UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

"Análisis bioinformático para la identificación de genes expresados diferencialmente en la progresión del carcinoma hepatocelular"

TESIS

QUE PARA OBTENER EL TÍTULO DE:

Licenciatura en Matemáticas Aplicadas y Computación

PRESENTA:

César Osvaldo Martínez Cantú

Asesores:

Dr. Jaime Arellanes Robledo

Dr. José Jaime Martínez Magaña

M. en C. José Antonio Coria Fernández

SANTA CRUZ ACATLÁN, NAUCALPAN, EDO. DE MÉXICO, 2022

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Financiamiento

El presente proyecto se realizó con recursos del Consejo Nacional de Ciencia y Tecnología – CONACYT-Ciencia de Frontera 2019, financiamiento número CF2019-53358, y del Instituto Nacional de Medicina Genómica – INMEGEN, financiamiento número 06/2017/I; otorgados al Dr. Jaime Arellanes Robledo.

Objetivos de la investigación

Hipótesis

El análisis bioinformático de la expresión genética global identifica genes que son alterados desde las etapas tempranas hasta el establecimiento del cáncer, durante la progresión del carcinoma hepatocelular en el ratón.

Objetivo general

Identificar genes expresados diferencialmente en la progresión del carcinoma hepatocelular, en un modelo de ratón a través del análisis bioinformático de la expresión genética global obtenida mediante la tecnología de microarreglos.

Objetivos específicos

1. Identificar genes expresados diferencialmente durante la progresión del carcinoma hepatocelular.

Resumen

Introducción: El carcinoma hepatocelular (CHC) es el cáncer primario de hígado más común y la segunda causa de muerte por cáncer a nivel mundial. Este tipo de cáncer es causado por diferentes factores de riesgo como la exposición al virus de la hepatitis, el consumo de alimentos contaminados con diferentes hepatotoxinas, el consumo excesivo de alcohol, entre otros. Uno de los grandes retos de esta enfermedad es que su desarrollo es asintomático por lo que generalmente es detectado en etapas avanzadas de su desarrollo.

Objetivo: Identificar genes expresados diferencialmente en la progresión del CHC en el ratón, desde etapas tempranas hasta el establecimiento del cáncer, a través del análisis bioinformático de la expresión genética global obtenida mediante la tecnología de microarreglos de ADN.

Metodología: Se realizó un análisis bioinformático de la expresión genética global explorada a las 6, 10, 14 y 18 semanas durante la progresión del CHC inducido por el carcinógeno dietilnitrosamina (DEN) en el ratón, mediante un script utilizando el lenguaje de programación R con el apoyo de librerías de bioconductor tales como: oligo, ArrayExpress y limma. Estas herramientas informáticas permitieron la lectura y procesamiento de datos de expresión génica de la progresión del CHC, obtenidos del análisis de microarreglos de ADN.

Resultados: Los análisis de componentes principales revelaron que los datos del grupo tratado con DEN por 6 semanas tienen una dimensionalidad más cercana al grupo control que con aquellos tratados por 10, 14, y 18 semanas. El mapa de calor confirmó que los grupos tratados por 10, 14 y 18 semanas con DEN tuvieron un cambio mayor en su expresión génica, mientras que los tratados por 6 semanas no presentaron cambios significativos. Se identificaron 40 y 11 genes expresados diferencialmente con un valor de 2 y 3 veces de cambio $(p \leq 0.05)$, respectivamente, respecto al grupo de control.

Conclusiones: Como se esperaba, en la medida que avanza el desarrollo del CHC, se inducen mayores cambios a nivel de la expresión génica global hepática. Por lo cual, al realizarse el análisis de microarreglos de ADN utilizando el script desarrollado, se identificaron genes expresados diferencialmente en la progresión del CHC, por lo que se propuso 2 listados de genes que podrían ser relevantes para el desarrollo de la enfermedad.

Agradecimientos

En esta sección me gustaría expresar mi agradecimiento a las personas que me apoyaron de una u otra manera para poder concluir mi proyecto de titulación. Agradezco profundamente a mis padres por darme la oportunidad de estudiar y culminar una carrera.

Agradezco a mis asesores por el tiempo y apoyo brindado en la culminación del presente trabajo.

Agradezco a la Universidad Nacional Autónoma de México por el conocimiento y experiencias, que pude obtener al ser parte de una de sus instituciones como la Facultad de Estudios Superiores Acatlán. También me gustaría agradecer al Instituto Nacional de Medicina Genómica (INMEGEN) por haberme dado la oportunidad de colaborar en uno de sus proyectos.

Además, quiero agradecer a mis amigos por su apoyo y ánimos brindados para lograr concluir este proyecto.

Índice General

Glosario

ADN: Ácido Desoxirribonucleico. Material que contiene la información hereditaria.

Adenina (A): Compuesto químico que las células usan para elaborar los elementos fundamentales del ADN y el ARN.

ARN: Ácido ribonucleico. Estructura similar al ADN, pero con tareas diferentes como es la transferencia de información del genoma.

Arreglo: Conjunto de datos o estructura de datos, ubicados generalmente de manera consecutiva.

Bases Nitrogenadas: Cualquier compuesto químico que constituye el ácido nucleico.

CHC: Carcinoma hepatocelular. Tipo común de cáncer de hígado.

Cirrosis: Enfermedad crónica e irreversible del hígado que se origina a causa de la destrucción de las células hepáticas.

Citosina (C): Compuesto químico que las células usan para elaborar los elementos fundamentales del ADN y el ARN.

Codón: Nombre que se le da a un tramo de 3 nucleótidos.

DEN: Dieltilnitrosamina, químico que induce el cáncer.

Epiteliales: Tipo de célula que recubre la superficie del organismo.

Fibrosis: Formación patológica de tejido en un órgano del cuerpo.

Fold change: Valor de cambio.

Genómica estructural: Identificación y estudio de las variantes estructurales de secuencia en los genomas.

Genómica funcional: Determina la función biológica de los genes y sus productos

Genómica individual: Estudio de secuenciación y análisis del genoma de un individuo

Genómica comparativa: Estudio comparativo de los genomas estructural y funcionalmente en organismos

Grupo Control: Grupo sano al que no se le administró ningún químico, utilizado para realizar comparaciones contra otro grupo de estudio.

Grupo Tratamiento: Grupo experimental al que se le administran diferentes químicos con el objetivo de realizar comparaciones.

Guanina (G): Compuesto químico que las células usan para elaborar los elementos fundamentales del ADN y el ARN.

Hipocondrio: Desde el punto de vista etimológico, lugar bajo las costillas. Desde la perspectiva anatómica, se refiere a los cuadrantes superiores del abdomen que están bajo las parrillas costales que lo cubren parcialmente.

Homeostasis: Estado de equilibrio entre todos los sistemas del cuerpo que se necesitan para sobrevivir y funcionar correctamente.

Metadato: Toda aquella información descriptiva sobre el contexto, calidad, condición o características de un recurso, dato u objeto que tiene la finalidad de facilitar su recuperación, autentificación, evaluación, preservación y/o interoperabilidad.

Normalización: Proceso para identificar y eliminar errores sistemáticos.

Oligonucleótidos: Secuencia corta de ADN o ARN, con cincuenta pares de bases o menos

Ómicas: Estudio de la totalidad o de conjunto de algo, como genes, organismos de un ecosistema, proteínas, e incluso la relación entre estos.

Paliativo: Tratamiento médico para pacientes con cáncer terminal.

Proteoma: Es la totalidad de proteínas expresadas en una célula particular bajo condiciones de medio ambiente y etapas de desarrollo específicas.

Ribosoma: Partícula celular hecha de ARN y proteína que sirve como el sitio para la síntesis de proteínas en la célula.

RMA (Promedio robusto multi arreglos - Robust Multiarray Average): Método utilizado para realizar el preprocesamiento de los microarreglos de affymetrix.

SDRF: Formato de relación de datos y muestras - *Sample and Data Relationship Format*

Sondas: Posición dentro de una matriz de información que contiene una hebra de ADN.

Transcriptoma: Conjunto de moléculas de ARN mensajero y de ARN no codificante presentes en una célula o tejido concreto.

Timina (T): Compuesto químico que las células usan para elaborar los elementos fundamentales del ADN y el ARN.

VHB (infección): Virus que causa la hepatitis (inflamación del hígado).

Introducción

Una de las principales características de las células cancerosas es la proliferación descontrolada. De acuerdo con la definición de Meza-Junco, el cáncer comprende un grupo de enfermedades caracterizadas por proliferación autónoma de células neoplásicas que tienen varias alteraciones, incluyendo mutaciones e inestabilidad genética (Meza-Junco et al, 2006).

Desarrollar un tratamiento para dicha enfermedad se vuelve complejo debido a la manera en cómo progresa el cáncer. La OMS resalta lo siguiente: una característica definitoria del cáncer es su rápida multiplicación de células anormales, las cuales pueden propagarse a otros órganos. Este proceso se denomina metástasis y es la principal causa de muerte por cáncer (OMS, 2018). Sin embargo, la OMS también explica que el cáncer se reproduce por la transformación de células normales en células tumorales por un proceso que se ejecuta en varias etapas, el cual empieza en la progresión de una lesión precancerosa a un tumor maligno (OMS, 2018).

Como resultado de estos acontecimientos se han explorado otras alternativas como son las ciencias genómicas, las cuales se centran en el estudio del material genético. Esto debido a que el genoma contiene toda la información hereditaria de cualquier organismo. En este sentido, una de las estrategias para investigar el comportamiento del cáncer es la identificación de genes expresados diferencialmente en su progresión en modelos experimentales y en muestras provenientes de seres humanos a través de análisis bioinformáticos de la expresión genética global obtenida mediante la tecnología de microarreglos (Lewin, 2004).

Así, con el fin de apoyar los análisis genómicos se han utilizado disciplinas complementarias como la bioinformática. Aunque no existe una definición universalmente aceptada de bioinformática, actualmente el término denota un campo relacionado con la aplicación de técnicas, algoritmos y herramientas de tecnología de la información para resolver problemas en ciencias biológicas (Chapman & Hall, 2012). Por lo tanto, con el apoyo de estas

herramientas, se pueden identificar genes que podrían jugar un papel relevante en la progresión de la enfermedad. Como consecuencia, se podrá conocer la alteración de la expresión de genes específicos que puedan ser de utilidad para explorar nuevas alternativas de tratamiento en un futuro cercano.

Para realizar estudios sobre el comportamiento de la expresión genética global de un órgano específico como el hígado, se utiliza una tecnología llamada microarreglos de expresión genética (también son conocidos como Microchips de expresión). El uso de microarreglos permite conocer lo que sucede cuando las células son alteradas por un estímulo externo o por eventos de transformación maligna. De tal manera estamos aprendiendo que la estimulación de una célula o daño celular regula la expresión de una amplia variedad de genes que codifican una proteína (Soto, 2003).

