etapa II y los rojos se reprimen en las muestras en etapa II, o viéndolo de otra forma, los genes en color verde son aquellos que se reprimen en las muestras en etapa III y en color rojo aquellos que se sobreexpresan en las muestras en etapa III. Con base en este filtro se obtuvieron 11 genes, mismos que se enlistan en la tabla 1.

Genes Sobreexpresados en etapa II	Genes Sobreexpresados en etapa III
PSPH	ADGRG7
LOC101929036 / PAH	UGT2B17
KRT23	CLCA1
MAP7D2	XIST
-	CNTN3
-	FLJ22763 / C3orf85
-	REG1A

Tabla 1: Genes sobreexpresados en etapa II y en etapa III. Genes obtenidos mediante un p-value de 0.05 y un tamaño del efecto de 0.4 en logaritmo base 2 (1.5 veces aproximadamente)

5. Resultados

Gráficos generados en R. A continuación, en las figuras 1 a 8 se muestran las gráficas generadas a partir del código de R que se describió en la sección anterior

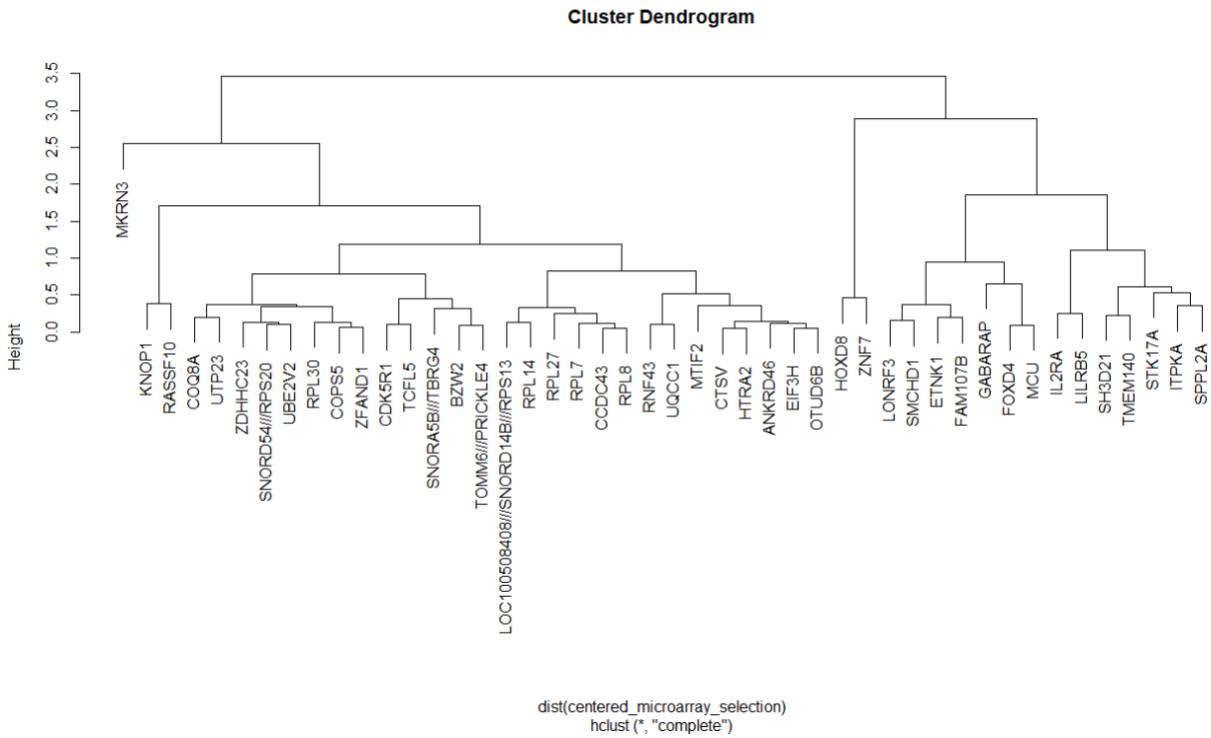


Figura 1: Dendograma de donde se muestra la agrupación de los 46 genes seleccionados con un p-value menor a 0.05 en ambos grupos. Estos valores están agrupados por la relación de los valores de expresión entre ellos.

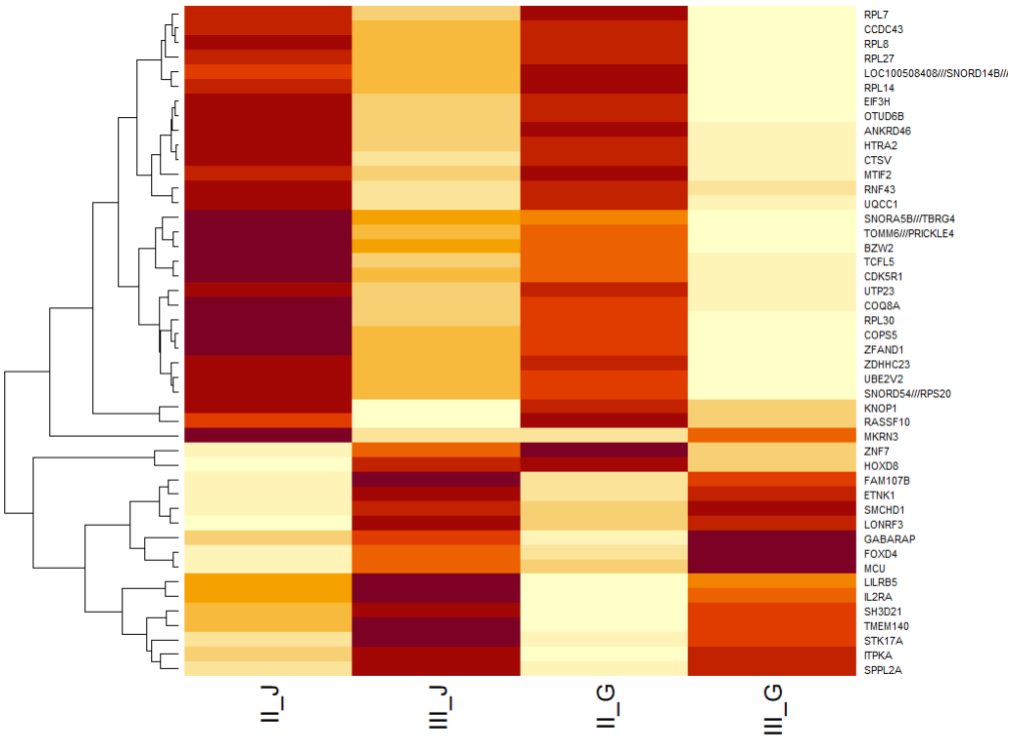


Figura 2: Mapa de calor de los valores de expresión de los 46 genes seleccionados a partir de la agrupación del dendograma. Donde II_J representa al grupo de adultos jóvenes con cáncer de colon en etapa II, III_J representa al grupo de adultos jóvenes con cáncer de colon en etapa III, II_G representa al grupo de adultos grandes con cáncer de colon en etapa II y III_G representa al grupo de adultos grandes con cáncer de colon en etapa III. Mientras más rojo es un color significa que ese gen se expresa más en ese grupo, y entre más amarillo sea un color, significa que ese gen se expresa en menor medida.

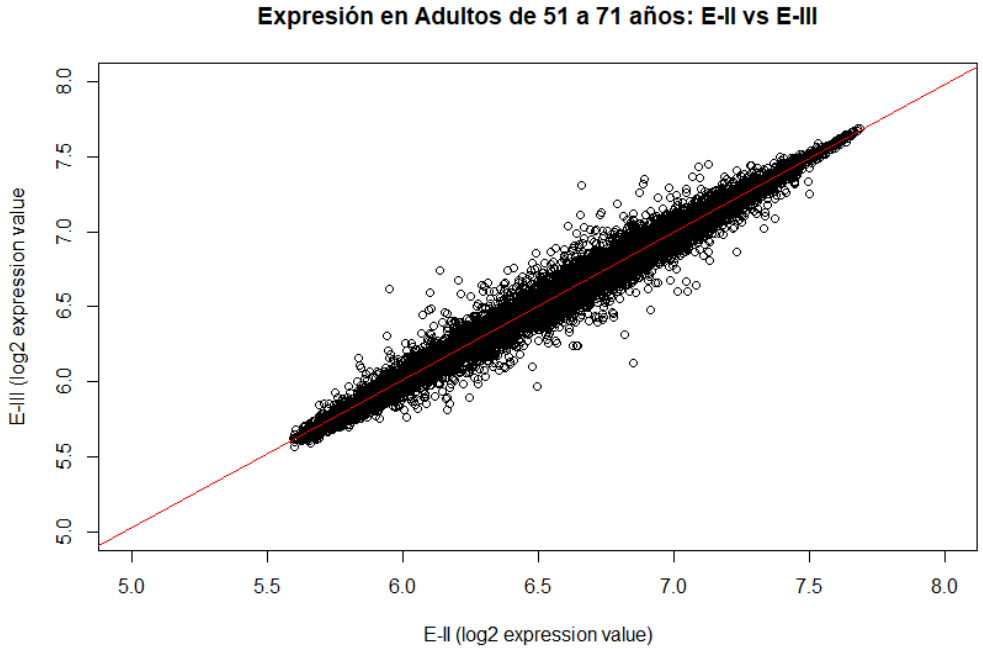


Figura 3: Diagrama de dispersión de los valores de expresión de los adultos de 51 a 71 años, con los valores de expresión de cáncer de colon en etapa II en el eje *x* y los valores de expresión de cáncer de colon en etapa III en el eje *y*

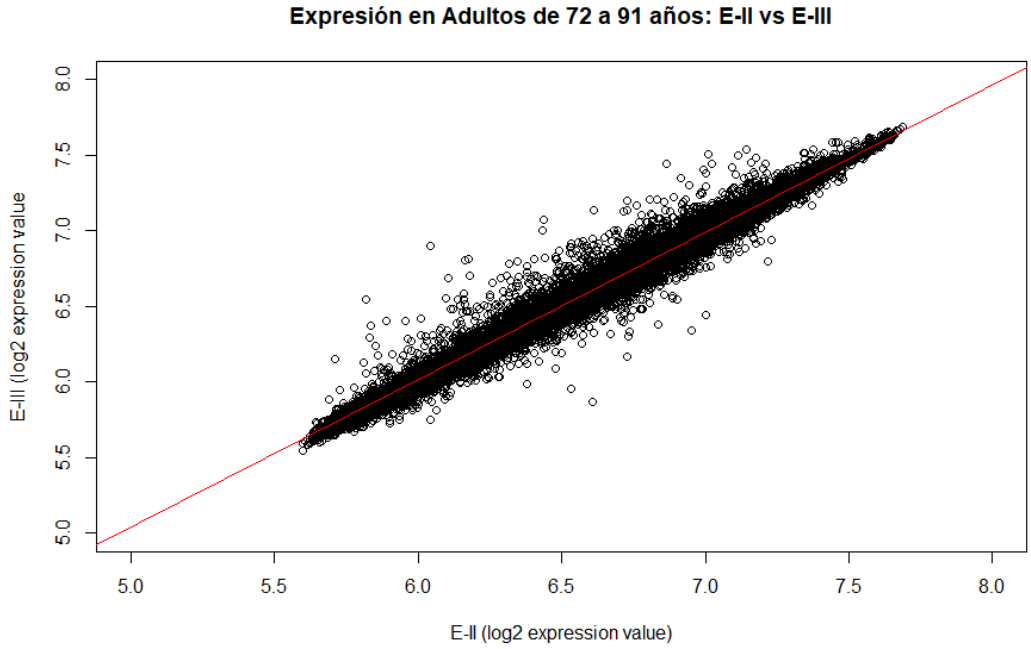


Figura 4: Diagrama de dispersión de los valores de expresión de los adultos de 72 a 91 años, con los valores de expresión de cáncer de colon en etapa II en el eje *x* y los valores de expresión de cáncer de colon en etapa III en el eje *y*