Esta tesis presenta un análisis sobre la progresión del carcinoma hepatocelular (CHC) inducido en el ratón, dado que el CHC inducido por la dietilnitrosamina (DEN) en modelos de ratones tiene características histológicas y genéticas cercanas a las observadas en el CHC humano (Lee et al., 2004). Los animales fueron sometidos a los efectos del carcinógeno DEN a una dosis de 20 mg/kg en un lapso de 6 a 18 semanas, con el objetivo de interpretar el comportamiento de expresión genética. Estos parámetros fueron obtenidos de los resultados del trabajo elaborador por el Biólogo, Sergio Fuentes, donde indica lo siguiente: La dosis de DEN de mayor efectividad para inducir la hepatocarcinogénesis fue la de 20 mg/ks ya que indujo la producción de fibras de colágeno, tumores hepáticos y metástasis pulmonar, a las 18 semanas de administración (Fuentes Hernandez, S., 2018).

Adicionalmente, se explica la manera en cómo fue utilizado el lenguaje de programación R para realizar el análisis bioinformático de la expresión genética global obtenida mediante análisis de microarreglos proporcionado por investigadores del Instituto Nacional de Medicina Genómica (INMEGEN).

El capítulo 1 describe las características del hígado y sus principales funciones; así como también, las características del cáncer con especial énfasis en el CHC; lo anterior, como tema central de esta investigación.

El capítulo 2 expone un contexto general de los conceptos biológicos; los cuales serán usados durante el presente trabajo.

El capítulo 3 explica a detalle el análisis de microarreglos de expresión génica; así como las diferentes etapas involucradas en el análisis de dicha tecnología, y con esto, obtener un número de genes relevantes en la progresión cronológica del CHC.

El capítulo 4 ejemplifica a detalle la manera de utilizar el lenguaje de programación R, con el objetivo de explotar los metadatos compartidos mediante la tecnología de microarreglos de expresión génica.

Planteamiento del problema

El CHC es una de las enfermedades más letales a nivel mundial. En gran medida se debe a que el desarrollo de esta enfermedad es asintomático; como consecuencia, se detecta generalmente en etapas avanzadas de su desarrollo. Por lo tanto, detectar alteraciones en el transcriptoma de su progresión, desde etapas tempranas hasta el establecimiento del cáncer, representa una ventana de oportunidad para identificar de manera específica genes que jueguen un papel central en este proceso. En este sentido el presente trabajo tiene la finalidad de realizar un análisis bioinformático de la expresión genética global obtenida mediante un análisis de microarreglos realizado a partir de muestras provenientes de diferentes etapas del desarrollo del CHC en el ratón.

Capítulo 1

Cáncer de Hígado

1.1. El Hígado

El hígado es un órgano con un peso que varía de 1.5 a 2 kg y se localiza en el hipocondrio derecho. Anatómicamente se divide en cuatro lóbulos; el derecho, el izquierdo, el cuadrado y el de Spiegel o lóbulo caudado. El sistema vascular está constituido principalmente por la vena porta y la arteria hepática. La vena porta suministra cerca del 70% del flujo sanguíneo y el 40% del oxígeno mientras que la arteria hepática suministra el 30% del flujo sanguíneo y el 60% del oxígeno (Torres Mena, J., E., 2016). De acuerdo con la definición de Rojas Lemusa se puede decir que el hígado es la glándula más grande del cuerpo y en él se identifican dos estructuras principales, el parénquima y el estroma (Rojas Lemusa, M., 2017). La estructura de tipo estroma es prácticamente el tejido del hígado, el cual recibe el nombre de "cápsula de Glisson" dado que cubre la superficie externa del hígado. Posteriormente, este tipo de estructura se comunica con la de tipo parénquima la cual es la responsable de realizar las funciones del hígado, como es la transmisión de los elementos vasculares. Dichas estructuras se encuentran formadas por 3 tipos celulares principales, que en orden de abundancia son: hepatocitos, células de Kupffer (macrófagos) y células de Ito (lipocitos) (Rojas Lemusa, M., 2017). Cada una de estas células se encarga de realizar diferentes tareas para apoyar en las funciones del hígado, entre las cuales se enumeran las siguientes:

1. Digestión de los alimentos: Extrae los nutrientes esenciales para la digestión, los cuales son carbohidratos, lípidos y proteínas. Al mismo tiempo, el hígado secreta bilis para ayudar a descomponer las grasas (Kern Parma, 2018).

2. Almacenamiento de energía: Convierte los nutrientes extraídos en forma de azúcar, de tal modo, que el organismo podrá utilizarlos cuando sea necesario (Kern Parma, 2018).
3. Eliminación de sustancias tóxicas: Además tiene la capacidad de filtrar y eliminar las toxinas provenientes de lo que consumimos, como pueden ser grandes cantidades de alcohol o la ingesta de medicamentos (Kern Parma, 2018).

Una de las células con más relevancia en el hígado es el hepatocito, dado que es metabólicamente muy activo y es, quizá, la célula más versátil del organismo. Todos y cada uno de los hepatocitos que conforman al hígado cumplen íntegramente con las funciones que se describieron (y con muchas otras), lo que implica una alta relevancia en la homeostasis del ser humano (Rojas Lemusa, M., 2017).

Al ser un órgano con una gran cantidad de células funcionalmente activas, lo hace propenso a ser un blanco perfecto de alteraciones debido a su contacto directo con factores hepatotóxicos tanto exógenas como endógenas como el virus de la hepatitis y las especies reactivas de oxígeno, respectivamente. Estos factores pueden inducir alteraciones como la fibrosis, cirrosis y el CHC, el cual representa entre el 85% y 90% de los casos de cáncer primario de hígado y la segunda causa de muerte por cáncer en el mundo (Torres Mena, J., E., 2016).

Una forma adecuada de prevenir estos padecimientos es la alimentación balanceada, en la cual se debe evitar la ingesta de alimentos que contengan una alta cantidad de grasas y de ser posible no ingerir bebidas alcohólicas en grandes cantidades. En resumen, como para la mayoría de las enfermedades, la buena alimentación y una vida activa sin excesos, son importantes para mantener el organismo saludable (Kern Parma, 2018).

1.2. El cáncer

El cáncer se clasifica de acuerdo con el tejido y tipo celular del cual surge. Si proviene de células epiteliales es denominado carcinoma, el que se origina de células de tejido conectivo o musculares se llama sarcoma, el que se origina en el cerebro y sistema nervioso central se denomina cáncer neuro ectodérmico y, por último, el que surge de células de la sangre y linfa

se denomina cáncer del tejido hematopoyético, como la leucemia, el mieloma y el linfoma (Dangoor Education, 2020).

Los diferentes tipos de células que alberga el hígado pueden formar tumores malignos (cancerosos) y benignos (no cancerosos), debido a una alteración en el ciclo celular. Estos tumores pueden desarrollarse en cualquier parte del hígado, de tal modo que podrán ser tratados con base a un diagnóstico previo (Sociedad Americana Contra el Cáncer, 2019).

Es importante mencionar que el hígado tiene una gran capacidad regenerativa, solamente superada en la placenta, pero esto no llega a ser suficiente cuando es agredido constantemente y de forma crónica, como sucede con la ingesta de alcohol, infecciones virales o enfermedades autoinmunes, las cuales puede inducir un daño hepático irreversible (Rojas Lemusa, M., 2017).

Sin embargo, no se puede asumir del todo que el cáncer de hígado únicamente puede ser originado por una ingesta excesiva de alcohol o por la falla regenerativa celular, también puede ser hereditario, tomando en cuenta la perspectiva de Hannban y Weinberg, donde explican que las células cancerosas son la base de la enfermedad; inician tumores y conducen la progresión tumoral, acumulan mutaciones oncogénicas en genes supresores de tumores, por lo que el cáncer es una enfermedad genética (Hannhan y Weinberg, 2011).

Dada la naturaleza del cáncer, al ser una de las enfermedades que en la mayoría de los casos no es curable, éste se ha convertido en un gran reto de salud mundial; sin embargo, cuando se diagnostica de manera temprana puede ser tratado y curado de quien lo padezca. En algunos casos el trasplante del órgano afectado representa una alternativa, siempre y cuando la enfermedad no se haya propagado a otras partes del cuerpo. De tal manera que el tratamiento de esta enfermedad sigue siendo un enigma en el sector de las ciencias médicas. Una alternativa para enfrentar dicha enfermedad ha sido a través de la medicina genómica, debido a que el cáncer surge principalmente de las alteraciones genéticas las cuales ocasionan una proliferación celular sin control. La Sociedad Americana del Cáncer o ACS (por sus siglas en ingles American Cancer Society) lo explica como la muerte de células que se han desgastado o dañado, y nuevas células toman su lugar. El cáncer surge cuando las células dañadas comienzan a crecer sin control, dichas células siguen creciendo y propagándose a células normales o sanas (ACS, 2020).

A continuación, se describen las 3 etapas del desarrollo del cáncer:

1. **Etapa de iniciación:** Es cuando un agente genotóxico externo provoca un daño al ADN (también se conoce como mutación) en algunas ocasiones estos daños son heredados.
2. **Etapa de promoción:** En esta etapa otros agentes, ya sea externos o internos promueven la expansión clonal de células iniciadas.
3. **Etapa de progresión:** En este punto se empieza a desarrollar el tumor, y le confiere capacidades invasivas.

1.3. CHC (Carcinoma Hepatocelular)

El CHC se desarrolla en el 90% de los casos en pacientes con cirrosis; una importante etapa precancerosa que puede promover el desarrollo del cáncer. Sin embargo, en el 10% de los casos restantes el CHC se ha observado en pacientes sin cirrosis relacionado principalmente con el síndrome metabólico, infección por VHB o con alguna etiología no conocida (Fuentes Hernández, S., 2018).

La cirrosis puede ocasionar que el sistema inmunitario no funcione bien, lo que puede llevar a infecciones secundarias. La ascitis (líquido) en el abdomen, puede infectarse con bacterias que se encuentran normalmente en el intestino, y la cirrosis también puede ocasionar que los riñones no funcionen de la manera adecuada (National Institutes of health, 2002).

Sus principales factores de riesgo son la infección crónica por virus de hepatitis B y C, la hepatopatía crónica por alcohol. El incremento de los nuevos casos es notorio y se estima que continuará aumentando (Ochoa Carrillo, F., J., 2012). Así como lo menciona Ochoa Francisco, en la actualidad han aumentado los pacientes que padecen de dicha enfermedad. El riesgo es mayor en los pacientes con hepatitis viral. A nivel mundial, el virus de hepatitis B es el principal factor de riesgo. Al menos el 50% de los casos de CHC ocurren, en aquellos que fueron infectados por este virus (González Aguirre, A., J., 2013).

A pesar de los esfuerzos realizados para curar esta enfermedad, hasta la fecha sólo existen dos principales formas de tratarlo: la resección quirúrgica o curativo y la paliativa.

Xóchitl Celaya explica la manera de realizar estos tratamientos de la siguiente forma, la primera consiste en una cirugía para extirpar la parte afectada del hígado, siempre y cuando el tumor sea menor a tres centímetros y no se haya diseminado. La segunda se realiza cuando el hígado no responde al tratamiento anterior y se requiere de un trasplante (Celaya, X., 2014).

Para ejemplificar tales consideraciones, en la **Figura 1**, se representan, las posibles vías para decidir si se debe optar por un tratamiento curativo o paliativo.

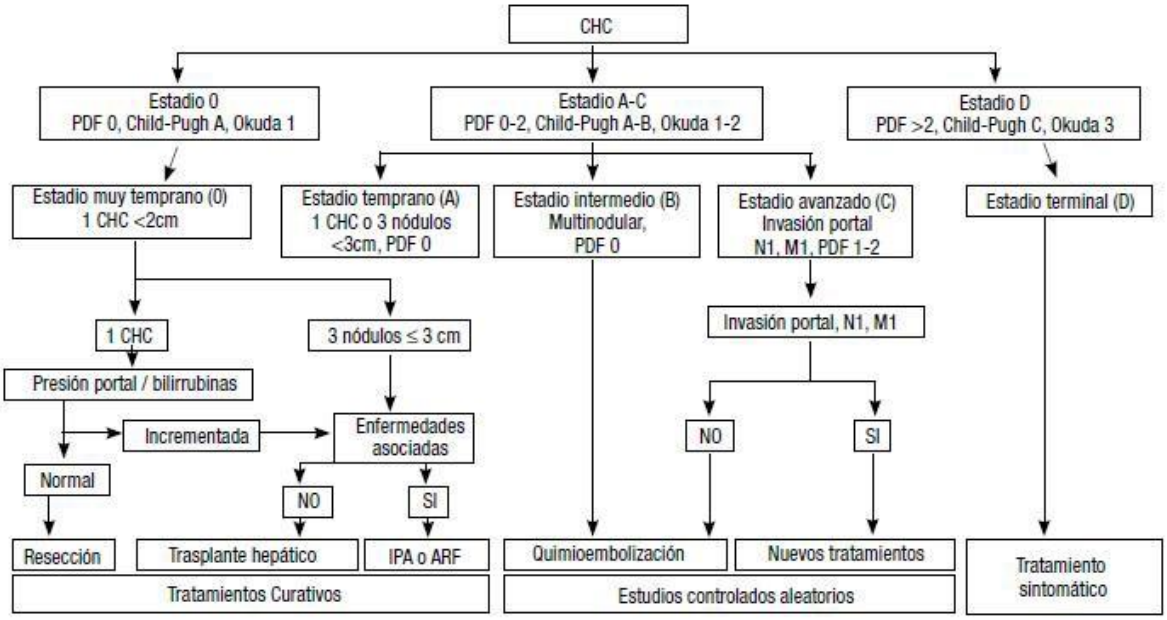

Figura 1. Esquema representativo de las vías de tratamiento del CHC (curativo y paliativo). Los tratamientos paliativos son en el caso de un estado terminal, o en un estado avanzado del mismo, dado que estos requieren de una mayor atención, los cuales serían los flujos del "Estado terminal" y "Estado avanzado" que se observan en el diagrama, mientras que el tratamiento curativo es en etapas tempranas. Recuperado de (Uribe Esquivel, M., 2010).

Capítulo 2

Expresión Génica

De acuerdo con (BIOINNOVA, 2016), la expresión génica se define como el proceso mediante el cual la información codificada en un gen se utiliza para dirigir el montaje de proteínas. Mientras que otras instituciones lo explican como la obtención de información de un gen para construir un producto funcional, dicho proceso se conoce como expresión génica (Khanacademy, 2020). Sin embargo, Benjamín Pierce explica que uno de los temas principales de la genética molecular es el dogma central que establece que la información genética fluye del ADN al ARN y de allí a las proteínas (Benjamín, A., 2010). Los anteriores artículos explican que la expresión génica es el proceso por el cual se genera una proteína, pero Benjamín menciona una tarea más en dicho proceso, el cual es la regulación génica. La cual explica de la siguiente manera: las bacterias llevan información genética para sintetizar muchas proteínas, pero sólo un subconjunto de esta información genética se expresa. Cuando el ambiente cambia, se expresan nuevos genes y se sintetizan las proteínas apropiadas para el nuevo ambiente (Benjamín, A., 2010). Por lo cual, se puede describir que dicho proceso consta de una adaptabilidad a un cambio. Dado que crean la proteína para una necesidad en específico, pero cuando ya no están en contacto con ese entorno o agente, los genes que se encargan de codificar esa proteína, se "apagan". Por lo cual el objetivo del presente trabajo es identificar los genes expresados diferencialmente durante la progresión del carcinoma hepatocelular.

Las características de nuestro cuerpo, la forma en cómo nos relacionamos con nuestro entorno y la manera en cómo nos adaptamos al cambio, son procesos evolutivos, cuya transformación empieza en los genes. Toda la información de un ser vivo se encuentra alojada en el ADN. Los diferentes organismos presentan mecanismos celulares, encargados de identificar dicha secuencia para así dar origen a las proteínas como se puede ver en la **Figura 2**. El estudio de los mecanismos de control de la expresión génica y la función de los genes está adquiriendo una gran importancia. Los esfuerzos de los genéticos y biólogos moleculares

se han centrado en la caracterización, secuenciación y estudio de nuevos genes, obteniendo avances significativos (Rafael, O., 2004).

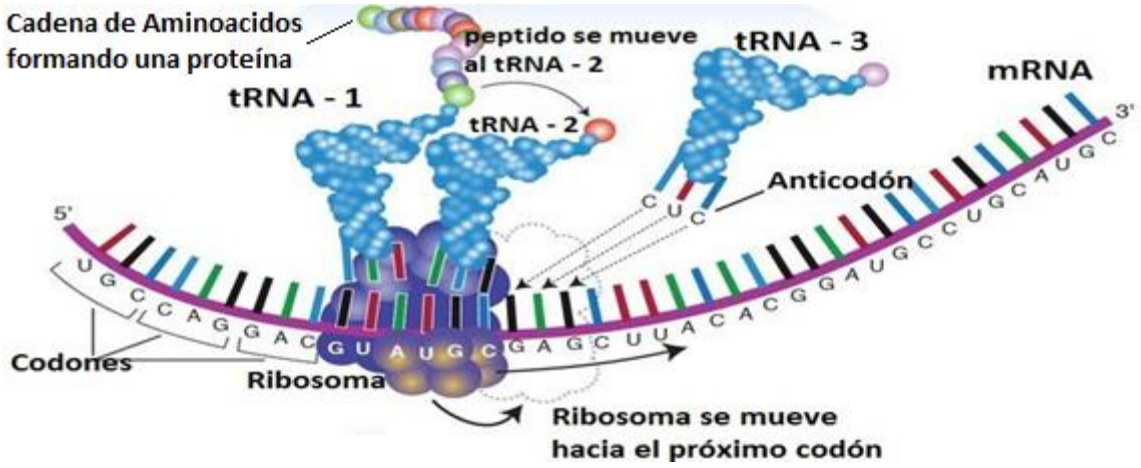

Figura 2. Visualización del proceso de expresión genética. En la imagen se puede apreciar la generación de aminoácidos, que forma la estructura primaria de una proteína. Recuperado de (National Humann Genome Research Institute, s.f.).

2.1. La célula

La célula es la unidad mínima de los seres vivos, ésta se encarga de realizar las funciones de nutrición, relación y reproducción. Adicionalmente, también son las encargadas de contener el material hereditario y transmitirlo (bioenciclopedia, s.f.).

Una de las partes de mayor interés para el análisis efectuado en el presente trabajo, es el núcleo de las células debido a que dentro de él se encuentra el nucléolo. Gran parte del ADN se encuentra en el nucléolo, de igual forma es donde se elabora el ARN (NIH, s.f.).

2.1.1 Clasificación de las células

La presencia o ausencia de un núcleo celular, define la clasificación de las células (Estela Raffino, M., 2020). Por lo tanto, se puede definir que solo existen 2 clasificaciones celulares, las cuales son las procariotas cuyo material genético se encuentra en el citoplasma y las eucariotas cuyas células poseen un núcleo que contiene su material genético (Estela Raffino, M., 2020). En la **Figura 3** se observa un ejemplo de célula eucarionte.

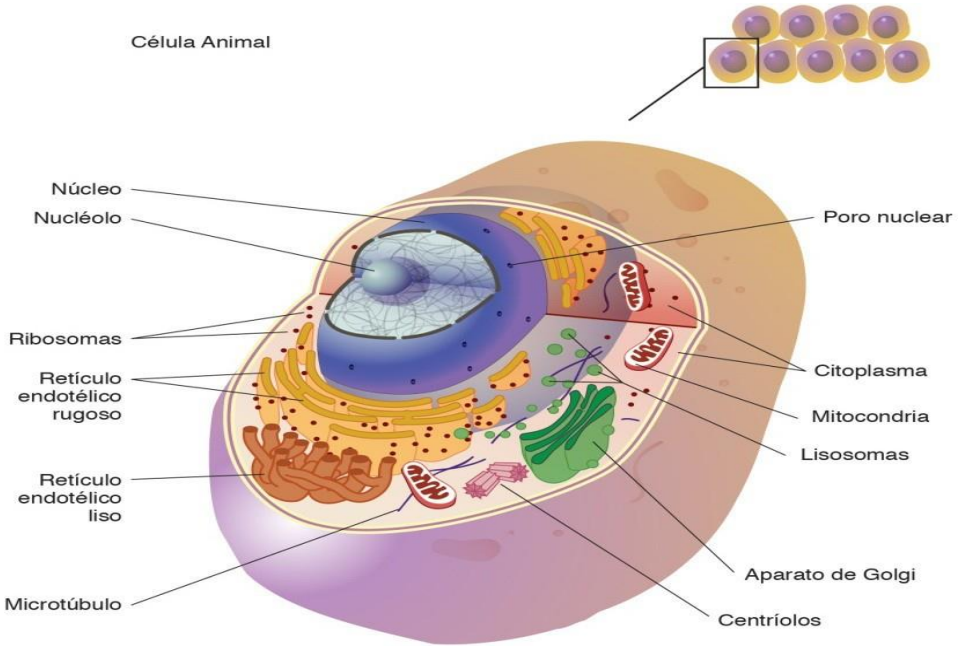

Figura 3. Representación estructural de una célula animal. Recuperado de (National Human Genome Research Institute, s.f.).