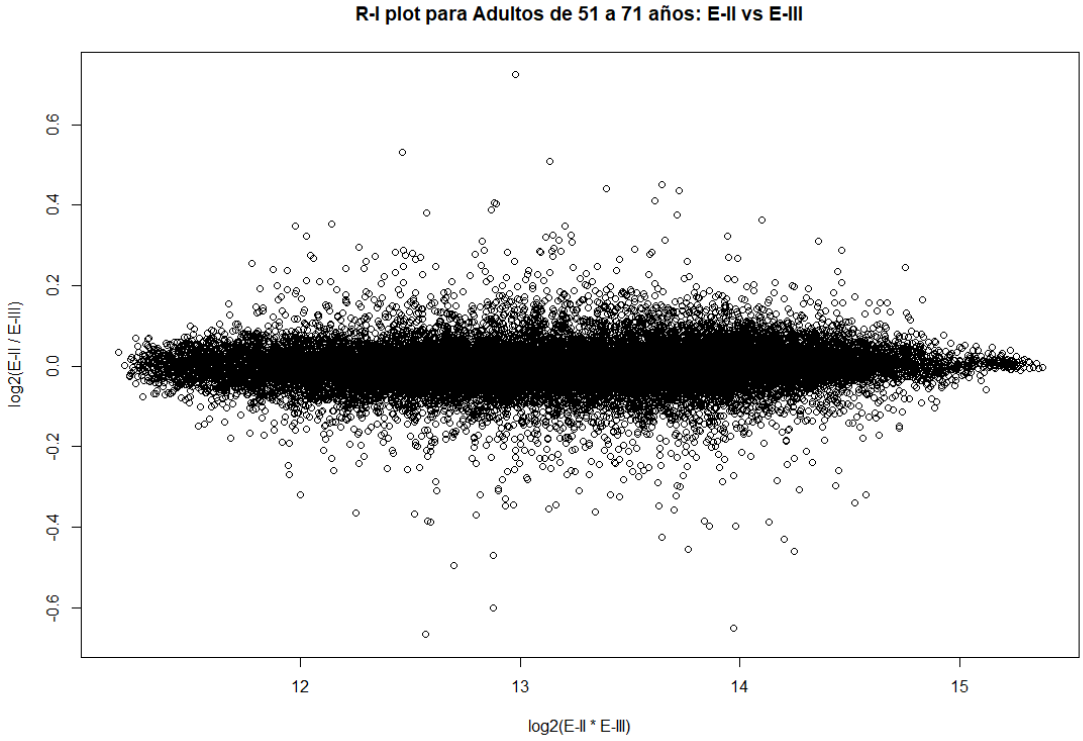


Figura 5: Diagrama de R-I de los valores de expresión de los adultos de 51 a 71 años, donde se muestra la correlación de los valores de expresión de cáncer de colon en etapa II y etapa III, entre más alejados del 0 en el eje *y*, hay una menor correlación entre ellos

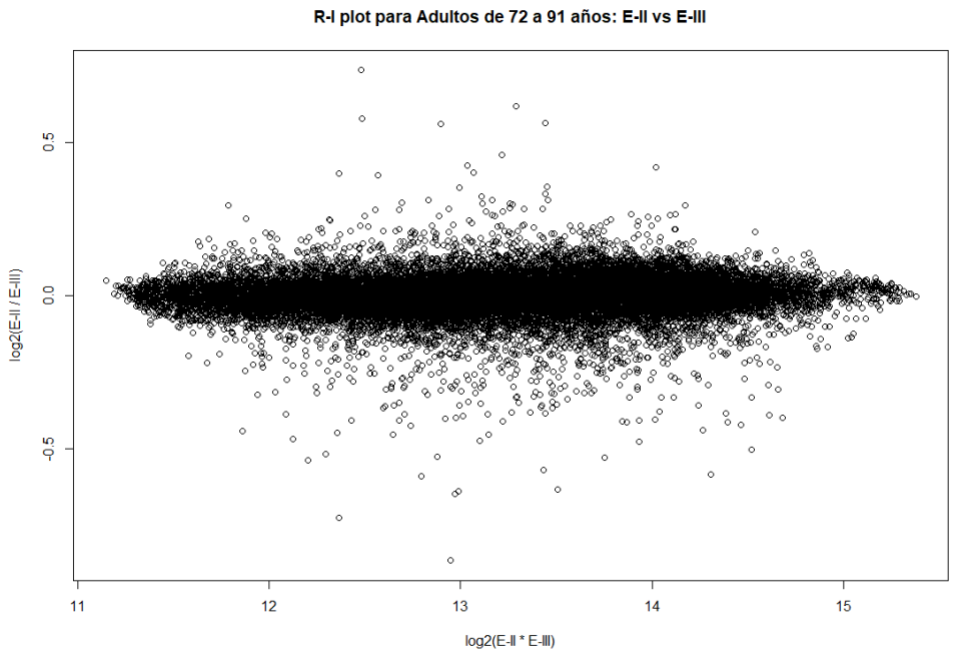


Figura 6: Diagrama de R-I de los valores de expresión de los adultos de 51 a 71 años, donde se muestra la correlación de los valores de expresión de cáncer de colon en etapa II y etapa III, entre más alejados del 0 en el eje *y*, hay una menor correlación entre ellos

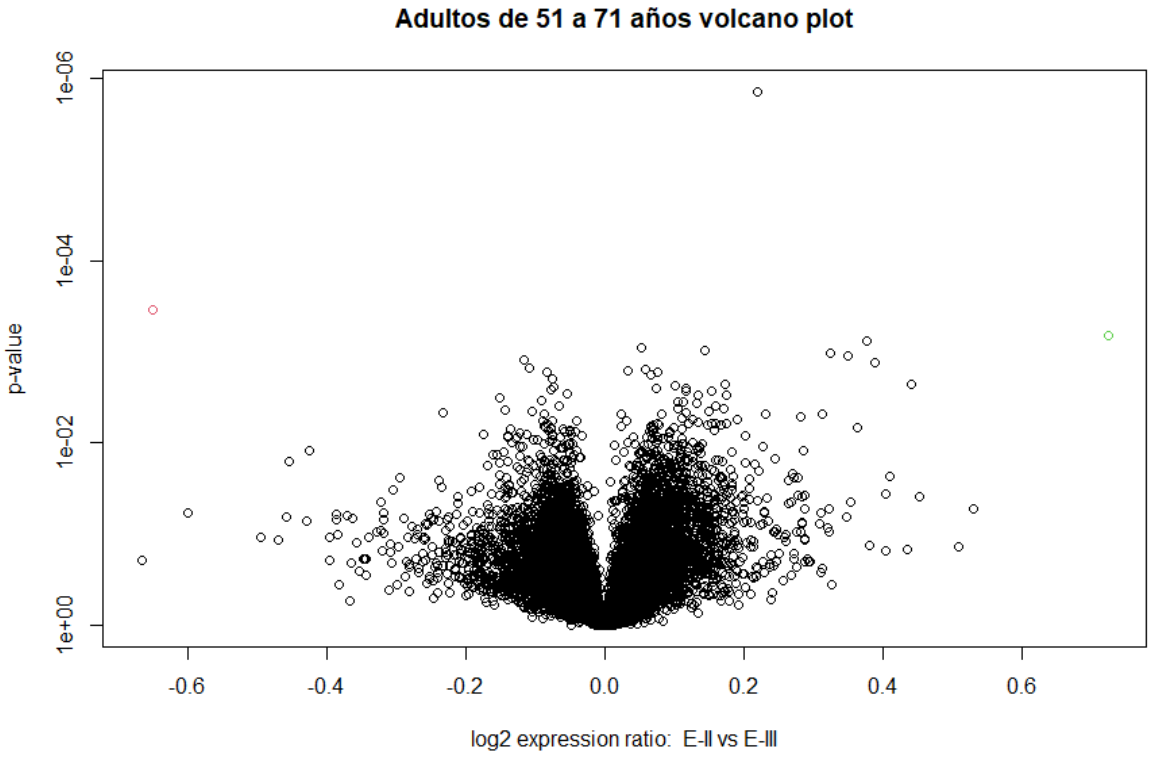


Figura 7: Diagrama de volcán de los valores de expresión de los adultos de 51 a 71 años, en color rojo se muestran los valores con un p-value menor a 0.05 y un tamaño del efecto menor al logaritmo base 2 de 1.4 (sobreexpresión de cáncer de colon en etapa III) y en color verde se muestran los valores con un p-value menor a 0.05 y un tamaño del efecto mayor al logaritmo base 2 de 1.4 (sobreexpresión de cáncer de colon en etapa II)

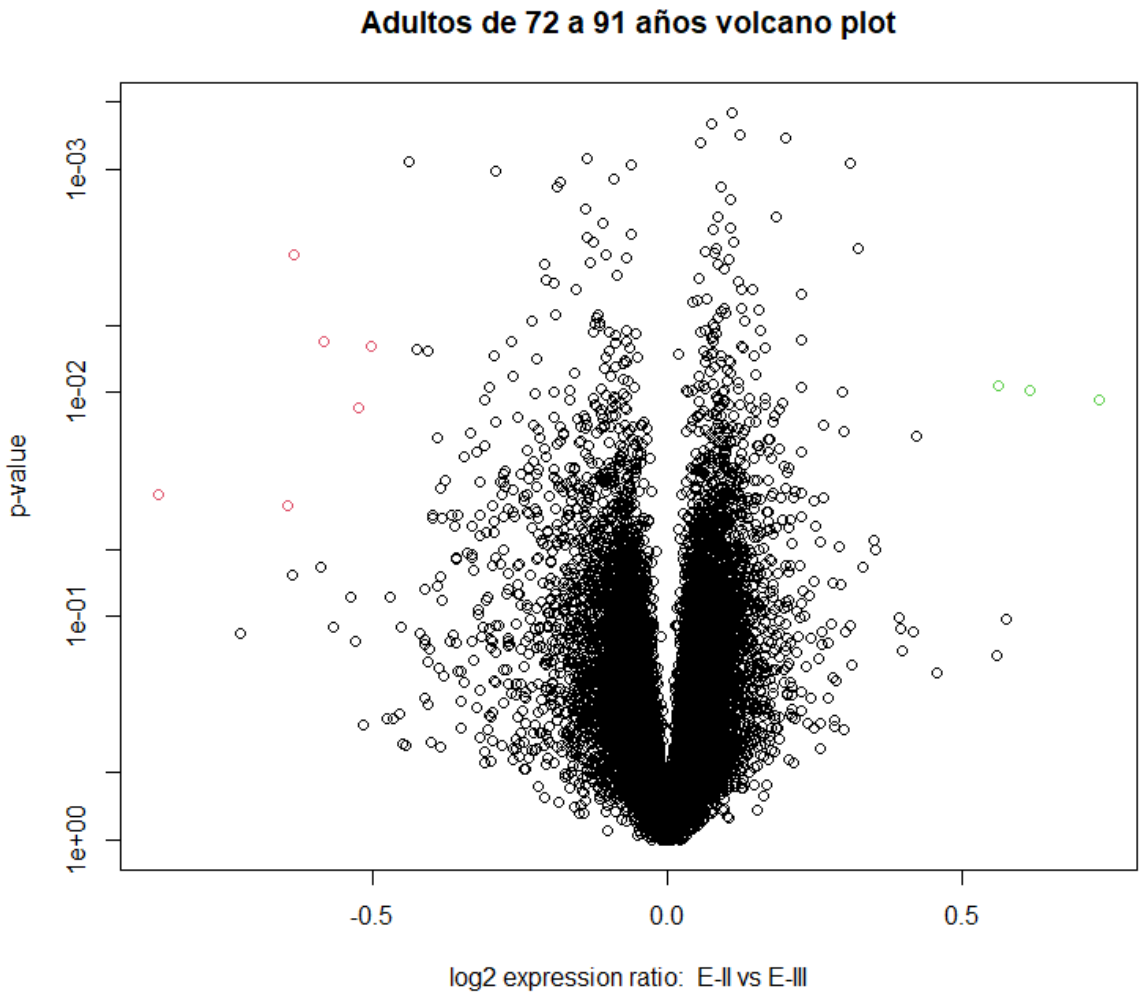


Figura 8: Diagrama de volcán de los valores de expresión de los adultos de 72 a 91 años, en color rojo se muestran los valores con un p-value menor a 0.05 y un tamaño del efecto menor al logaritmo base 2 de 1.4 (sobreexpresión de cáncer de colon en etapa III) y en color verde se muestran los valores con un p-value menor a 0.05 y un tamaño del efecto mayor al logaritmo base 2 de 1.4 (sobreexpresión de cáncer de colon en etapa II)

Con base en los genes obtenidos a partir del análisis estadístico, se pudieron identificar 2 grupos de genes cuyo reconocimiento sirve en distintas aplicaciones. Uno de ellos consiste en los 46 genes que se encuentran tanto en las muestras de adultos jóvenes y de adultos grandes y que tienen un p-value menor a 0.05 (ANKRD46, BZW2, CCDC43, CDK5R1, COPS5, COQ8A, CTSV, EIF3H, ETNK1, FAM107B, FOXD4, GABARAP, HOXD8, HTRA2, IL2RA, ITPKA, KNOP1, LILRB5, LOC100508408/SNORD14B/RPS13. LONRF3, MCU, MKRN3, MTIF2, OTUD6B, RASSF10, RNF43, RPL14, RPL27, RPL30, RPL7, RPL8, SH3D21, SMCHD1, SNORA5B/TBRG4, SNORD54/RPS20, SPPL2A, STK17A, TCFL5, TMEM140, TOMM6/PRICKLE4, UBE2V2, UQCC1, UTP23, ZDHHC23, ZFAND1, ZNF7). El segundo grupo, consiste en los 11 genes encontrados que tuvieron un p-value menor a 0.05 y además tuvieron un tamaño del efecto del logaritmo base 2 de 1.4 (PSPH, ADGRG7, LOC101929036 / PAH, UGT2B17, KRT23, CLCA1, MAP7D2, XIST, CNTN3, FLJ22763, REG1A).