2.2. El Gen

Algunos autores definen al gen como un segmento o región de ADN, que codifica ciertas instrucciones, las cuales permiten a una célula producir un producto en específico (Chapman & Hall, 2012). Por otro lado, Dolores Corella lo define de la siguiente forma, unidad de herencia que ocupa una posición concreta en el genoma y que posee una estructura determinada (Dolores, C., 2017). En la presente tesis, se realizó un análisis sobre esos "segmentos" de ADN, para conocer si los ARNm que generan cambian su expresión. Esto se descubrió gracias a estudios realizados en la Drosophilia melanogaster (la mosca del vinagre), dichos estudios permitieron identificar muchos caracteres heredables concretos -genes-, demostrando que se encuentran localizados y alineados en el núcleo de las células, en unos orgánulos conocidos como cromosomas, cada gen se encuentra en una parte en específico de estos orgánulos (Ginés Morata, s.f.). En la sección anterior se comentó sobre la célula y su clasificación, las cuales son eucariotas y procariotas. Para el objetivo de esta tesis se centró en las células eucariotas, las cuales contienen un núcleo donde se encuentra el

material genético. El punto central que se retomó en el presente trabajo es conocer los genes mayormente alterados durante la progresión del CHC.

2.3. ADN (Ácido Desoxirribonucleico)

El instituto nacional del cáncer explica el ADN como moléculas del interior de las células que contienen información genética y la transmiten de una generación a otra. También conocido como ácido desoxirribonucleico (NIH, s.f.). Mientras que Cañedo Rubén explica que el ADN es el componente químico primario de los cromosomas de donde se forman los genes. Además de ser el elemento que controla todos los procesos celulares como la alimentación, la reproducción y la transmisión de caracteres entre padres a hijos (Cañedo, R., s.f.). Por lo cual, se puede concluir que en el ADN se encuentra toda la información que será transmitida de generación en generación. Dado que el ADN cuenta con una estructura codificada en 4 bases nitrogenadas: adenina (A), Guanina (G), citosina (C) y timina (T) (Jiménez, L., 2003). La estructura química de estas bases dicta que una molécula de adenina se une estrictamente con una de timina, mientras que una de citosina se une con una de guanina, dicha relación de moléculas es conocida como las reglas de Watson-Crick (Chapman & Hall, 2012). En la **Figura 4** se observa la manera en cómo estas bases nitrogenadas se unen de acuerdo con las reglas de Watson-Crick, para así formar una cadena de ADN.

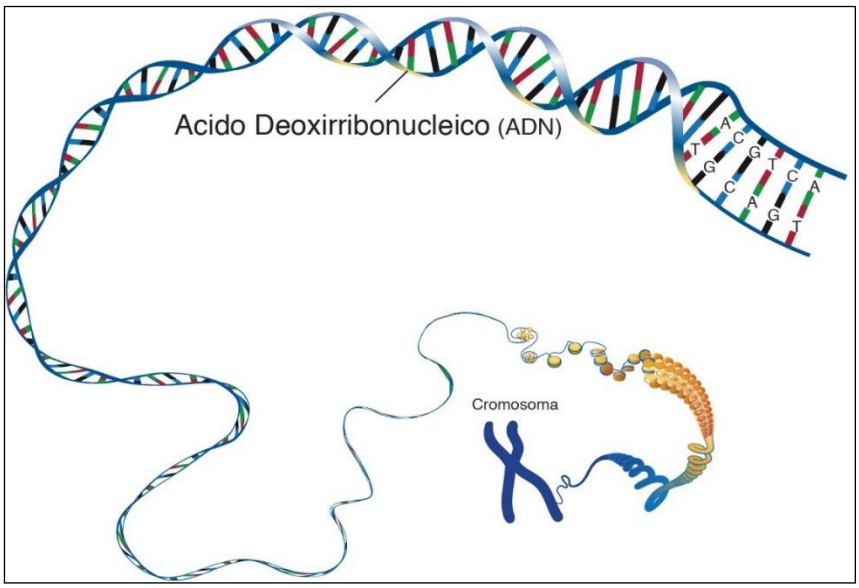

Figura 4. Representación de la construcción del ADN. La presente imagen muestra cómo se unen las bases para formar el ADN, que a su vez son parte de un cromosoma. Recuperado de (National Human Genome Research Institute, s.f.).

Algunas de las particularidades con las que cuenta el ADN además de tener el código que permite decodificar la información de los ARNm, el ADN contiene información para regular la expresión de los genes (Jiménez, L., 2003). Adicionalmente la Universidad Internacional de Valencia define otras funciones que tiene el ADN, las cuales son la capacidad de hacer copias de sí mismo para transferir la información genética de una célula a la célula hija o, dicho de otra forma, la replicación. La codificación de proteínas para cada célula interviene en el control del metabolismo celular, mediante la ayuda del ARNm y la síntesis de proteínas (Universidad Internacional de Valencia, 2018). Un proceso anormal que puede ocurrir en el ADN son las mutaciones, las cuales pueden ser un paso de la evolución de una especie.

2.3.1. Errores en el ADN

La secuencia de ADN de algún cromosoma puede tener errores. Como los siguientes:

1. **Mutaciones:** Una mutación por sustitución ocurre simplemente cuando un par de bases se convierten en un par diferente, alterando la secuencia (Soberón, X., 1999). Por ejemplo, se puede suponer una cadena como la siguiente “CG”, la cual sería una

cadena normal, la mutación ocurre cuando un elemento de la secuencia, como la citosina "C" se sustituye anormalmente por la guanina "G", la secuencia cambiará a esta forma "GG", lo cual codifica un gen con características diferentes al que normalmente se produce.

2. **Eliminaciones:** Consiste en eliminar un par de bases con relación a la original (Soberón, X., 1999), Tomando el mismo ejemplo que la mutación se puede suponer que la cadena "CG", pasa por esta situación y es eliminada la citosina "C", de tal manera, sólo quedaría la Guanina "G". Con lo cual no solo se tendría una secuencia incompleta, sino más pequeña de lo que debe de ser.
3. **Inserciones:** Ocurre cuando en la secuencia normal entra un par de bases más, que originalmente no se encontraban (Soberón, X., 1999). Por ejemplo, la secuencia "AT" pasa por esta situación añadiendo un par de citosinas generando el siguiente resultado "ATCC" con lo cual, ahora se tiene una cadena más larga de lo esperado.

Dolores Corella define estos procesos como mutaciones y lo explica de la siguiente manera: es el cambio de la secuencia original que, en algunos casos, tiene como consecuencia que el gen ya no codifique adecuadamente la proteína, por lo tanto, no será funcional (Soberón, X., 1999).

Estas son algunas de las situaciones por las cual pasa el ADN y su importancia para el ciclo de síntesis de proteínas, como lo menciona Luis Felipe Jiménez la información del ADN primero se transcribe; es decir, se copia selectivamente a moléculas de ARN y posteriormente la información de algunas de estas moléculas se traducen a proteínas (Jiménez, L., 2003). Lo mencionado será explicado en las siguientes secciones.

2.4. ARN (Ácido Ribonucleico)

Se han identificado fundamentalmente tres clases de ARN, el ARN mensajero (ARNm), el cual representa de 3 a 5% del ARN total celular, el ARN de transferencia (ARNt), con un porcentaje de 5 a 7% del ARN total celular, el ARN ribosomal (ARNr), que es el más abundante y cuyo porcentaje de ARN total celular oscila entre 85 a 90% (Jiménez, L.F., 2003). Cada una de estas moléculas es la encargada de realizar una tarea en específico para llevar a cabo la síntesis de una proteína, citando a Luis Felipe Jiménez, cada molécula de

ARNm contiene información para la secuencia de aminoácidos de una proteína, mientras que las moléculas de ARNr y de ARNt forman parte de la maquinaria celular que traduce la información de los ARNm a proteínas (Jiménez, L.F., 2003). Como se ha dicho la información genética se encuentra en el ARNm, los cuales serán fundamentales en el análisis del presente trabajo, debido a que un error en ese mensaje ocasionará que la proteína sea incorrecta. Adicionalmente existe una molécula más que se encarga de la síntesis del ARN la cual es conocida como ARN polimerasa. Jiménez Felipe cita a Losick y Chamberlon explicando que dicho ARN se mueve a lo largo del molde de ADN y sintetiza ARN, hasta que encuentra una secuencia específica de terminación. El proceso de ARN debe de partir de una hélice de ADN, una de las cadenas sirve de molde para formar la cadena sencilla de ARNm cuando se transcribe para expresar la información genética en una célula dada (García, F., s.f.). En la **Figura 5** se puede ver visualmente el proceso que sucede para generar una molécula de ARNm, para ejemplificar las consideraciones de los diferentes autores, sabemos que al ser el ARN una molécula muy parecida a una de las cadenas del ADN, la información dada por la secuencia de nucleótidos correspondiente a uno o varios genes se transfiere a una secuencia complementaria en el proceso de síntesis de ARN. Este proceso se denomina transcripción y está mediado por la enzima ARN polimerasa (Mateos, D., s.f.). Ahora bien, Daniel Mateos explica que el proceso para sintetizar un ARN es denominado transcripción (este término se explica en la sección 2.5); sin embargo, la ARN polimerasa es la encargada de unir nucleótido por nucleótido para crear el ARNm. Después sigue la labor del ARNr el cual menciona Luis Jiménez de la siguiente manera: el ARNr es informativo y conformacional, además de que existen múltiples pruebas génicas y bioquímicas que señalan que el ARNr es un elemento fundamental en el mecanismo de la traducción, ya que esta clase de ARN son elementos consecutivos y conformacionales de los ribosomas (Jiménez, L.F., 2003). Los ribosomas leen el mensaje transcripto y lo traducen (el proceso de traducción se explica en la sección 2.6) para así darle paso a la siguiente molécula, que es el ARNt, los cuales catalizan la formación de uniones peptídicas cuyo producto, luego de muchos ciclos, es una proteína (Rinflerch, A., 2008).

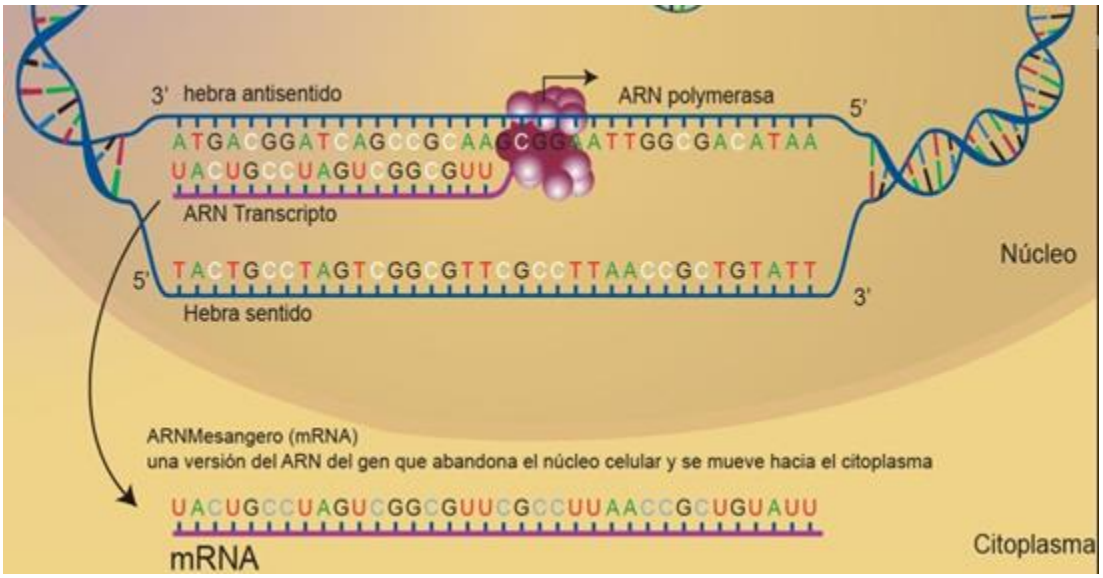

Figura 5. Ejemplo de la síntesis de un ARNm. Donde se observa un ARN polimerasa transcribiendo una hebra de ADN en un ARNm. Recuperado de (National Human Genome Research Institute, s.f.).

2.5. Transcripción

La transcripción es el proceso por el cual se transmite la información del ADN al ARN. Es llevado a cabo por la ARN polimerasa que utiliza como molde una de las dos hebras del ADN (hebra codificante) (Morales. I., 2017). Adicionalmente, dicho proceso comienza en unos sitios de la secuencia denominados promotores y está regulado con precisión a nivel celular (Mateos. D., s.f.). A manera de resumen, el proceso de transcripción significa convertir la información de una secuencia de ADN a una secuencia de ARN, por medio de la ARN polimerasa, lo cual puede ser observado en la **Figura 6**. Del mismo modo, Betancor en sus artículos sobre "Genética Bacteriana" explica lo dicho de la siguiente manera, durante la transcripción, las reglas del apareamiento de bases son aplicadas por la ARN polimerasa para sintetizar un producto complementario a una cadena del ADN usada como molde, que es el ARN (Betancor. L., s.f.). Finalmente, cuando la ARN polimerasa llega al final de un gen, este se desprende y la cadena de ARNm se encuentra lista para ser procesada por los ribosomas donde ocurre la acción de "traducción", en este punto es donde comenzará a crearse una proteína.

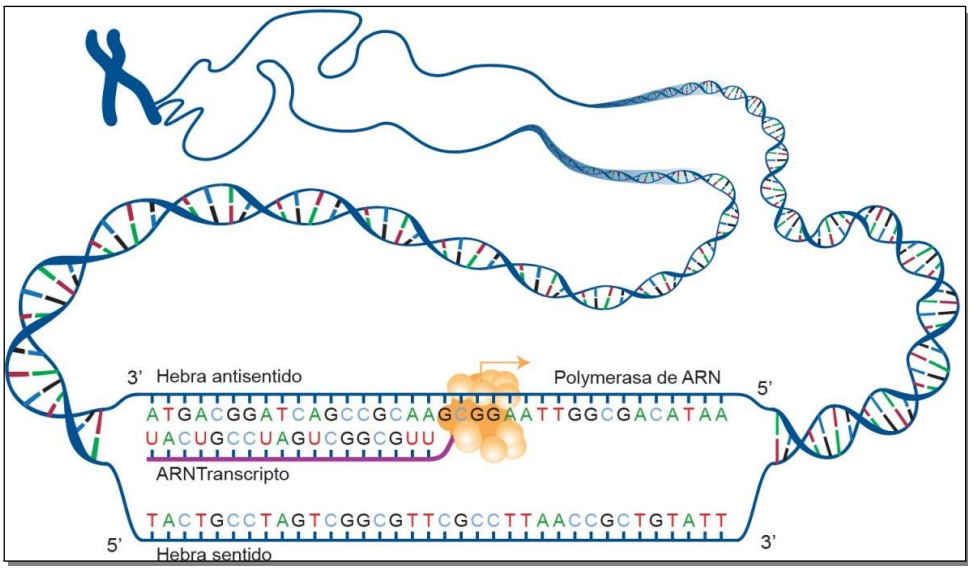

Figura 6. Proceso de transcripción de un gen. Recuperado de (National Human Genome Research Institute, s.f.).

2.6. Traducción

La traducción es el proceso por el cual la información genética que se ha transcrito del ADN a un ARN va a ser procesada para formar una proteína. Este proceso tiene lugar en los ribosomas (Morales. I., 2017). La traducción toma el resultado del proceso de la transcripción, pero en este participan fundamentalmente tres tipos de ARN: el ARN ribosomal (ARNr), el ARNm que es el portador de la información génica y los ARN de transferencia (ARNt), que son unos adaptadores específicos para cada tipo de aminoácido (Mateos. D., s.f.). Finalmente, Chapman explica que la información contenida de un ARNm se mapea de una secuencia de nucleótido de ARN a una secuencia de aminoácidos, formando así una proteína (Chapman & Hall, 2012). Para ilustrar lo mencionado por los diferentes autores, en la **Figura 7** se observa el proceso de traducción el cual parte de una cadena de ARNm como se ha dicho, y en lugar de la ARN polimerasa se encuentra un ribosoma traduciendo codón por codón, para que después empiecen los ARNt a mandar la información a los aminoácidos, de tal manera que se creará una proteína.

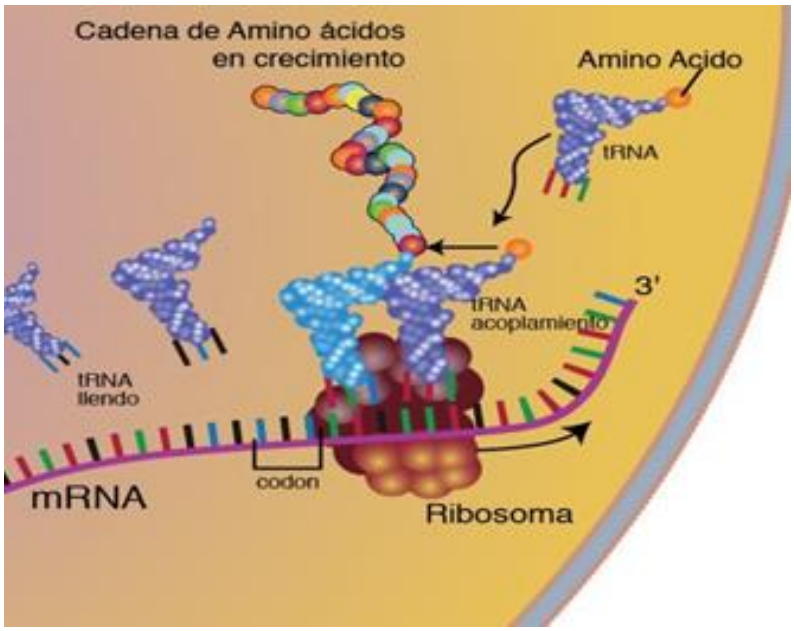

Figura 7. Proceso de traducción. Donde se puede observar el ribosoma recorriendo la cadena de ARNm, para formar una proteína por medio de los ARNt. Recuperado de (National Human Genome Research Institute, s.f.).

2.7. Las Proteínas

Las proteínas son moléculas informacionales, pero a diferencia del ADN, que es la molécula en donde reside la información génica, en las proteínas reside la información funcional de la célula (Mateos. D., s.f.). Estos procesos de síntesis de proteínas sólo ocurren cuando alguna célula lo necesita, en un nivel fundamental, las moléculas de proteínas se asocian al material genético y convierten condiciones ambientales en señales para activar o desactivar genes específicos (Jiménez. L., 2003). Con respecto al análisis realizado en la presente tesis, se tiene un interés particular en aquellos genes expresados, en otras palabras, los que se encuentran activos, así como lo menciona Luis Jiménez esto se encuentra ligado a las funciones de las proteínas ya que puede suceder algún fallo durante el proceso de síntesis de una proteína, lo cual ocasiona que un gen no se active en un proceso necesario. La importancia de las proteínas radica en que son macromoléculas que desempeñan el mayor número de funciones en las células de los seres vivos. Forman parte de la estructura básica de tejido, durante todos los procesos de crecimiento y desarrollo, crean, reparan y mantienen los tejidos corporales; además de desempeñar funciones metabólicas y reguladoras

(González. L., 2007). Es necesario recalcar que si alguna función de las proteínas se ve alterada o afectada por una mala síntesis puede ocasionar fallas en sus funcionamientos, como bien lo menciona Laura González, las proteínas realizan funciones metabólicas y reguladoras tales como la asimilación de nutrientes, transporte de oxígeno, eliminación de materiales tóxicos, entre otras. Son los elementos básicos del cuerpo, esenciales en todo el metabolismo.

Capítulo 3

Microarreglos

Con el objetivo de entender el comportamiento del genoma del ser humano se han implementado nuevas estrategias de biología molecular, para evaluar de manera integral el comportamiento de los genes a través de análisis de herramientas tecnológicas como lo son los microarreglos.

Un microarreglo de ADN sirve para determinar la expresión genética completa de un tejido en un momento determinado. A esta evaluación se le denomina transcriptoma (Vallin Plous, C., 2007). Por otro lado, Jorge Ramírez define un microarreglo de la siguiente manera: conjunto ordenado de genes en una pequeña superficie de 10 000 cm^2, con lo cual se pueden analizar grandes cantidades de información en un solo experimento (Ramírez, J., 2003). Adicionalmente Luis Benítez menciona que los microarreglos permiten usar la información de secuencias del genoma para medir en forma paralela y cuantitativa la expresión de los genes, por medio de los ARN mensajeros (Benítez. L., 2004). Así, la tecnología de microarreglos ha sido de gran ayuda para las ciencias genómicas, las cuales están conformadas por la genómica estructural, funcional, individual y comparativa.

Para el presente trabajo se aplicó la genómica funcional, la cual se define como: la determinación del número de genes, transcritos o proteínas (ADN, ARN mensajeros y proteínas) presentes y expresados en una célula-tejido específico, en un momento fisiológico determinado: Transcriptoma y Proteoma, respectivamente (Salcedo. M., 2003). Los microarreglos se integran en un chip de ADN, en los que se encuentra una placa donde se ubican inmovilizados los genes. Este material se encuentra distribuido de manera similar a una matriz, donde cada posición contendrá una hebra de ADN, comúnmente esa posición es la que tiene el material genético y se conoce como “sondas”. Cada posición de esa matriz se lee, y se cuantifican las intensidades de fluorescencia, de cada fluorocromo en un escáner especial. Desde el punto de vista de Salcedo se pueden observar diferentes colores que tienen

significados diferentes; rojo para una sobre-expresión, verde para sub-expresión y amarillo cuando no hay cambio en la expresión, entre la muestra normal y la muestra problema (Salcedo, M., 2003). Cabe mencionar que dicho significado de los colores dependerá de la manera en cómo se quieran presentar los resultados. Salcedo usó esos colores para representar sus resultados, en otros artículos se utiliza el color negro para no mostrar cambio entre las muestras. En el presente trabajo se utilizó de la siguiente manera: rojo para indicar una sobre-expresión, verde para indicar una sub-expresión y negro para indicar que no tuvo cambio alguno.