Análisis de relación de los 46 genes (STRING). Los 46 genes mencionados anteriormente, con base en su p-value, se puede decir que se expresan con una gran diferencia entre la etapa II y la etapa III de cáncer de colon. Por lo tanto, estos están relacionados a los efectos del cáncer de colon en las personas, cada etapa con sus respectivos genes que sobreexpresa. Con base en esto, se puede deducir que estos genes pueden servir como biomarcadores de detección de cáncer de colon. Esto debido a que son genes que se diferencían con mucha frecuencia en estas etapas del cáncer. Y además, como se muestra en la figura 9, tras un análisis de la expresión de estos genes en la plataforma STRING [5], se encontró que algunos de estos interactúan entre sí, por lo que se comprueba que pueden ser vitales para el diagnóstico e identificación del cáncer de colon, pues incluso, muchos de estos genes están asociados a la formación de proteínas y ciclo celular en el colon. Y además, los genes que están relacionados entre sí nos pueden servir para entender los factores relacionados a la evolución del cáncer de colon.

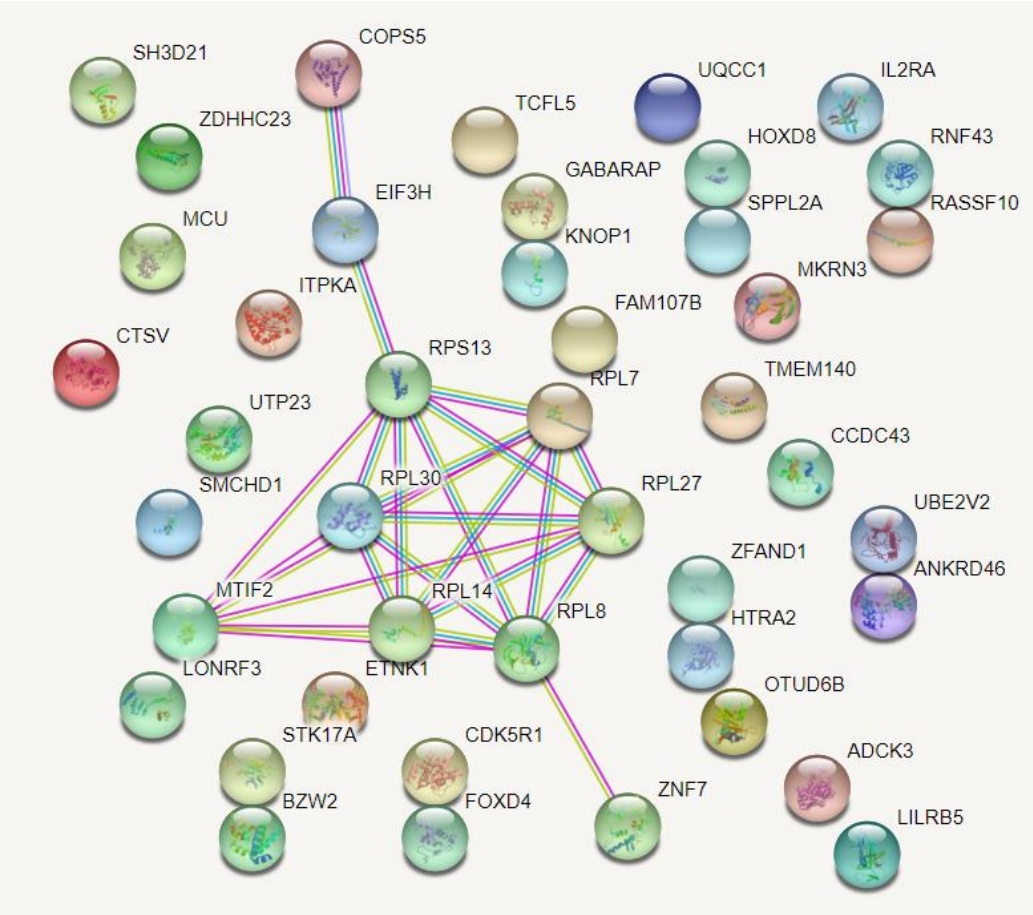


Figura 9: Interacción de los 46 genes seleccionados con base en la base de datos de STRING

Función celular de los 11 genes (KEGG). A partir de un análisis en *Kyoto Encyclopedia of Genes and Genomes* (KEGG), se encontraron los procesos y funciones celulares a los que estaban ligados el grupo de los 11 genes, estos se muestran en la tabla 2.

Nombre del Gen	Función general
MAP7D2	Función no disponible en KEGG. Nombre: MAP7 domain-containing protein 2. [6]
REG1A	Función no disponible en KEGG. Nombre: Regenerating family member 1 alpha. [7]
PSPH	Nombre: Phosphoserine phosphatase. Metabolic pathways. Biosynthesis of amino acids. Glycine, serine and threonine metabolism. [8]
PAH / LOC101929036	Nombre: Phenylalanine hydroxylase. Phenylalanine metabolism. Phenylalanine, tyrosine and tryptophan biosynthesis. Metabolic pathways. [9]
KRT23	Nombre: Keratin 23. Estrogen signaling pathway. Staphylococcus aureus infection. [10]
ADGRG7	Función no disponible en KEGG. Nombre: Adhesion G protein-coupled receptor G7. [11]
UGT2B17	Nombre: UDP glucuronosyltransferase family 2 member B17. Pentose and glucuronate interconversions. Steroid hormone biosynthesis. Metabolic pathways. Biosynthesis of cofactors. [12]
CLCA1	Nombre: Chloride channel accessory 1. Renin secretion. Pancreatic secretion. [13]
XIST	Función no disponible en KEGG. Nombre: XIST antisense RNA. [14]
CNTN3	Función no disponible en KEGG. Nombre: Contactin 3. [15]
FLJ22763 / C3ORF85	Función no disponible en KEGG. Nombre: Chromosome 3 open reading frame 85. [16]

Tabla 2: Función de los 11 genes con base en los datos de KEGG

Función celular de los 11 genes (NCBI y Reactome). Análisis de *National Center for Biotechnology Information* (NCBI) y *Reactome*, se encontraron los procesos y funciones celulares a los que estaban ligados el grupo de los 11 genes.

Nombre del Gen	Función general
MAP7D2	Microtubule-associated protein 7, Epithelial microtubule-associated protein of 115 kDa. [17]
REG1A	Encodes a protein that is secreted by the exocrine pancreas. [18]
PSPH	The protein encoded by this gene belongs to a subfamily of the phosphotransferases. This encoded enzyme is responsible for the third and last step in L-serine formation. [19]
PAH / LOC101929036	The encoded phenylalanine hydroxylase enzyme hydroxylates phenylalanine to tyrosine and is the rate-limiting step in phenylalanine catabolism. [20]
KRT23	The keratins are intermediate filament proteins responsible for the structural integrity of epithelial cells and are subdivided into cytokeratins and hair keratins. [21]
ADGRG7	Biased expression in duodenum (RPKM 18.5). [22]
UGT2B17	This gene encodes a member of the uridine diphospho glucuronosyltransferase protein family. [23]
CLCA1	The encoded protein is expressed as a precursor protein that is processed into two cell-surface-associated subunits, although the site at which the precursor is cleaved has not been precisely determined. [24]
XIST	X inactivation is an early developmental process in mammalian females that transcriptionally silences one of the pair of X chromosomes, thus providing dosage equivalence between males and females. [25]
CNTN3	Broad expression in brain (RPKM 2.7), prostate (RPKM 1.8). [26]
FLJ22763 / C3ORF85	Biased expression in duodenum (RPKM 15.1). [27]

Tabla 3: Función de los 11 genes con base en los datos de NCBI y Reactome

Análisis de relación de los 11 genes (STRING). A partir de la lista de 11 genes, de igual manera se usó la base de datos STRING para analizar la relación de estos genes entre sí y cómo influyen los unos con los otros, esta interacción se muestra en la figura 10. Y las funciones de estos genes obtenidas a partir de STRING [5] se ubican en la figura 11.

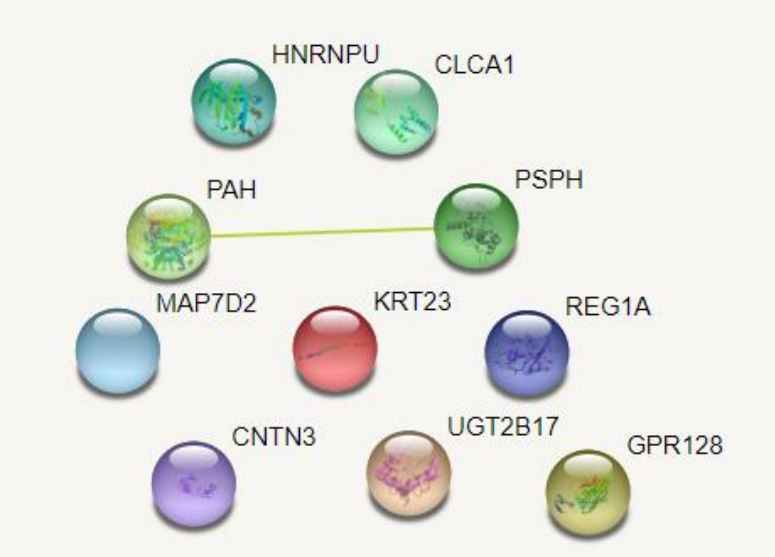


Figura 10: Interacción de los 11 genes seleccionados con base en la base de datos de STRING

KRT23	*Keratin, type I cytoskeletal 23; Keratins, type I (422 aa)*
UGT2B17	*UDP-glucuronosyltransferase 2B17; UDPGT is of major importance in the conjugation and subsequent elimination of potentially toxic xenobiotics and endogenous compounds. The major substrates of this isozyme are eugenol > 4-methylumbelliferone > dihydrotestosterone (DHT) > androstane-3-alpha,17-beta-diol (3-alpha-diol) > testosterone > androsterone (ADT); Minor histocompatibility antigens (530 aa)*
GPR128	*Adhesion G-protein coupled receptor G7; Orphan receptor (797 aa)*
PAH	*Phenylalanine-4-hydroxylase; Phenylalanine hydroxylase (452 aa)*
PSPH	*Phosphoserine phosphatase; Catalyzes the last step in the biosynthesis of serine from carbohydrates. The reaction mechanism proceeds via the formation of a phosphoryl-enzyme intermediates; HAD Asp-based non-protein phosphatases (225 aa)*
CLCA1	*Calcium-activated chloride channel regulator 1; May be involved in mediating calcium-activated chloride conductance. May play critical roles in goblet cell metaplasia, mucus hypersecretion, cystic fibrosis and AHR. May be involved in the regulation of mucus production and/or secretion by goblet cells. Involved in the regulation of tissue inflammation in the innate immune response. May play a role as a tumor suppressor. Induces MUC5AC; Chloride channel accessory (914 aa)*
HNRNPU	*Heterogeneous nuclear ribonucleoprotein U; DNA- and RNA-binding protein involved in several cellular processes such as nuclear chromatin organization, telomere-length regulation, transcription, mRNA alternative splicing and stability, Xist-mediated transcriptional silencing and mitotic cell progression. Plays a role in the regulation of interphase large-scale gene-rich chromatin organization through chromatin- associated RNAs (caRNAs) in a transcription-dependent manner, and thereby maintains genomic stability. Required for the localization of the long non-coding Xist RNA on the inacti [...] (825 aa)*
MAP7D2	*MAP7 domain containing 2 (773 aa)*
REG1A	*Lithostathine-1-alpha; Might act as an inhibitor of spontaneous calcium carbonate precipitation. May be associated with neuronal sprouting in brain, and with brain and pancreas regeneration; C-type lectin domain containing (166 aa)*
CNTN3	*Contactin-3; Contactins mediate cell surface interactions during nervous system development. Has some neurite outgrowth-promoting activity (By similarity); Belongs to the immunoglobulin superfamily. Contactin family (1028 aa)*

Figura 11: Funciones de los 11 genes seleccionados con base en la base de datos de STRING

Genes importantes para el análisis (Panther). Con base en la información recaba en STRING, se decidió ampliar la lista de genes para incluir a los 11 genes (PSPH, ADGRG7, LOC101929036 / PAH, UGT2B17, KRT23, CLCA1, MAP7D2, XIST, CNTN3, FLJ22763, REG1A) y agregar los genes que están relacionados entre sí de la figura 9 (COPS5, EIF3H, RPS13, RPL7, RPL27, RPL8, RPL14, RPL30, MTIF2, ZNF7). Una vez obtenidos estos genes, se usó la base de datos de Panther [28] para obtener su información acerca de la función molecular, procesos biológicos, componentes celulares, clase de proteína y las rutas de estos genes. En la figura 12 se encuentran los genes agrupados por sus funciones moleculares, como unión (8), actividad catalítica (4), regulador de funciones moleculares (1), actividad de transducción molecular (2), actividad molecular estructural (5), actividad regulatoria de traducción (1) y transporte (1). En la figura 13 se encuentran los genes agrupados por los procesos biológicos a los que están ligados como regulación biológica (3), procesos celulares

(9), procesos del sistema inmune (1) interacción entre organismos (1), procesos metabólicos (5), respuesta a estímulos (2) y señalización (1). En la figura 14 se encuentran los genes agrupados por componentes celulares como entidad anatómica celular (14), intracelular (11) y complejo de proteínas (7). Y finalmente, en la figura 15 se encuentran los genes agrupados por clase de proteína, como citoesqueleto (1), interconversión enzimática (1), traductoras (9), transmembranal (1) y de transporte (1).