En la **Figura 8** se muestra el proceso de análisis de un microarreglo donde se compara el tejido de estudio contra el tejido control. Primero se separa el ARNm de ambos tejidos, y obtener el ADNc (ADN complementario) de cada tejido. Las moléculas de ADNc se deben de marcar con un compuesto fluorescente de cianina (Cy3 generalmente es un color cercano a rojo y Cy5 que es un color entre naranja y amarillo), diferente por cada tejido para después mezclar ambos ADNc y se da inició a la hibridación (unirse) a un ADNc en específico de la placa (Vallin, C., 2007).

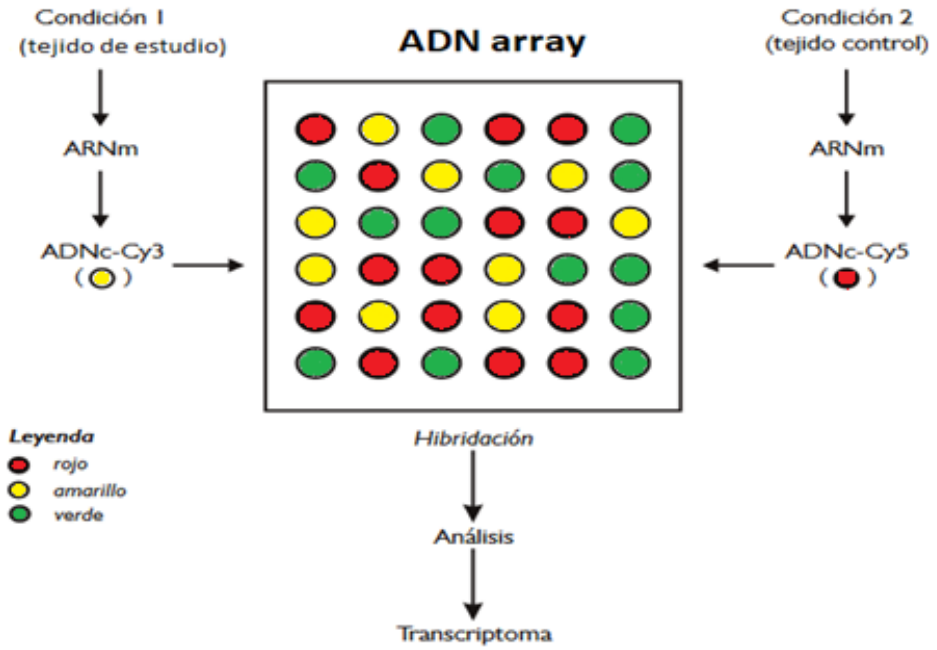

Figura 8. Metodología de un microarreglo de ADN. Se muestra el proceso de análisis de 2 tejidos (el de interés contra control), la muestra de estudio tendrá el pigmento Cy3 (tendrá un color amarillo), mientras que la muestra control tendrá un pigmento Cy5 (tendrá un color rojo). En la imagen se pueden observar círculos de color verde, que representan los genes que hibridaron, después de dicho proceso se puede proceder al análisis para después conocer el transcriptoma, que determinará la expresión genética del tejido en un momento determinado. Recuperado de (Vallin. C., 2007).

Las aplicaciones de los microarreglos se utilizan comúnmente para:

1. **Monitorear la expresión génica**
2. **Detección de mutaciones y polimorfismo**
3. **Secuenciación**
4. **Diagnóstico clínico-detección de microorganismos**
5. **Tamizaje en toxicología**
6. **Seguimiento de terapias**

A pesar de que existen varias tecnologías para la producción de microarreglos las más utilizadas por los investigadores son ADNc, y Affymetrix (Miranda. J., Bringas. R., 2008). Para el presente trabajo, se utilizaron microarreglos de Affymetrix, dado que las sondas en los arreglos son más homogéneas y menos variables que en los arreglos de ADNc (Miranda. J., Bringas. R., 2008).

El esquema del proceso de análisis de un microarreglo se representa en la **Figura 9**.

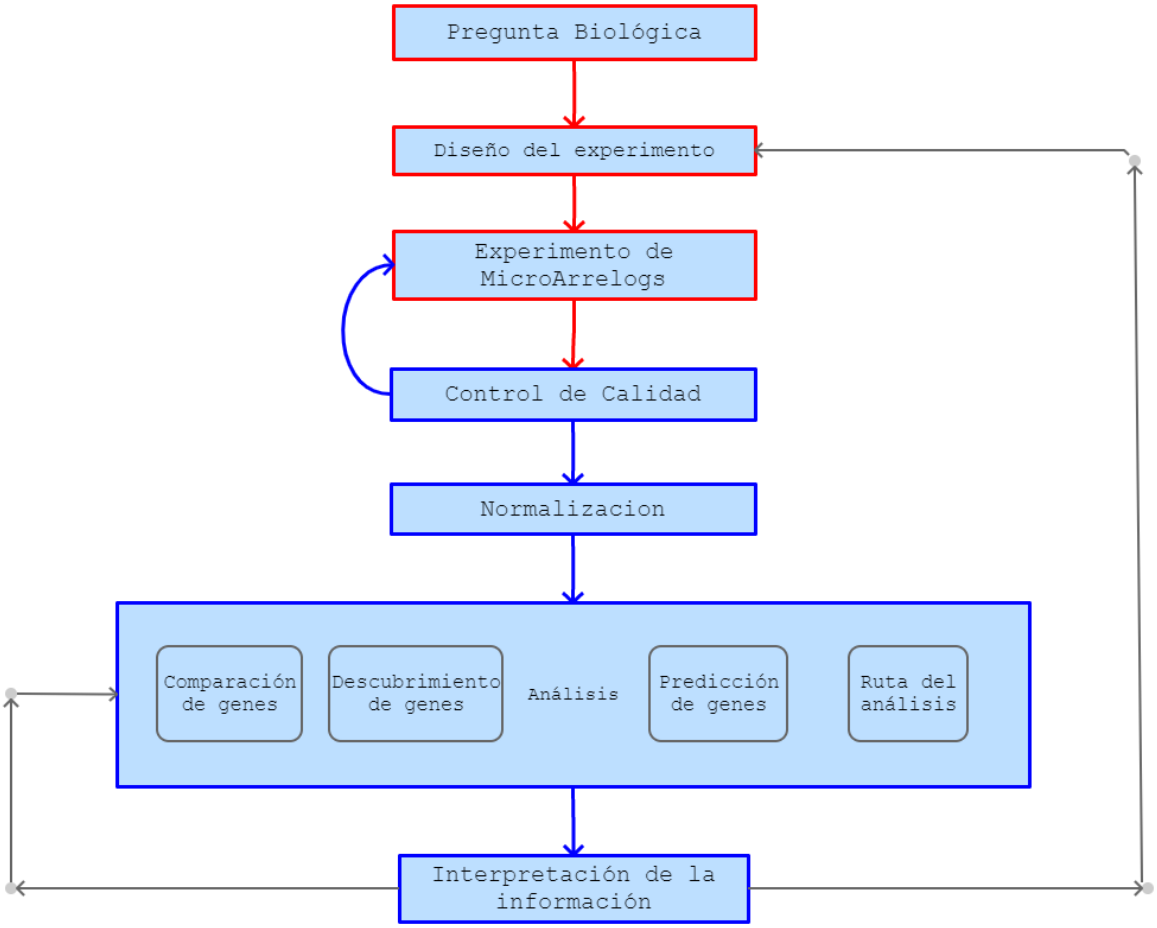

Figura 9. Esquema general para la realización de un experimento de microarreglos. En rojo se marcan los pasos de la primera etapa requerida, que son la definición y realización del experimento. Mientras que en azul se marca la fase de análisis. Recuperado de (Gonzalo, R. Sánchez. A., 2018).

3.1. Tipos de microarreglos

Dependiendo de cómo se hibridan las muestras, hay dos tipos de microarreglos: microarreglos de un canal o un color (en los que se pone una única muestra por arreglo) y microarreglos de dos colores o dos canales (en los que se ponen dos muestras etiquetadas con distintas moléculas fluorescentes) (Acevedo. R., Álvarez. E., et al, 2007).

En los microarreglos de dos canales el objetivo es comparar la cantidad relativa en las dos muestras estudiadas. Por ejemplo, una muestra antes del tratamiento y otra después, una muestra normal y una de tumor, etc. El valor resultante para cada sonda suele ser un valor relativo (normalmente llamado valor relativo de cambio o fold changue) que representa en cuál de las dos muestras hay más cantidad de la secuencia dada. Su principal ventaja es que dan un resultado claro sin necesidad de muchas muestras. Sin embargo, resulta difícil incluir más controles o hacer otras comparaciones posteriores, por lo que sólo son recomendables en estudios en los que haya un claro punto de referencia (p.ej. muestra de tejido sano y una tumoral de un mismo paciente) (Salcedo, M., 2003).

Por otro lado, los microarreglos de un canal siguen un enfoque distinto: medir o estimar la cantidad absoluta total en cada muestra para tener más libertad para realizar distintas comparaciones entre ellas. De este modo, el valor resultante se aproxima a una medida absoluta de la concentración de señal en las muestras. Sin embargo, para realizar la comparación entre múltiples arreglos son necesarias normalizaciones robustas y por ello normalmente también preprocesamientos más complejos.

En cuanto al tipo de fabricación de los microarreglos, también hay dos tipos principales: los microarreglos *spootted* y los sintetizados *in situ*.

Los microarreglos *spotted* se producen con un robot que deposita las secuencias previamente sintetizadas sobre una placa de cristal. Suelen tener menos reproducibilidad y son de dos canales. Esta tecnología es una de las primeras en aparecer y ha ido cayendo en desuso. Sin embargo, son más flexibles (se pueden hacer arreglos personalizados más fácilmente) y pueden tener secuencias de nucleótidos más largas. Los microarreglos sintetizados *in situ* son arreglos producidos sintetizando los oligonucleótidos directamente sobre el soporte. Aunque

hoy en día hay distintas tecnologías para producirlos, la inicial desarrollada por la compañía Affymetrix se basa en fotolitografía, el sistema utilizado para crear circuitos integrados (Villan. C., 2007). Esta técnica consiste en ir añadiendo los nucleótidos uno a uno, utilizando máscaras que tapan/destapan las ubicaciones de cada sonda para que el nucleótido sólo se deposite en aquellas que corresponde.

Estos arreglos son mucho más precisos y permiten mayor densidad de sondas. Sin embargo, la longitud de la sonda está más limitada (25-100 bases, según la tecnología), y en el caso de querer un arreglo personalizado se tiene que contactar a la compañía.