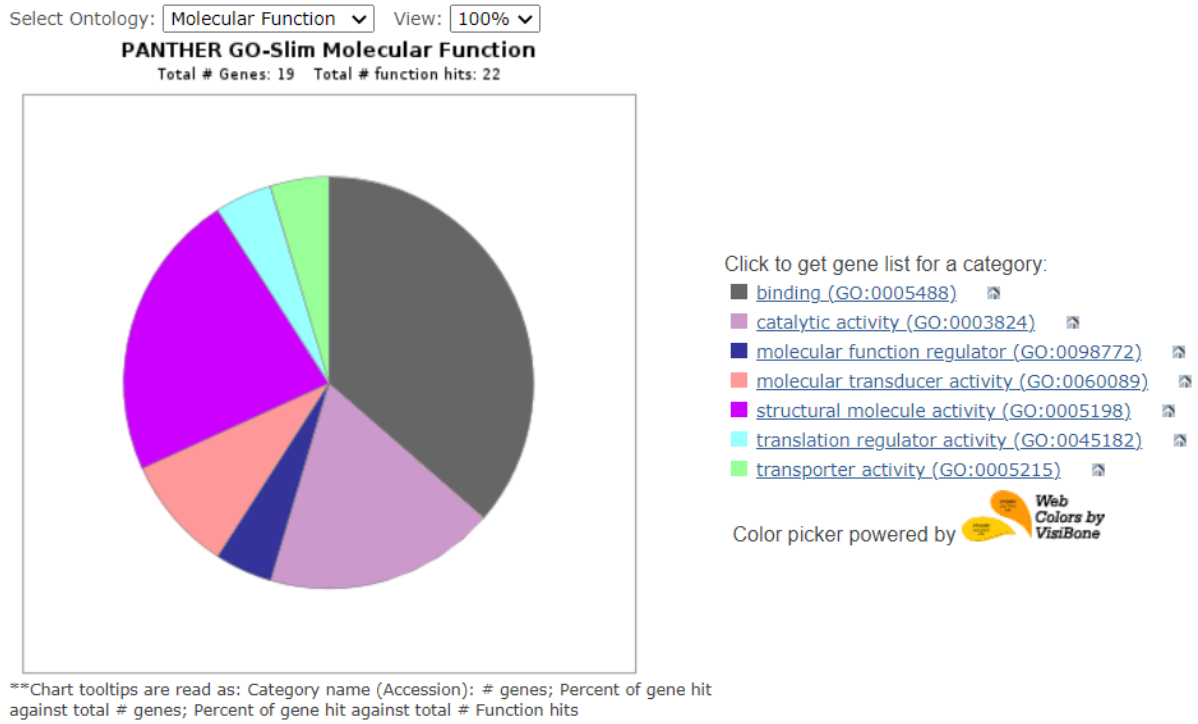


Figura 12: Gráfica de pastel que representa la distribución de las funciones moleculares de los genes obtenidos.

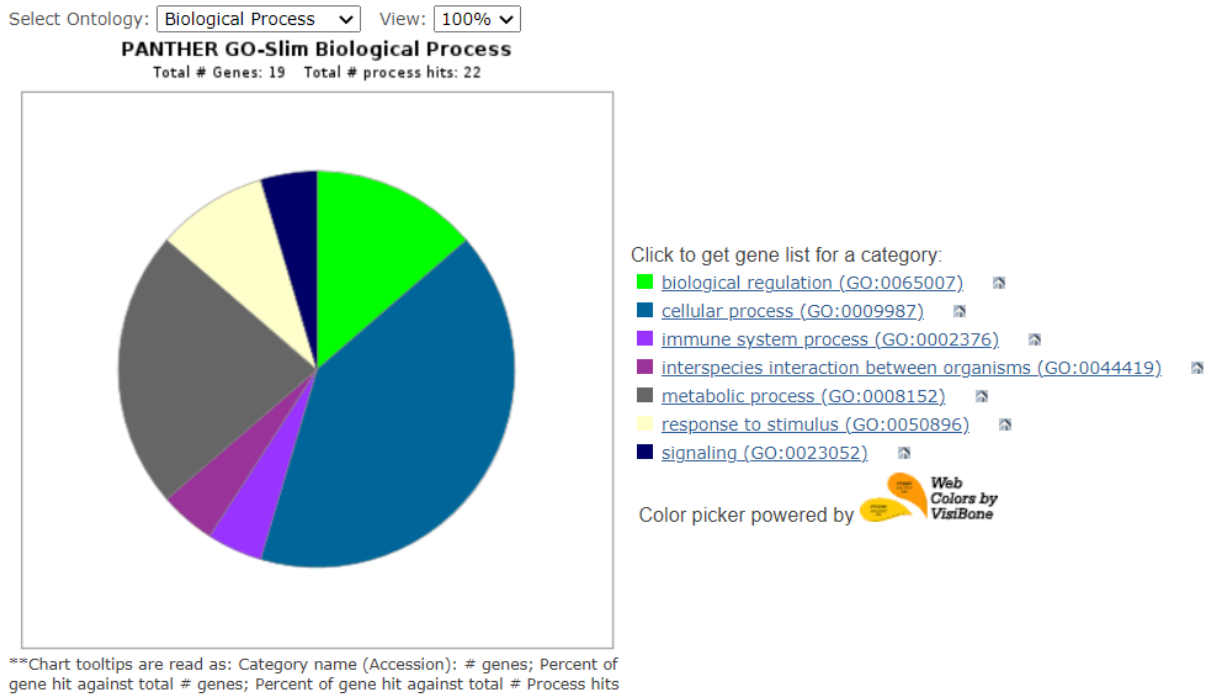


Figura 13: Gráfica de pastel que representa la distribución de los procesos biológicos de los genes.

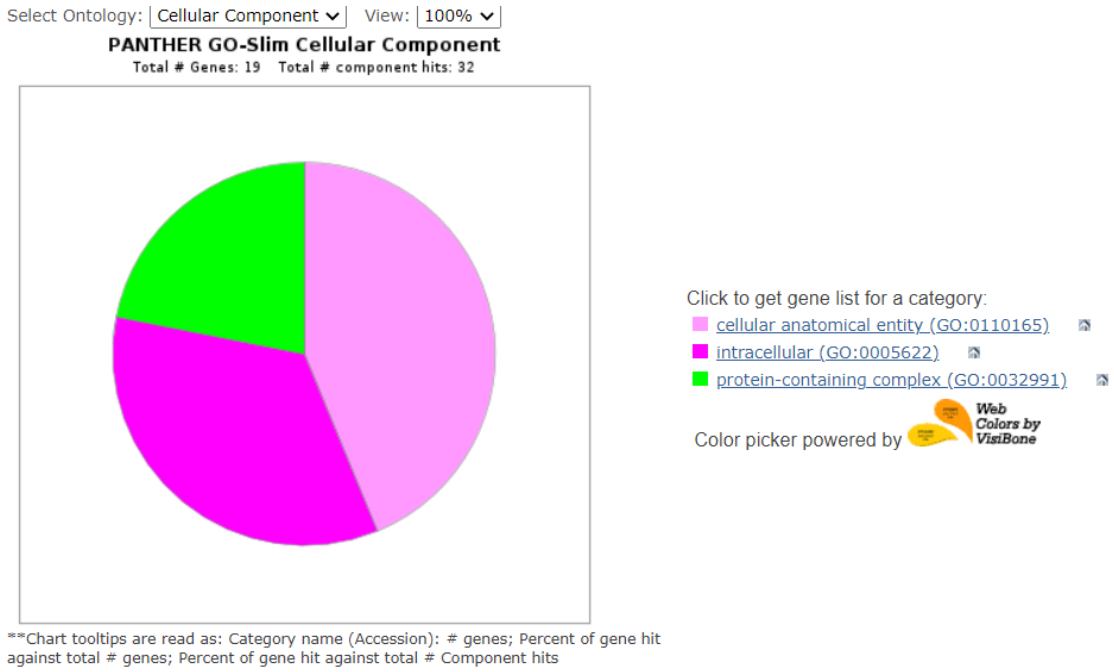


Figura 14: Gráfica de pastel que representa la distribución de los componentes celulares presentes en los genes.

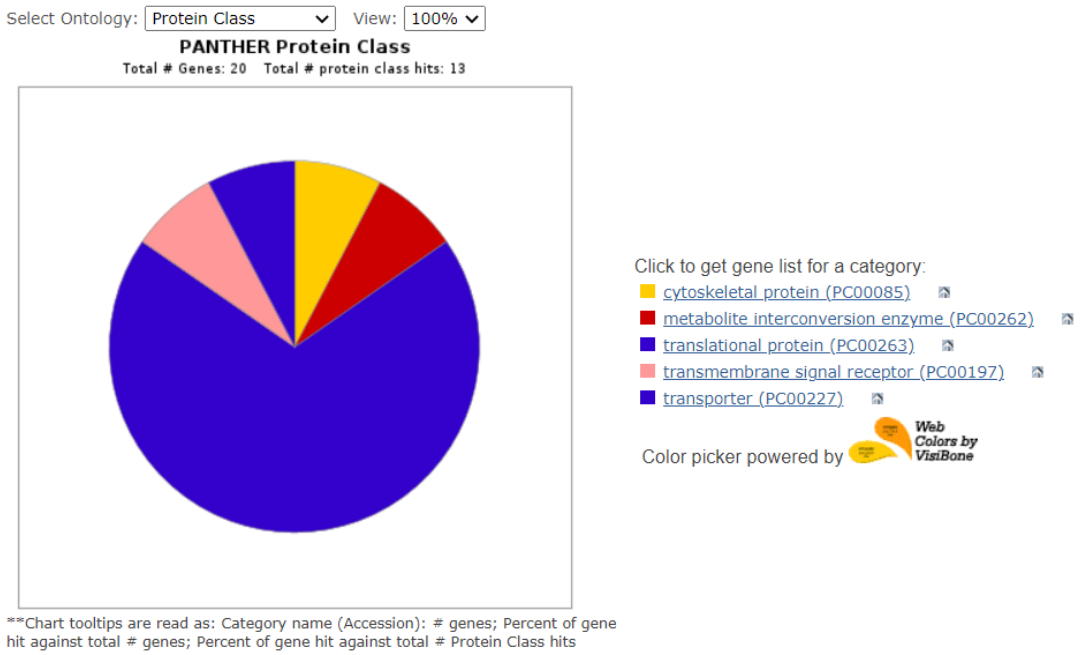


Figura 15: Gráfica de pastel que representa la distribución de los tipos de proteínas en los genes.

Órganos donde se expresan los genes principales para el estudio de la enfermedad (NCBI). Con base en toda la información recabada en KEGG, NCBI, STRING y PANTHER de los genes mencionados anteriormente, se seleccionaron 5 genes principales que debido a sus características, tienen un gran potencial para servir como genes clave o biomarcadores para

entender la evolución de la enfermedad en las personas: PSPH, PAH, REG1A, MAP7D2, UGT2B17. Y como últimos datos para la realización del análisis de los genes se necesita saber el o los tejidos donde tienen una mayor expresión estos genes. En las figuras 16 a 20, extraídas de la plataforma NCBI, se muestran los tejidos donde más se expresan estos genes.

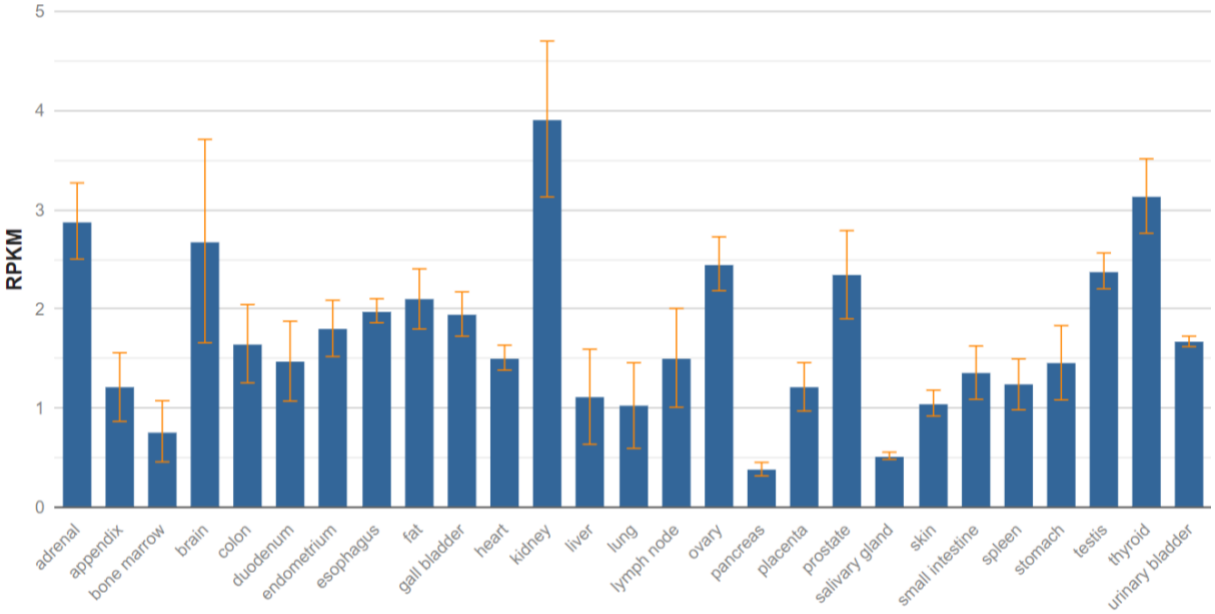


Figura 16: Gráfica que representa el nivel de expresión del gen PSPH en diversos tejidos [19].