3.2. Extracción de características

En cualquier estudio con tecnologías ómicas, la obtención de metadatos sin procesar es sólo uno de los pasos iniciales. Los datos obtenidos directamente del escáner o de la máquina de secuenciación tienen que ser preprocesados, normalizados y convertidos en valores representativos de los parámetros biomoleculares que se están midiendo (Ramírez. J. et al., 2003). Por ejemplo, un valor que represente la concentración de un transcrito en una muestra. Una vez los datos hayan sido preprocesados, estarán listos para analizar y responder las preguntas del estudio. La elección de los métodos de preprocesamiento y análisis depende tanto de la plataforma como del tipo de dato.

En el caso de los microarreglos, los datos crudos son los valores de fluorescencia en cada celda, que a su vez representan el valor de hibridación de una sonda. Sin embargo, aparte de la señal de fluorescencia propia de la hibridación de las sondas, también suele haber una ligera fluorescencia remanente en todo el arreglo. Esta fluorescencia puede ser producida por las sondas que han hibridado a pesar de no ser perfectamente complementarias, y por lo tanto hay que corregirla. Este paso se denomina corrección de la señal de fondo.

Posteriormente, hay que calcular la señal correspondiente a cada sonda. En el caso de los arreglos de dos colores, esto se hace calculando la señal relativa entre los colores (por ejemplo, el valor de cambio – "*fold change*"). En el caso de los microarreglos de un canal, hay que calcular un valor de señal para cada conjunto de sondas que represente un gen o unidad de medida (resumen – "*sumarizacion*"). La agrupación de las sondas se suele definir

en el archivo de descripción del chip (CDF, archivo de definición del chip - "*chip definition file*") que indica qué conjunto de sondas (*probe-set*) corresponden a un mismo gen. Sin embargo, el valor obtenido para cada sonda no se puede utilizar directamente. Hay muchos factores que pueden haber alterado la medida: desde la cantidad de muestra con la que se ha hibridado el arreglo, manipulación de la muestra, efectos del escaneado, ubicación de la sonda en el arreglo, proporción de nucleótidos en la sonda, etc.

Puesto que muchos de estos efectos son sistemáticos, se pueden compensar con un proceso de normalización dentro de cada arreglo y entre arreglos para pasar todos los datos a escalas comparables.

La **Figura 10** es un ejemplo de una matriz de expresión, en donde los genes se encuentran en la retícula.

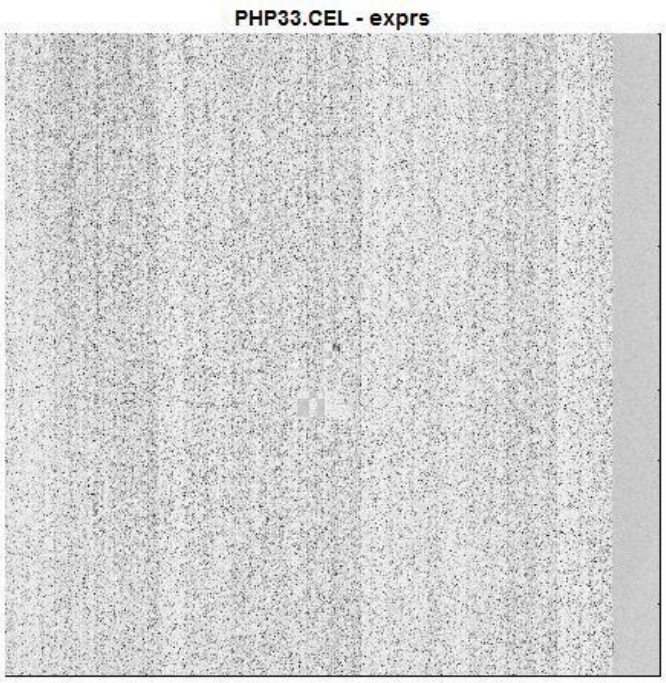

Figura 10. Imagen resultante del proceso de hibridación de un microarreglo de ARNm. Donde los puntos negros son aquellos genes que hibridaron y los blancos aquellos que no lo hicieron.

3.3. Control de calidad

Una vez realizado el análisis de imagen y calculadas las intensidades de cada señal, se hacen imprescindibles las transformaciones de estos datos primarios, debido a que las intensidades de cada señal, además de reflejar los niveles de ARNm, pueden contener sesgos asociados a la forma de impresión en el chip, a la eficiencia del marcaje de las muestras y a otras fuentes de variabilidad. Algunas de estas transformaciones son el filtrado y normalización de los datos y la aplicación de preprocesamientos. Como primer paso se recomienda el filtrado de los datos primarios para eliminar los valores que probablemente sean el producto de errores. Se utilizan criterios como el cálculo del coeficiente de variación (CV) para cada gen, que se calcula como la desviación estándar (SD) entre la media de un conjunto de razones de expresión de múltiples señales del mismo gen: CV = SD/Media, y se eliminan los genes con CV mayores que un determinado umbral.

Es decir, tomando los siguientes valores de expresión: (3108, 3698, 1976, 4092, 3112, 4392, 2415, 3499, 4518, 3718, 2506, 4129, 2980, 3031, 2790), se calcula el coeficiente de variación para conocer el porcentaje de variabilidad. Calculando la desviación estándar se obtiene el valor de 757.2436 con una media de 3335.7333, por lo tanto, al sustituir los valores se obtiene CV = 757.2436/3335.7333 = 0.2270, lo cual al hablar de porcentajes se multiplica por 100, teniendo como resultado CV = 22.70% de variabilidad entre los datos, lo cual quiere decir que todos los genes del primer conjunto menor a este umbral son menos expresados. Haciendo una primera inspección visual del conjunto de datos se puede observar que el valor menor es 1976, por lo tanto, este gen debe de ser descartado al tener un valor bajo.

Otro criterio consiste en eliminar los valores de intensidad mayores que un umbral, que pudieran ser valores de señales sobresaturadas. También se sugiere realizar una inspección visual de las imágenes para detectar efectos en los arreglos y eliminar las intensidades correspondientes antes de proceder a la normalización.

El control de calidad se inicia con la inspección visual de los datos, para verificar qué áreas no fueron hibridadas. Entre las técnicas existentes para realizar este proceso se encuentran métodos estadísticos de análisis de datos, los cuales son utilizados para realizar un diagnóstico preliminar de la información, con el objetivo de identificar matrices

problemáticas, ciertos falsos positivos y datos que no tendrán relevancia en el experimento. Entre los gráficos utilizados se tienen análisis de componentes principales (ACP – *análisis component principal*) y diagrama de cajas (*boxplots*).

El análisis de componentes principales consiste en: reducir la dimensionalidad de un conjunto de datos de una gran cantidad de variables interrelacionadas, al tiempo que conserva la mayor cantidad posible de la variación presente en el conjunto de datos (Jollife, I., 2002).

El análisis de componentes principales es una técnica muy utilizada en campos como reconocimiento de imágenes y encontrar patrones en datos de grandes dimensiones. Otro uso dado para ACP es la comprensión de datos, ya que puede reducir la dimensión de vectores de datos sin pérdidas de información, esta técnica es muy utilizada en el análisis de datos. Del total de factores o variables que se representan por las dimensiones se eligen los que recogen el porcentaje de variabilidad que se considere suficiente, denominados componentes principales. En tanto que como método de *clustering*, ACP busca las características principales que permitan separar en grupos a los datos, a través de la reducción de dimensiones. También puede utilizarse como un método previo a la aplicación de otro algoritmo de *clustering* al simplificar el conjunto de datos. Algoritmo de análisis de componentes principales:

Datos de entrada: Conjunto de M vectores de N datos.

1. Obtener la matriz de covarianza.
2. Obtener los eigenvectores y eigenvalores de la matriz de covarianza.
3. Ordenar la matriz de eigenvectores de acuerdo con los eigenvalores.
4. Seleccionar los primeros N componentes y crear los vectores de N-dimensión.

Para que se pueda realizar el ACP, es necesario que las variables presenten factores comunes. Es decir, que estén muy correlacionadas entre sí. Los coeficientes de la matriz de correlaciones deben ser grandes en valor absoluto. Como este método permite reducir el número de dimensiones de una matriz es utilizado para graficar vectores de grandes dimensiones utilizando 2 ó 3 dimensiones de los componentes más importantes. En la **Figura 11** se puede observar un ejemplo del gráfico de ACP.

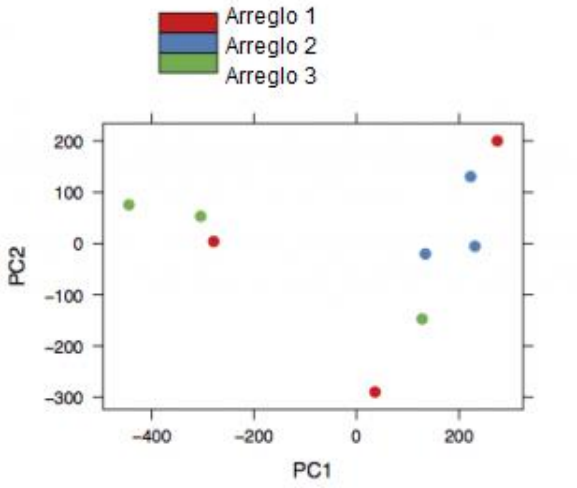

Figura 11. Gráfico de ACP. Se expresan los valores en puntos de diferente color, ya sea dispersos o agrupados dependiendo de su varianza. Recuperado de (EMBL-EBI, s.f.).

El diagrama de caja representa los tres cuartiles, los valores mínimos y máximos de los datos, sobre un rectángulo, alineado horizontal o verticalmente y describe varias características al mismo tiempo, tales como la dispersión y simetría. Es útil para visualizar aspectos de la distribución de diferentes grupos o categorías en una o más series de datos cuantitativos.

En la **Figura 12** se observa un ejemplo del gráfico de cajas.

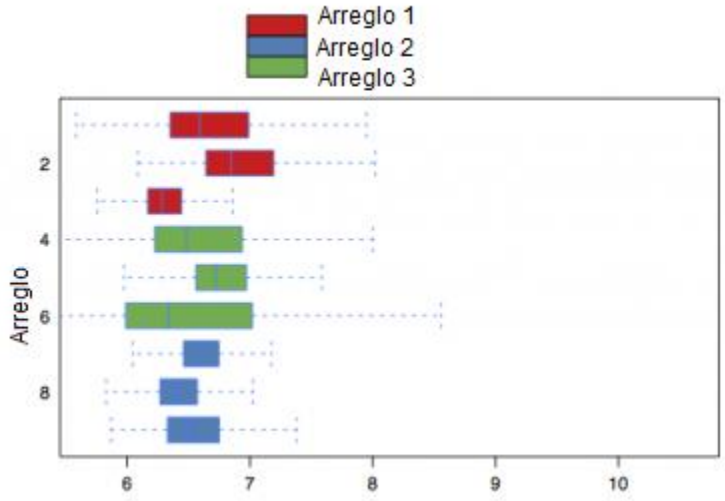

Figura 12. Diagramas de cajas: Cada color representa un grupo de genes y la cantidad de cajas indica el número de muestras, cada caja cuenta con una línea negra en el centro que es la varianza de esa muestra, mientras que el tamaño de la caja es la limitación de los valores donde se encuentra la varianza. Recuperado (EMBL-EBI, s.f.).