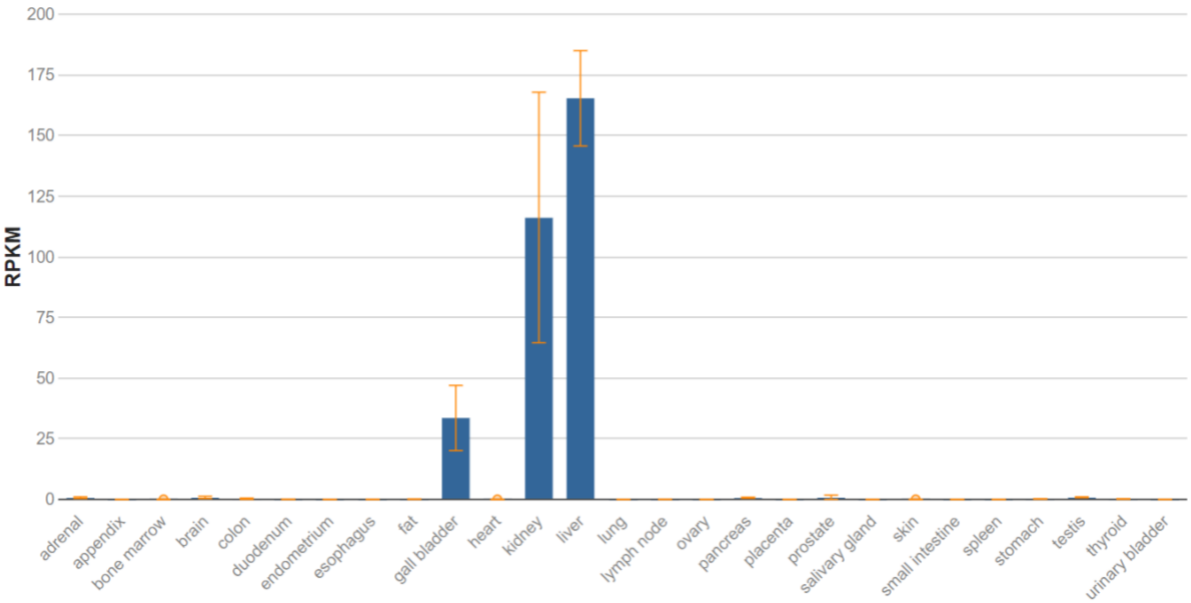


Figura 17: Gráfica que representa el nivel de expresión del gen PAH en diversos tejidos [20].

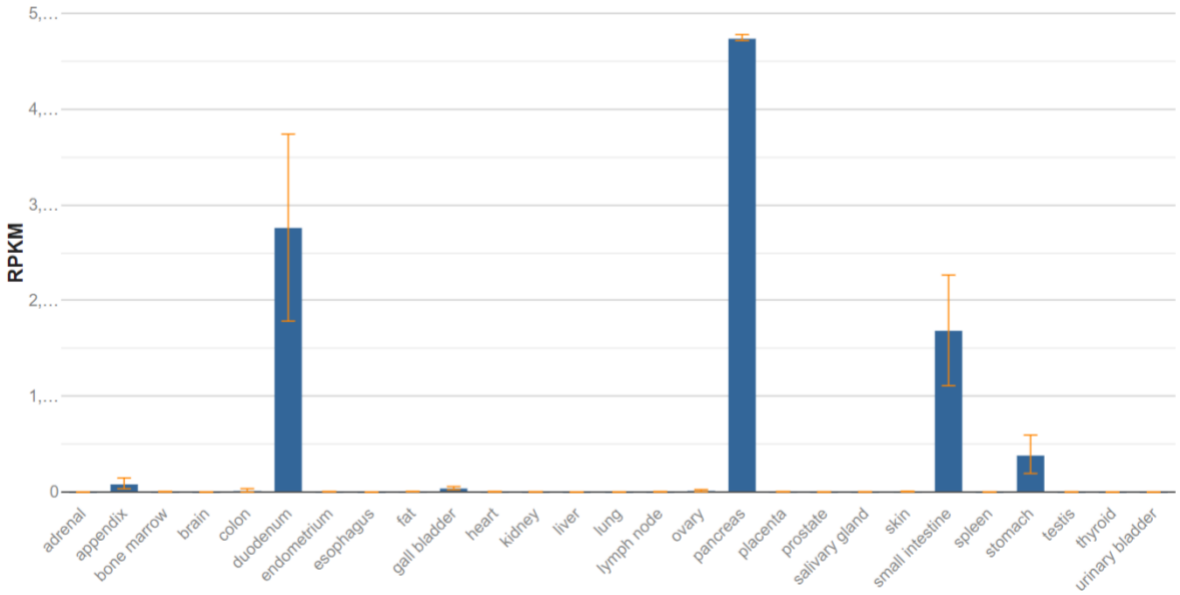


Figura 18: Gráfica que representa el nivel de expresión del gen REG1A en diversos tejidos [18].

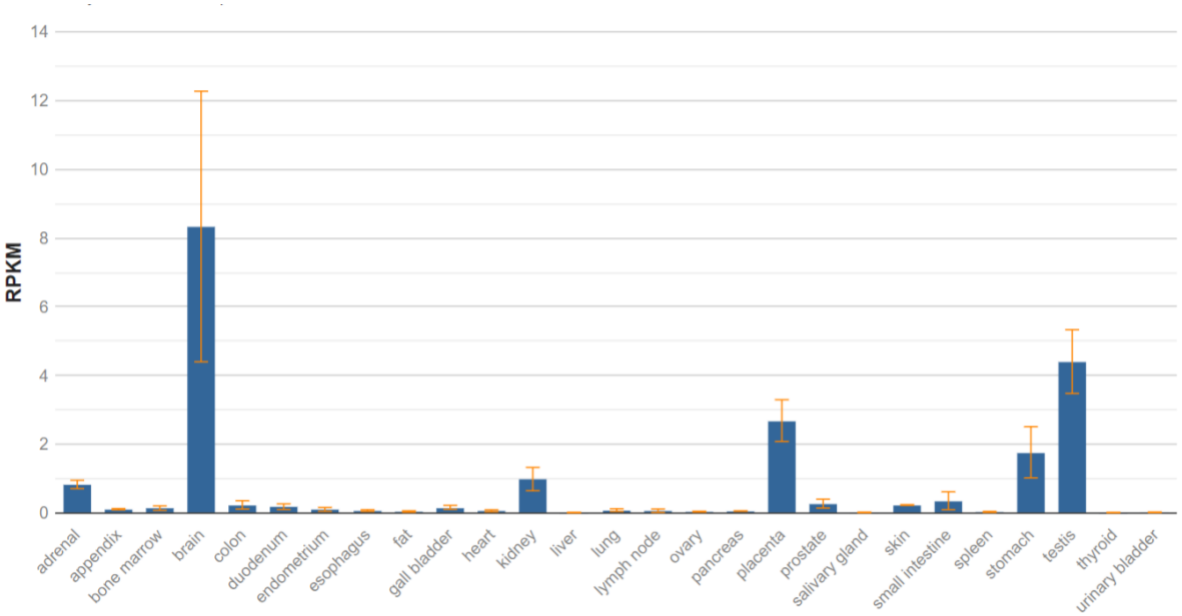


Figura 19: Gráfica que representa el nivel de expresión del gen MAP7D2 en diversos tejidos [29].

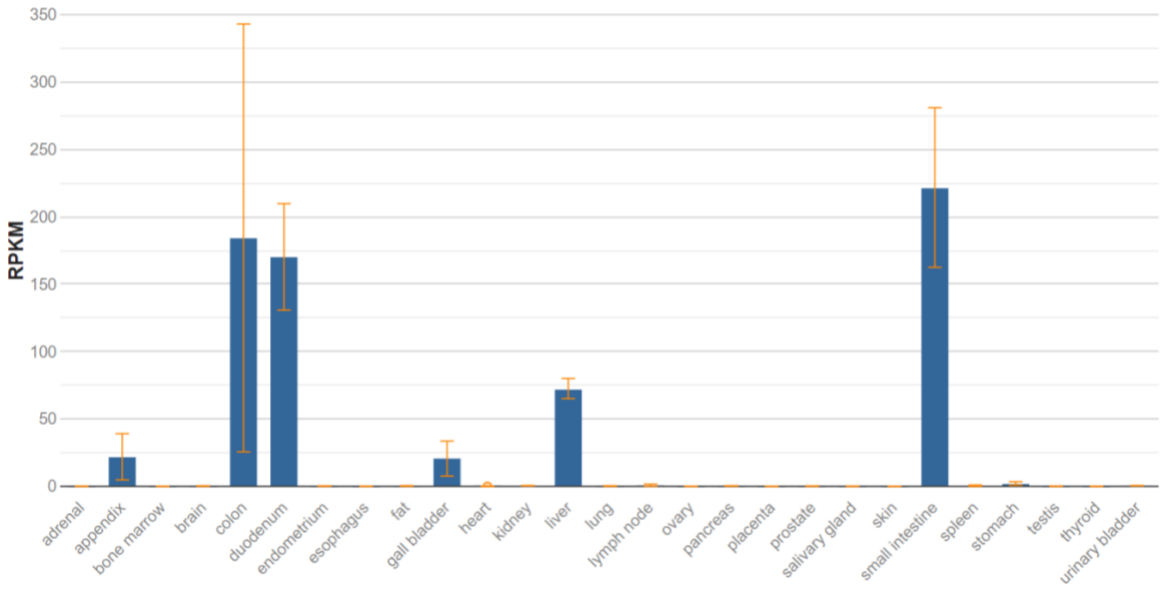
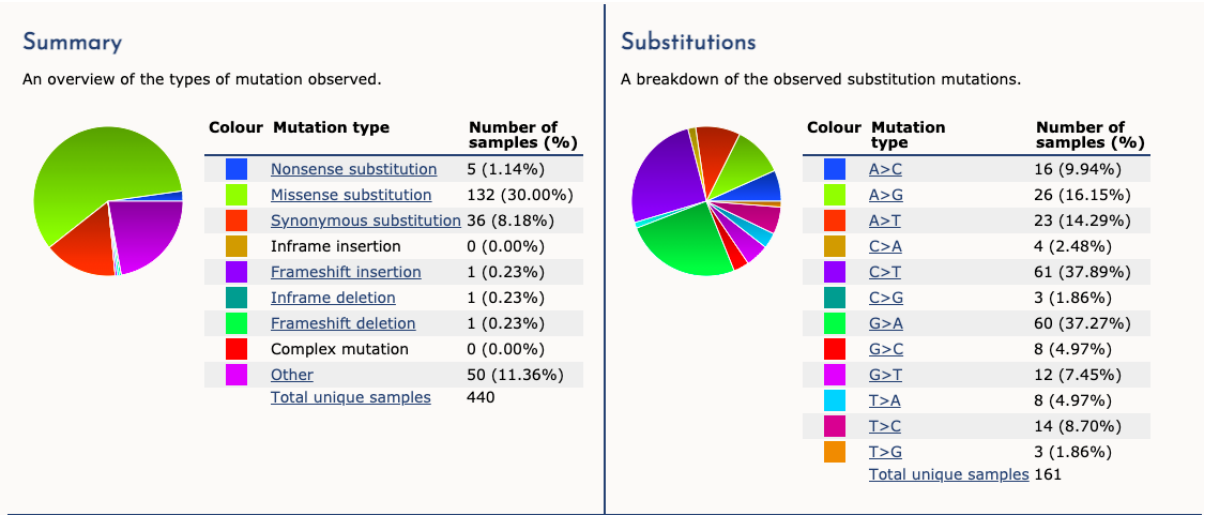


Figura 20: Gráfica que representa el nivel de expresión del gen UGT2B17 en diversos tejidos [23].

Mutaciones de los genes considerados como más importantes (COSMIC). A partir de esta lista de 5 genes considerados como los más importantes, se analizaron las mutaciones más comunes de estos genes usando la base de datos de COSMIC. En las figuras 21 a 25 se observan las mutaciones que pueden tener estos genes.

PSPH



Summary

An overview of the types of mutation observed.

Colour	Mutation type	Number of samples (%)
	Nonsense substitution	5 (1.14%)
	Missense substitution	132 (30.00%)
	Synonymous substitution	36 (8.18%)
	Inframe insertion	0 (0.00%)
	Frameshift insertion	1 (0.23%)
	Inframe deletion	1 (0.23%)
	Frameshift deletion	1 (0.23%)
	Complex mutation	0 (0.00%)
	Other	50 (11.36%)
	Total unique samples	440

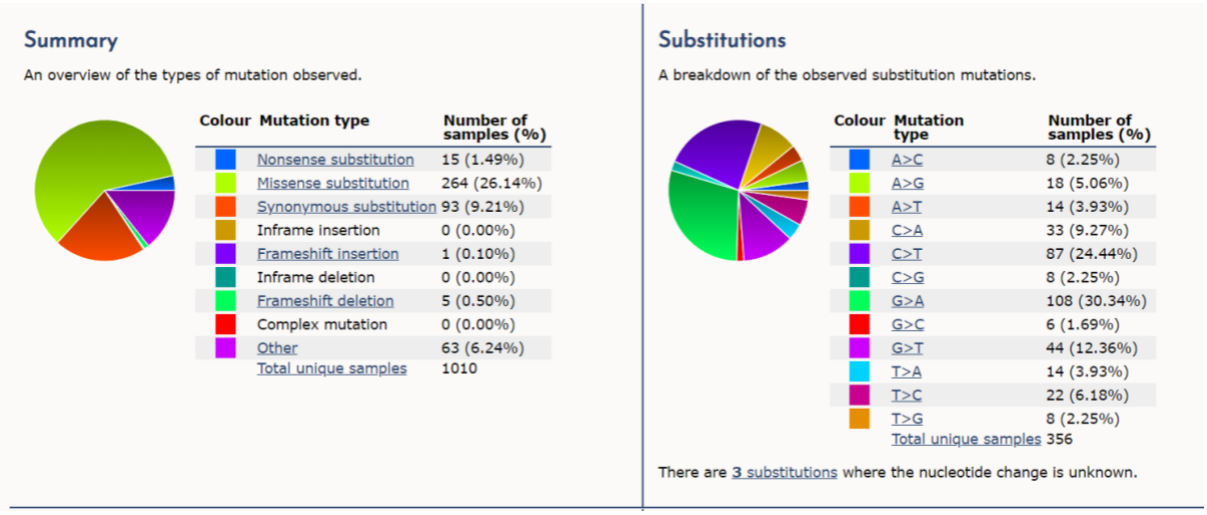
Substitutions

A breakdown of the observed substitution mutations.

Colour	Mutation type	Number of samples (%)
	A>C	16 (9.94%)
	A>G	26 (16.15%)
	A>T	23 (14.29%)
	C>A	4 (2.48%)
	C>T	61 (37.89%)
	C>G	3 (1.86%)
	G>A	60 (37.27%)
	G>C	8 (4.97%)
	G>T	12 (7.45%)
	T>A	8 (4.97%)
	T>C	14 (8.70%)
	T>G	3 (1.86%)
	Total unique samples	161

Figura 21: Gráficos de pastel que muestran la distribución de las mutaciones según su tipo, presentes en el gen PSPH [30].

LOC101929036 / PAH



Summary

An overview of the types of mutation observed.

Colour	Mutation type	Number of samples (%)
	Nonsense substitution	15 (1.49%)
	Missense substitution	264 (26.14%)
	Synonymous substitution	93 (9.21%)
	Inframe insertion	0 (0.00%)
	Frameshift insertion	1 (0.10%)
	Inframe deletion	0 (0.00%)
	Frameshift deletion	5 (0.50%)
	Complex mutation	0 (0.00%)
	Other	63 (6.24%)
	Total unique samples	1010

Substitutions

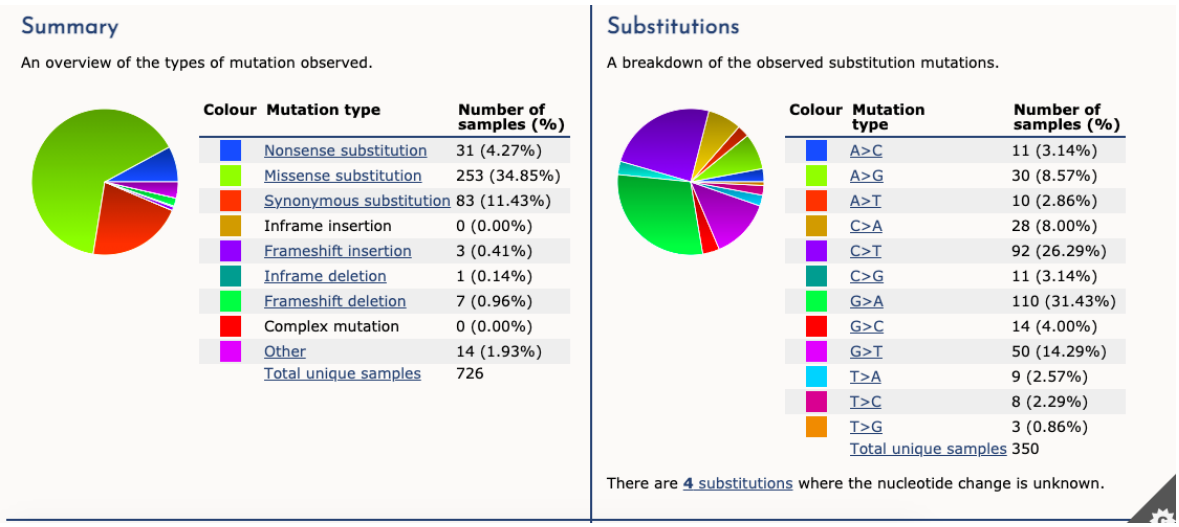
A breakdown of the observed substitution mutations.

Colour	Mutation type	Number of samples (%)
	A>C	8 (2.25%)
	A>G	18 (5.06%)
	A>T	14 (3.93%)
	C>A	33 (9.27%)
	C>T	87 (24.44%)
	C>G	8 (2.25%)
	G>A	108 (30.34%)
	G>C	6 (1.69%)
	G>T	44 (12.36%)
	T>A	14 (3.93%)
	T>C	22 (6.18%)
	T>G	8 (2.25%)
	Total unique samples	356

There are 3 substitutions where the nucleotide change is unknown.

Figura 22: Gráficos de pastel que muestran la distribución de las mutaciones según su tipo, presentes en el gen PAH[31].

MAP7D2



Summary

An overview of the types of mutation observed.

Colour	Mutation type	Number of samples (%)
	Nonsense substitution	31 (4.27%)
	Missense substitution	253 (34.85%)
	Synonymous substitution	83 (11.43%)
	Inframe insertion	0 (0.00%)
	Frameshift insertion	3 (0.41%)
	Inframe deletion	1 (0.14%)
	Frameshift deletion	7 (0.96%)
	Complex mutation	0 (0.00%)
	Other	14 (1.93%)
	Total unique samples	726

Substitutions

A breakdown of the observed substitution mutations.

Colour	Mutation type	Number of samples (%)
	A>C	11 (3.14%)
	A>G	30 (8.57%)
	A>T	10 (2.86%)
	C>A	28 (8.00%)
	C>T	92 (26.29%)
	C>G	11 (3.14%)
	G>A	110 (31.43%)
	G>C	14 (4.00%)
	G>T	50 (14.29%)
	T>A	9 (2.57%)
	T>C	8 (2.29%)
	T>G	3 (0.86%)
	Total unique samples	350

There are 4 substitutions where the nucleotide change is unknown.

Figura 23: Gráficos de pastel que muestran la distribución de las mutaciones según su tipo, presentes en el gen MAP7D2[32].

Genes sobreexpresados en etapa III.

UGT2B17

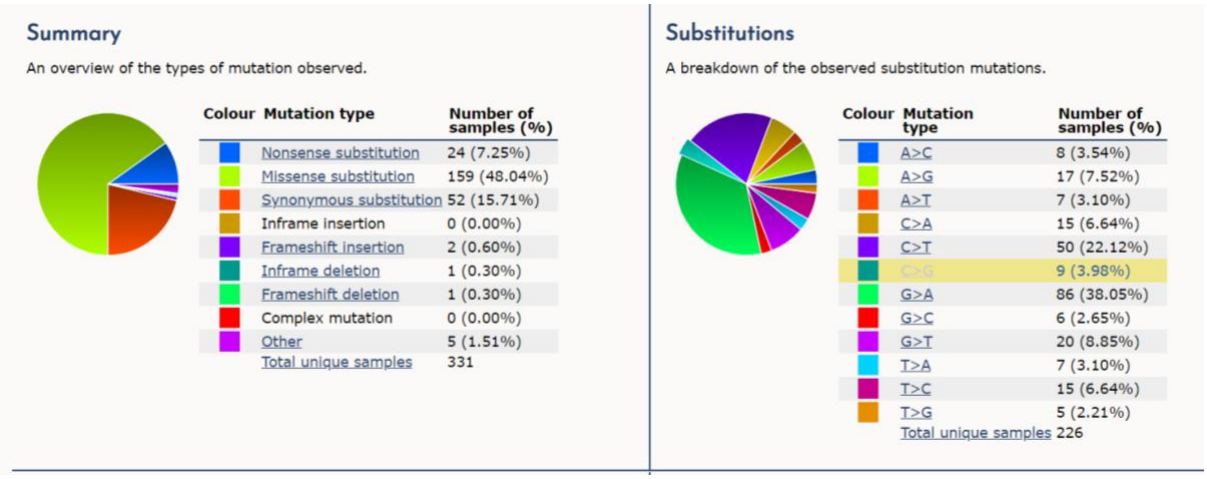


Figura 24: Gráficos de pastel que muestran la distribución de las mutaciones según su tipo, presentes en el gen UGT2B17[33].

REG1A

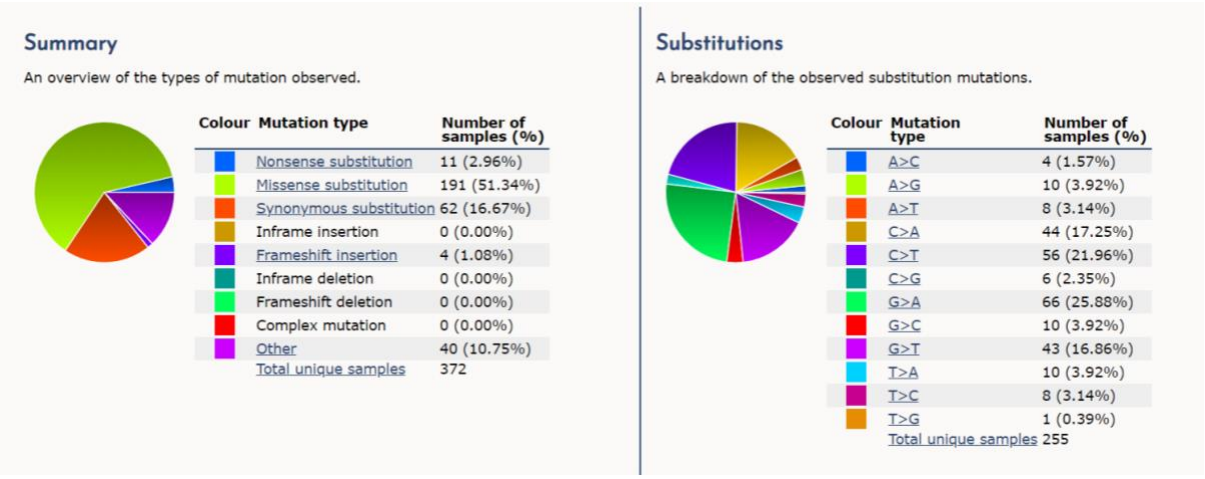


Figura 25: Gráficos de pastel que muestran la distribución de las mutaciones según su tipo, presentes en el gen REG1A [34].

Resumen de los genes 'importantes': A partir de los datos recabados en las secciones anteriores, y la elección de los genes que se consideraron con mayor importancia o relevancia para el cáncer de colon, se resumió la información obtenida en las tablas 4 a 8 que se muestran a continuación para poder visualizar sus características principales.

Resumen de las características de **PSPH**	
Característica	Explicación
Grupo en que se sobreexpresa	Cáncer de colon en etapa II
Resumen de su función (STRING, KEGG, NCBI)	Se encarga de la biosíntesis del aminoácido no esencial L-serina a partir de carbohidratos
Tejidos donde se expresa más	Riñón, Suprarrenal, Tiroides y Cerebro
Interactúa con	PAH
Función molecular (Panther)	Ion binding, Hydrolase activity
Procesos biológicos (Panther)	Metabolismo
Mutación más común (Cosmic)	Mutación sin sentido

Tabla 4: Resumen de las características e información más destacable del gen PSPH.

Resumen de las características de **PAH**	
Característica	Explicación
Grupo en que se sobreexpresa	Cáncer de colon en etapa II
Resumen de su función (STRING, KEGG, NCBI)	Este gen codifica la enzima fenilalanina hidroxilasa, misma que se encarga de hidroxilar la fenilalanina en tirosina y se encarga del catabolismo de la fenilalanina
Tejidos donde se expresa más	Hígado, Riñón, Vesícula biliar
Interactúa con	PSPH
Función molecular (Panther)	-
Procesos biológicos (Panther)	-
Mutación más común (Cosmic)	Mutación sin sentido

Tabla 5: Resumen de las características e información más destacable del gen PAH.

Resumen de las características de **REG1A**	
Característica	Explicación
Grupo en que se sobreexpresa	Cáncer de colon en etapa III
Resumen de su función (STRING, KEGG, NCBI)	Funciona como un inhibidor de precipitaciones espontáneas de calcio y carbonato, que está asociado a la regeneración del páncreas
Tejidos donde se expresa más:	Páncreas, Duodeno, Intestino delgado y estómago
Interactúa con	-
Función molecular (Panther)	Carbohydrate binding, Carbohydrate derivative binding, signalyng receptor activity
Procesos biológicos (Panther)	Regulación de procesos biológicos, Proliferación de células, Respuesta inmune, Respuesta a otros organismos
Mutación más común (Cosmic)	Mutación sin sentido

Tabla 6: Resumen de las características e información más destacable del gen REG1A.

Resumen de las características de **MAP7D2**	
Característica	Explicación
Grupo en que se sobreexpresa	Cáncer de colon en etapa II
Resumen de su función (STRING, KEGG, NCBI)	Es un gen cuya proteína está asociada a procesos de los microtúbulos en células epiteliales.
Tejidos donde se expresa más	Cerebro, Testículos, Placenta, Riñón
Interactúa con	-
Función molecular (Panther)	-
Procesos biológicos (Panther)	Biogénesis, Procesos asociados a microtúbulos
Mutación más común (Cosmic)	Mutación sin sentido

Tabla 7: Resumen de las características e información más destacable del gen MAP7D2.

Resumen de las características de **UGT2B17**	
Característica	Explicación
Grupo en que se sobreexpresa	Cáncer de colon en etapa III
Resumen de su función (STRING, KEGG, NCBI)	Este gen codifica a una proteína encargada de la eliminación de compuestos externos con potencial a ser dañinos para el cuerpo.
Tejidos donde se expresa más	Colon, Intestino delgado, Duodeno e Hígado
Interactúa con	-
Función molecular (Panther)	Transferase activity,
Procesos biológicos (Panther)	-
Mutación más común (Cosmic)	Mutación sin sentido

Tabla 8: Resumen de las características e información más destacable del gen UGT2B17.

6. Análisis de Resultados

Selección del p-value y el tamaño del efecto. En primer lugar, como se puede observar con base en la información de los genes obtenidos, los valores de expresión de los genes son muy parecidos hasta el punto que son pocos los genes que se pudo encontrar estadísticamente que se expresan más en un grupo que en otro, ya fuera en personas con cáncer de colon en etapa II o personas con cáncer de colon en etapa III, incluso cuando el p-value fue de 0.5 y el tamaño del efecto fue de 1.5. Esto se debió a que no se usó alguna muestra control con la cual se pudieran contrastar estos valores para obtener grandes diferencias, sino que fueron pocos los cambios encontrados. Por lo tanto, a partir de lo anterior se puede decir que las células de cáncer de colon en etapa II y en etapa III muestran muy pocas diferencias entre sí y por consiguiente, son pocos los cambios generados en las células para que se den estos cambios contrastantes entre la etapa II y la etapa III. Es por esto que para este análisis se usó un p-value de 0.05 y un tamaño del efecto de 1.5, ya que se estimó que una diferencia significativa entre estos grupos realmente no vería reflejada como una diferencia grande entre los valores de expresión de los genes.

Biomarcadores. La lista de los 46 genes obtenidos corresponden a los genes comunes donde hay una diferencia significativa entre los valores de expresión de las muestras con cáncer de colon en etapa II y etapa III para las personas jóvenes y las grandes, pero el tamaño del efecto no es tomado en cuenta. Por lo tanto, esta lista de genes puede ser útil para ser usada como biomarcadores de detección de cáncer en ambas etapas. Esto es muy importante, ya que con la identificación de la sobreexpresión de estos genes en una persona se puede detectar con mayor anticipación el cáncer en las personas. Aunque, si estos genes se sobreexpresan en la etapa II y etapa III del cáncer de colon, es posible que no se sobreexpresen en la etapa I, misma que suele ser donde hay una menor cantidad de síntomas, por lo tanto, aunque estos genes

pueden ser útiles como biomarcadores, su potencial para ser usados en las etapas más tempranas del cáncer de colon puede ser muy baja. Aunque no se puede asegurar del todo que estos puedan ser usados como biomarcadores de la etapa I del cáncer de colon, sí se podrían usar estos genes para diferenciar alguna enfermedad ligada al colon, o al sistema digestivo en general, que muestre síntomas parecidos al cáncer de colon.

Selección de los genes considerados como 'importantes'. Después de hacer el análisis STRING de la lista de los 46 se observa en la figura 9, se encontró que había algunos genes que estaban relacionados entre sí, por lo que aunque el tamaño del efecto no fuera muy grande, se decidió analizarlos debido a la interacción entre estos, misma que tal vez podría ser clave para entender las causas de la evolución del cáncer de colon. Sin embargo, al analizar las funciones de estos genes se encontró que la razón por la que estaban relacionados era porque están relacionados a la producción de diferentes tipos de ribosomas, razón por la cual están relacionados entre sí. Y aunque una sobreproducción de ribosomas puede derivar en una cantidad anómala de células, esto puede deberse a efectos del cáncer en sí, y no a diferencias significativas entre las etapas de cáncer a analizar. Por lo tanto, a pesar de que estos genes estuvieran relacionados entre sí, no son útiles para entender la evolución del cáncer

Una vez descartados los 10 genes relacionados entre sí de la lista de 46 genes, se intentó acotar la lista de los 11 genes seleccionados (genes con p-value menor a 0.05 y un tamaño del efecto mayor a 1.5) a sólo 5 genes, mismos que fueron seleccionados por sus funciones, procesos biológicos, y relación con otros genes. Es por esto que se obtuvieron los siguientes genes: PSPH, PAH, REG1A, MAP7D2, UGT2B17. Mismos cuyo resumen a grandes rasgos se encuentra en las tablas 4 a 8.

Análisis del gen PSPH: Este gen, como se resume en la tabla 4, se sobreexpresó en el grupo de personas con cáncer de colon en etapa II. y se encarga principalmente de la síntesis de L-serina, misma que se encarga de metabolizar otras macromoléculas como las grasas. Al estar sobreexpresado en la etapa II, y por lo tanto suprimido en la etapa III, podría agravar las condiciones de las personas que padecen cáncer de colon, ya que al no poder metabolizar las grasas, estos nutrientes de su cuerpo se acumularían en diferentes secciones del sistema digestivo, y por lo tanto, tendría una menor capacidad para la absorción de alimentos. Además, según Kuniaki Sato et.al en *Phosphoserine Phosphatase Is a Novel Prognostic Biomarker on Chromosome 7 in Colorectal Cancer* [35], este gen está relacionado muy fuertemente a la metástasis del cáncer de colon. Por lo tanto, la sobreexpresión de este gen puede ser vista como un biomarcador para identificar qué se trata de un cáncer maligno con altas probabilidades de invadir otros tejidos. E incluso, según la investigación *PSPH Mediates the Metastasis and Proliferation of Non-small Cell Lung Cancer through MAPK Signaling Pathways* [36] de Li Liao et.al este gen no sólo está relacionado con el cáncer de colon, sino de otros tipos de cáncer como cáncer de pulmón. En conclusión, se puede afirmar que durante la etapa II de cáncer se produce una sobreexpresión de este gen que a su vez genera la metástasis del cáncer, mientras que en la etapa III de cáncer este gen se reprime, que ocasiona que los síntomas y condición del paciente se agraven debido a que no puede metabolizar ciertos nutrientes de los alimentos que consume.

Análisis del gen PAH: Retomando la información de la tabla 5, observamos que este gen se sobreexpresó en el grupo de personas con cáncer de colon en etapa II, además de que este se encuentra relacionado con el gen PSPH. Su función es codificar la enzima fenilalanina hidroxilasa. Al analizar esta información se pueden identificar diversos datos relevantes del gen. Comenzando de que este se suprime en personas con cáncer de colon en la etapa III. Se realizó una búsqueda de artículos sobre el gen y su relación con el cáncer de colon, en base a lo expuesto en el artículo *E3 Ubiquitin Ligase APC/C Cdh1 Regulation of Phenylalanine Hydroxylase Stability and Function* [37] realizado por Apoorvi Tyagi et. al, se menciona que de alterarse los niveles de PAH se genera una tóxica acumulación de Fenilcetonuria (PKU) en la sangre y el cerebro. De igual forma, en el artículo se menciona de manera breve el uso del gen PAH como biomarcador en relación al cáncer de colon, es por ello, que se considera que se requiere la realización de más investigaciones sobre el gen PAH y su relación como biomarcador en el cáncer de colon. Por otra parte, también es importante resaltar la sobreexpresión del gen en tejidos como el hígado, riñón y vesícula biliar.

Análisis del gen REG1A: Con base en los datos obtenidos acerca del gen en la tabla 6, se observa que este gen funciona como un inhibidor precipitaciones de calcio y carbonato, sobretodo en tejidos asociados al aparato digestivo como el páncreas, duodeno, intestino delgado y estómago. Además de que está asociado a la unión de carbohidratos, la proliferación de células y la regulación de procesos biológicos. Según la investigación *REG1A expression is a prognostic marker in colorectal cancer and associated with peritoneal carcinomatosis* [38] realizada por Christian Astrosini et. al este gen está altamente relacionado a la formación temprana de tumores en el colon y que incluso este gen también se sobreexpresa en los tejidos adyacentes al tumor. Además, está altamente relacionado con la metástasis del cáncer, y es por esto que también se encontró que está relacionado a la recurrencia de la enfermedad en personas que ya habían superado el cáncer, y también influye mucho en la supervivencia de la persona. Por lo tanto, a partir de lo anterior se puede afirmar que este gen se sobreexpresa más en la etapa III de cáncer de colon debido a que actúa como un detonante principal de los tumores en el colon, y está altamente relacionado con la invasión de otros tejidos. Es por esto que la sobreexpresión de este gen puede encontrarse desde etapas muy tempranas de su desarrollo y que es muy importante hacerse más estudios en caso de presentar una sobreexpresión de este gen aún sin tener, ya que una vez que este gen se haya sobreexpresado aún más, puede ser mortal debido a la alta posibilidad de metástasis y a la baja probabilidad estadística de supervivencia y la alta probabilidad de recurrencia.

Análisis del gen MAP7D2: Este gen, cuyo resumen se encuentra en la tabla 7, se expresó más en el grupo de cáncer de colon en etapa II, y lo más destacable de este gen es que está asociado a los procesos de los microtúbulos y a otros procesos en células epiteliales. Estos microtúbulos pueden jugar un papel muy importante en este estadío del cáncer, pues como señala Craig Blum et. al en el artículo *The expression ratio of Map7/B2M is prognostic for survival in patients with stage II colon cancer* [39], debido a que este gen está asociado al comportamiento de los microtúbulos, participa activamente al control de la división de las células hasta el punto que actualmente es uno de los genes con mayor potencial para causar metástasis y evolución del cáncer de colon debido a la escasez de un control de la división

celular, pues la sobreexpresión de este gen ocasiona que el periodo en que la célula se mantiene en interfase es menor. En resumen, este gen se sobreexpresa más en la etapa II del cáncer ya que sus funciones están asociadas a los procesos de los microtúbulos, y por lo tanto, están directamente relacionados a la división celular sin control que da origen a la formación de tumores, y la metástasis del cáncer.

Análisis del gen UGT2B17: Este gen, como se muestra en la tabla 8, se expresa sobre todo en las personas con cáncer de colon en etapa III, y su función principal es la eliminación de agentes externos al cuerpo que en su mayoría pueden ser dañinos. Y este gen se expresa en tejidos del sistema digestivo como el colon, el intestino delgado, el duodeno y el hígado. Sin embargo, a diferencia de los genes mencionados anteriormente, la sobreexpresión de este gen no es nociva, pues como indican Andrea Y. Angstadt et al *en The effect of copy number variation (CNV) in the phase II detoxification genes, UGT2B17 and UGT2B28, on colorectal cancer risk* [40], este gen lo que hace es combatir el cáncer de colon, pues se ha comprobado que una sobreexpresión de este gen está correlacionada con una disminución en el riesgo de cáncer de colon. Esto debido a que genera sustancias que sirven como antioxidantes que contrarrestan los efectos del cáncer. Aunque según la misma investigación, esta mutación se observa más en hombres que en mujeres. En otras palabras, la sobreexpresión de este gen en la etapa III de cáncer está relacionada con una respuesta del cuerpo contra el cáncer de colon, comportamiento más común en hombres que en mujeres, por lo que el estudio de los detonadores de la mutación de este gen pueden llevar a un tratamiento eficaz del cáncer de colon.

Alcances y limitaciones: Con los datos recabados en este análisis se pudo realizar un buen análisis de los genes seleccionados, sin embargo esta investigación puede estar sesgada por varias razones. Debido a que este análisis se realizó con una cantidad de muestras relativamente pequeña, es posible que algunos genes hayan sido falsos positivos, pues algunos genes no tenían características relacionadas al cáncer de colon. Además, las muestras tenían características físicas muy parecidas, pues aunque las edades variaron entre los 50 y 90 años, todos eran de la misma raza (caucásica) e incluso de la misma nacionalidad (danesa). Además, las funciones y procesos biológicos de algunos genes no se encontraban en las bases de datos seleccionadas para el análisis, esto se debe a que hay una gran cantidad de genes y para ello se debe hacer un estudio de cada gen y así obtener todas estas características de los genes. Es por esto que un estudio más amplio de los genes que se seleccionaron estadísticamente enriquecería futuros estudios. E incluso, la revisión por pares de este tipo de estudios puede confirmar o negar las hipótesis planteadas en el presente análisis

7. Reflexión

En conclusión a este trabajo de investigación, análisis científico y estadístico sobre la detección de biomarcadores de cáncer de colon utilizando 20 muestras aleatorias de la base de datos GSE31595[4] según los cuatro grupos seleccionados (adultos jóvenes con cáncer de colon en etapa II, adultos jóvenes con cáncer de colon en etapa III, adultos grandes con cáncer de colon en etapa II y adultos grandes con cancer de colon en etapa III) y sometiéndolas al software de análisis estadístico R, se obtuvo el resultado de una lista de 11 biomarcadores

(PSPH, ADGRG7, LOC101929036 / PAH, UGT2B17, KRT23, CLCA1, MAP7D2, XIST, CNTN3, FLJ22763, REG1A) gracias a que tuvieron un p-value menor a 0.05 y un tamaño de efecto del logaritmo base 2 de 1.4. Sin embargo, esta lista se redujo a cinco biomarcadores (PSPH, PAH, REG1A, MAP7D2, UGT2B17) tomando en cuenta los elementos mas destacables de los mismos como sus funciones, procesos biológicos y relación con otros genes. Para obtener esta lista de cinco biomarcadores se llevó a cabo un proceso extenso de investigación, iniciando por la selección de la base de datos, puesto que a pesar de todos los sitios que poseen esta clase de datasets, fue imperativo revisar todas las opciones y compararlas para realizar la mejor selección. Se consideró lo siguiente para esta selección: número de muestras, número de plataformas utilizadas en las muestras, si los datos estaban pre-tratados o no y las opciones de grupos a comparar. El dataset contiene un total de 37 muestras y de una sola plataforma, además de que se consideró relevante realizar una comparación de las muestras según la etapa del cáncer de colon diagnosticado. Posteriormente, se realizó el análisis más significativo del trabajo, el análisis estadístico en R, para poder realizar el mismo se tuvieron que utilizar parámetros específicos y consideraciones importantes, como el número de muestras a analizar, puesto que tener pocas podría perjudicar los resultados además, de que aunque estas pruebas eran aleatorias, se realizó el análisis en R con diversos conjuntos de estas para poder visualizar y seleccionar las que considerábamos proporcionaban mejores resultados a la investigación, esto en base a lo que se observaba en las gráficas, sobre todo en la gráfica de volcano, donde se observa qué genes se sobreexpresan y cuales se reprimen en cada grupo (etapa II y etapa III). Sin embargo, los parámetros más importantes para el análisis de R fueron primero, en la normalización de datos, el uso de trim=0.02, el cual eliminó una porción de los datos más grandes y pequeños. Otro parámetro importante fue el tamaño de efecto del logaritmo base 2. La prueba t-test para obtener un p-value, el cual, en este caso fue de 0.05, considerado un poco laxo para permitir tener más datos para los grupos que se estaban comparando, esto también influyó en la aceptación de la hipótesis nula, puesto que, siendo la hipótesis alternativa que los genes de expresión fueran diferentes entre los grupos, con ayuda del p-value se reducjo la lista a 46 genes los cuales estaban presentes en ambos grupos. Como ya se mencionó, estas consideraciones resultan muy importantes para la obtención de resultados, por ejemplo, en la prueba t-test, de mantener un p-value igual a 0.01, la cantidad de genes presentes en cada grupo (etapa II y etapa III) disminuye y cuando se quiere obtener la siguiente lista, estos grupos no tienen ningún gen en común y por lo tanto no se tendría ningún gen sobreexpresado o reprimido, afectando así los resultados del trabajo. Incluso, el efecto de los datos al logaritmo base 2, porque esto puede dar indicios de que quizás se seleccionaron datos que ya tienen un pretratamiento, afectando así el resto del análisis estadístico en R. Por último, considero de suma importancia recalcar uno de los objetivos implícitos en la realización de este trabajo, que es, como la aplicación de la ciencia de datos impacta en la resolución de problemáticas de salud pública. No solo el trabajo de nuestra autoría, sino todas las bases de datos, artículos e investigaciones que consultamos y exploramos demuestran el impacto de la ciencia de datos en las problemáticas de salud. Esta fue una investigación compleja, la cual, puedo verse completamente obstaculizada de no ser por las herramientas que se aplicaron, como el uso del software de análisis estadístico R, que contiene paquetes como *limma*, es gracias a la ciencia de datos que cientos de investigaciones se pueden llevar a cabo al día de hoy, no solo por la variedad de herramientas para manejar y analizar la información, también

por la cantidad de páginas que contienen todas estas bases de datos y están disponibles para realizar análisis estadísticos que permitan obtener resultados contundentes. Gracias a la ciencia de datos logramos identificar los biomarcadores, así como muchos otros investigadores, los cuales impactaron de manera positiva a un gran problema de salud, que es el cáncer, donde gracias a sus investigaciones se pueden implementar mejores protocolos de detección y tratamientos más efectivos, dando a los pacientes, alternativas que hace un par de años no se consideraban posibles. De igual forma, así como con este trabajo se muestra su impacto en el cáncer de colon, existen cientos de investigaciones enfocadas en diversas enfermedades con el objetivo de seguir aprendiendo de estas para poder tratarlas. La ciencia de datos no solo es una vía para las soluciones, también implica una gran ventaja de avance en periodos más cortos y por ello, el desarrollo de los diversos campos en ciencia de datos en indispensable para el avance en la sociedad.

Referencias

[1] Sánchez, M. (2019). *Cáncer de colon*. abril 18, 2021, de CuidatePlus Sitio web: https://cuidateplus.marca.com/enfermedades/cancer/cancer-de-colon.html

[2] Medline (2020) Cáncer de colon. abril 18, 2021, de MedlinePlus. Sitio web https://medlineplus.gov/spanish/ency/article/000262.htm

[3] Arvelo, F. (2014). *Biología del cáncer de colon.* abril 18, 2021, de Laboratorio de Cultivo de Tejidos y Biología de Tumores, Instituto de Biología Experimental, Universidad Central de Venezuela, Sitio web: https://ecancer.org/es/journal/article/520-biology-of-colorectal-cancer/pdf/es

[6] NCBI. (2011). *Gene Expression Profiles in Stage II and III Colon Cancer. Application of a 128-gene signature*. abril 18, 2021, de NCBI Sitio web: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31595

[5] STRING. (2021). *STRING - Search.* Abril 25 2021, de STRING Sitio web: https://string-db.org/

[6] KEGG. (2020). *Homo sapiens (human): 256714*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:256714

[7] KEGG. (2020). *Homo sapiens (human): 5967*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:5967

[8] KEGG. (2020). *Homo sapiens (human): 5723*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:5723

[9] KEGG. (2020). *Homo sapiens (human): 5053*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:5053

[10] KEGG. (2020). *Homo sapiens (human): 25984*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:25984

[11] KEGG. (2020). *Homo sapiens (human): 84873*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:84873

[12] KEGG. (2020). *Homo sapiens (human): 7367*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:7367

[13] KEGG. (2020). *Homo sapiens (human): 1179*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:1179

[14]KEGG. (2020). *Homo sapiens (human): 9383*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:9383
[15] KEGG. (2020). *Homo sapiens (human): 5067*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:5067
[16] KEGG. (2020). *Homo sapiens (human): 401081*. abril 20, 2021, de KEGG Sitio web: https://www.genome.jp/entry/hsa:401081
[17] Reactome. (2021). *UniProt:Q14244 MAP7.* abril 20, 2021, de Reactome Sitio web: https://reactome.org/content/detail/interactor/Q14244
[18] NCBI. (2021). *REG1A regenerating family member 1 alpha [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web:https://www.ncbi.nlm.nih.gov/gene/5967
[19] NCBI. (2021). *PSPH phosphoserine phosphatase [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web:https://www.ncbi.nlm.nih.gov/gene/5723
[20] NCBI. (2021). *PAH phenylalanine hydroxylase [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web:https://www.ncbi.nlm.nih.gov/gene/5053
[21] NCBI. (2021). *KRT23 keratin 23 [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web:https://www.ncbi.nlm.nih.gov/gene/25984
[22] NCBI. (2021). *ADGRG7 adhesion G protein-coupled receptor G7 [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web:https://www.ncbi.nlm.nih.gov/gene/84873
[23] NCBI. (2021). *UGT2B17 UDP glucuronosyltransferase family 2 member B17 [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web: https://www.ncbi.nlm.nih.gov/gene/7367
[24] NCBI. (2021). *CLCA1 chloride channel accessory 1 [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web:https://www.ncbi.nlm.nih.gov/gene/1179
[25] NCBI. (2021). *XIST X inactive specific transcript [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web:https://www.ncbi.nlm.nih.gov/gene/7503
[26] NCBI. (2021). *CNTN3 contactin 3 [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web:https://www.ncbi.nlm.nih.gov/gene/5067
[27] NCBI. (2021). *C3orf85 chromosome 3 open reading frame 85 [Homo sapiens (human)]*. Abril 20 2021, de NCBI Sitio web: https://www.ncbi.nlm.nih.gov/gene/401081
[28] Panther. (2021). PANTHER - *Gene List Analysis.* http://pantherdb.org/geneListAnalysis.do
[29] NCBI. (2021). *MAP7D2 MAP7 domain containing 2 [Homo sapiens (human)].* Abril 27 2021, de NCBI Sitio web: https://www.ncbi.nlm.nih.gov/gene/256714
[30] COSMIC. (2020, agosto 27). *PSPH Gene*. https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=PSPH
[31] COSMIC. (2020, agosto 27). *PAH Gene*. https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=PAH
[32] COSMIC. (2020, agosto 27). *MAP7D2 Gene*. https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=MAP7D2
[33] COSMIC. (2020, agosto 27). *UGT2B17 Gene*. https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=UGT2B17
[34] COSMIC. (2020, agosto 27). *REG1A Gene*. https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=REG1A

[35] Sato K, Masuda T, Hu Q, Tobo T, Kidogami S, Ogawa Y, Saito T, Nambara S, Komatsu H, Hirata H, Sakimura S, Uchi R, Hayashi N, Iguchi T, Eguchi H, Ito S, Nakagawa T, Mimori K. Phosphoserine Phosphatase Is a Novel Prognostic Biomarker on Chromosome 7 in Colorectal Cancer. Anticancer Res. 2017 May;37(5):2365-2371. doi: 10.21873/anticanres.11574. PMID: 28476802.
[36] Liao L, Ge M, Zhan Q, Huang R, Ji X, Liang X, Zhou X. PSPH Mediates the Metastasis and Proliferation of Non-small Cell Lung Cancer through MAPK Signaling Pathways. Int J Biol Sci. 2019 Jan 1;15(1):183-194. doi: 10.7150/ijbs.29203. PMID: 30662358; PMCID: PMC6329917.
[37] Tyagi A, Sarodaya N, Kaushal K, Chandrasekaran AP, Antao AM, Suresh B, Rhie BH, Kim KS, Ramakrishna S. E3 Ubiquitin Ligase APC/CCdh1 Regulation of Phenylalanine Hydroxylase Stability and Function. Int J Mol Sci. 2020 Nov 28;21(23):9076. doi: 10.3390/ijms21239076. PMID: 33260674; PMCID: PMC7729981.
[38] Astrosini C, Roeefzaad C, Dai YY, Dieckgraefe BK, Jöns T, Kemmner W. REG1A expression is a prognostic marker in colorectal cancer and associated with peritoneal carcinomatosis. Int J Cancer. 2008 Jul 15;123(2):409-413. doi: 10.1002/ijc.23466. PMID: 18452172.
[39] Blum C, Graham A, Yousefzadeh M, Shrout J, Benjamin K, Krishna M, Hoda R, Hoda R, Cole DJ, Garrett-Mayer E, Reed C, Wallace M, Mitas M. The expression ratio of Map7/B2M is prognostic for survival in patients with stage II colon cancer. Int J Oncol. 2008 Sep;33(3):579-84. PMID: 18695889; PMCID: PMC3399116.
[40] Angstadt AY, Berg A, Zhu J, Miller P, Hartman TJ, Lesko SM, Muscat JE, Lazarus P, Gallagher CJ. The effect of copy number variation in the phase II detoxification genes UGT2B17 and UGT2B28 on colorectal cancer risk. Cancer. 2013 Jul 1;119(13):2477-85. doi: 10.1002/cncr.28009. Epub 2013 Apr 10. PMID: 23575887; PMCID: PMC3686841.