3.4. Normalización

El proceso de normalización es la primera transformación aplicada a los datos de expresión y un paso esencial antes de pasar a su análisis. Esta transformación se realiza con el objetivo de minimizar las variaciones sistemáticas en la cuantificación de los niveles de hibridación de las muestras de ARNm, de manera que las diferencias biológicas se puedan distinguir fácilmente, y hacer comparables los niveles de expresión entre los chips. La normalización por lo general se aplica en el interior de cada chip o entre múltiples chips, y para ello se seleccionan los métodos y las variables o regiones del arreglo (conjunto de genes) que servirán para la estandarización de los datos.

Entre las técnicas utilizadas para la normalización de datos y la que fue utilizada en el presente trabajo es RMA (*Robust Multiarray Average*) el cual es un método no paramétrico muy robusto de gran eficacia y reproducibilidad, su uso generalizado se ha convertido prácticamente en un estándar de factor para los microarreglos de expresión de Affymetrix.

Una de las novedades que introdujo RMA, fue el no utilizar las sondas de mismatch (MM) para la corrección de la señal de fondo. Los arreglos de Affymetrix incluyen tanto las sondas para medir las secuencias de interés (perfect match PM), como sondas mismatch (diseñadas como controles) en las que se ha cambiado la base central. El objetivo de las sondas mismatch es obtener información sobre la capacidad de hibridación inespecífica. Los métodos existentes antes de que apareciese RMA calculaban la señal "real" usando la diferencia (Pm-MM) o la proporción entre ellas (PM/MM). RMA ignora las sondas MM y calcula la señal de fondo como una distribución a nivel global del arreglo. Una vez se ha descontado la señal de fondo de la señal observada, se tiene el valor de cada sonda (Miranda. J., 2008).

RMA usa entonces la normalización por cuantiles para hacer que la distribución de la intensidad de todos los arreglos sea la misma, y el logaritmo para facilitar el uso de métodos paramétricos. Finalmente, *sumariza* los valores normalizados utilizando un modelo lineal que asume que el efecto sonda (*probe effect*) es constante para cada sonda entre todos los arreglos. Tras pasar por este proceso, se obtiene un valor de expresión para cada gen o sonda (*probeset*) en cada microarreglo (Chapman & Hall, 2012).

Este valor permite comparar la señal de los *probesets* o genes entre los microarreglos que han sido normalizados juntos.

La normalización consta de 3 pasos: **Corrección del fondo:** Estimar y eliminar la intensidad del ruido de fondo, **Normalización global o local:** Asegurar que la mayoría de las sondas varíen igual y **Resumen:** Conversión de sondas o conjuntos de sondas a transcritos o genes (Morales General, I., 2017).

Entre los métodos de normalización existentes se tienen:

1. Global o lineal (aplicable a chips del tipo ADNc y Affymetrix): el factor de normalización es el mismo para todos los genes del chip.
2. Dependiente de la intensidad (aplicable a chips del tipo ADNc y Affymetrix): el factor de normalización depende de la intensidad de cada señal.
3. Dependiente de la localización (aplicable a chips del tipo ADNc): el factor de normalización depende de la localización de la señal en el chip.

3.5. Análisis de expresión diferencial

Estadística Inferencial

En este punto se busca determinar qué genes están expresados diferencialmente (DEGs) y si dicha expresión es significativa. Para tal fin se emplea una prueba estadística clásica, como la *t* de Student y se calcula para un gen, el valor observado (valor de la prueba que compara los grupos), y se repite la misma prueba en un gran número de veces, de forma que la asignación del grupo se cambie al azar. Al realizar este procedimiento se simula una situación, en el cual no existen diferencias, ya que la asignación de un grupo u otro a cada muestra es aleatoria. Por último, se calcula el valor de *p*, a partir del percentil que ocupe el valor observado en la distribución de los valores obtenidos por permutación. Este proceso se realiza por cada gen (Morales General, I., 2017).

La *t* de Student mide la expresión diferencial teniendo en cuenta la ratio entre señal (media) y ruido (varianza) en el experimento. A partir de este estadístico se puede calcular un *p-valor* que comparar con el nivel de significatividad para aceptar o rechazar la hipótesis nula en microarreglos.

Con el gráfico de volcán (*volcano plot*), se representa los genes según su expresión diferencial y significatividad estadística. En la **Figura 13** se puede observar un ejemplo de lo mencionado.

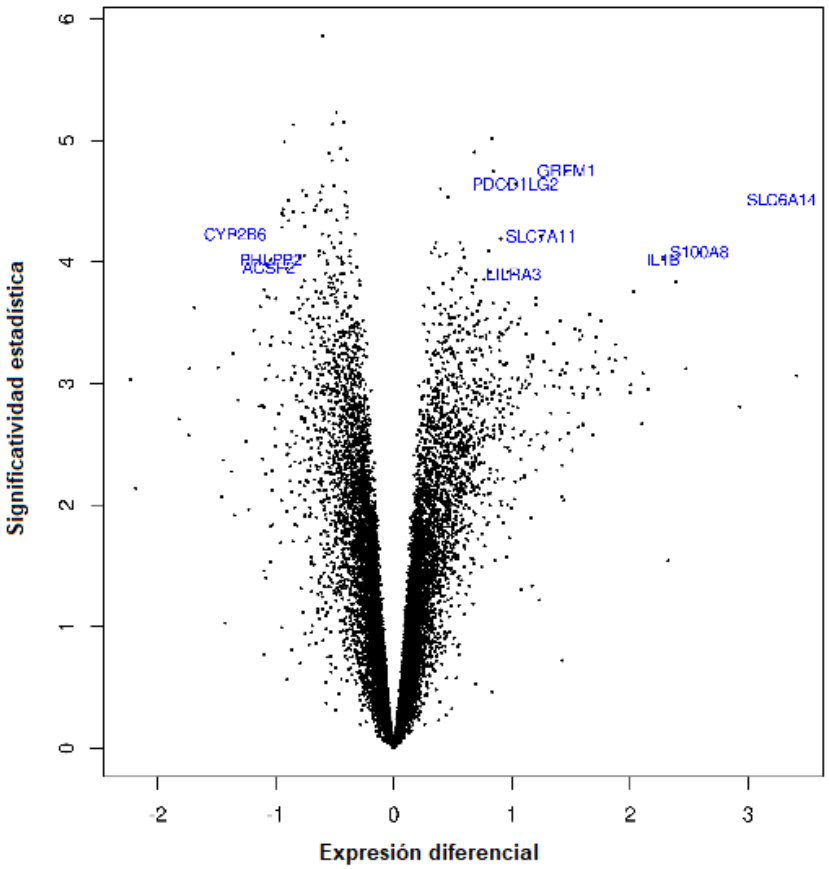

Figura 13. Gráfico de volcán. Utilizado para hacer un resumen del resultado de análisis de expresión diferencial. Recuperado de (Klaus Bernd, 2019).

Estadística Descriptiva

Se busca determinar grupos de genes que presentan patrones similares a través de Análisis no-supervisado, es decir, sin tener información de la estructura de los datos en el microarreglo.

Se emplean métodos de agrupamiento o clustering a datos de expresión para construir grupos de genes o muestras con perfiles de expresión similares utilizando una medida de distancia (la empleada en el análisis realizado fue la distancia euclidiana).

Los métodos de agrupamiento por lo general no requieren la información del grupo, clase o condición experimental a que pertenece cada muestra que se incluye en el análisis, sino que por el contrario pueden sugerir un nuevo agrupamiento de las muestras basado en el grado

de similitud entre los perfiles de expresión de los genes en estudio. Estos métodos, aplicados a datos de expresión, sirven para identificar grupos de genes coexpresados y patrones en los perfiles de expresión de las muestras, sin la necesidad de clases predefinidas que supervisen el análisis.

El método de agrupamiento más empleado en datos de microarreglos es el agrupamiento jerárquico. Este método no supervisado deriva una serie de particiones de los datos; en este caso, cada dato será el perfil de expresión de una muestra o gen. Existen varios tipos de métodos de agrupamiento jerárquicos, tales como el aglomerativo y el divisivo, los divisivos funcionan mejor para dividir los datos en pocos grupos de varios elementos. El resultado de estos métodos es una estructura de árbol o dendograma.

Para la construcción del árbol se tienen las técnicas:

Aglomerativa: se considera cada elemento por separado y se van uniendo los que tienen distancias más pequeñas.

Divisiva: técnica inversa, se considera todo el conjunto de elementos y se van separando los que tienen distancias más grandes.

La **Figura 14**, se presentan la construcción del árbol antes mencionados.

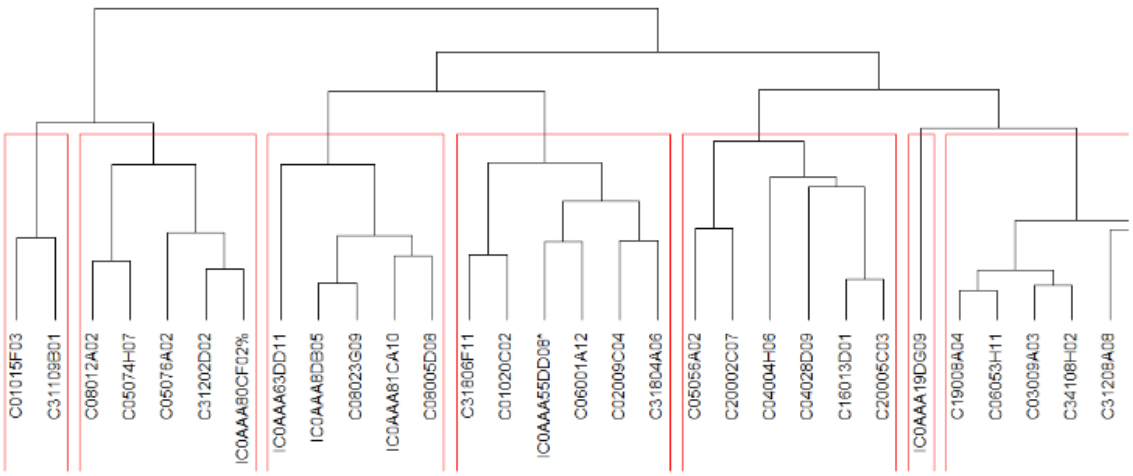

Figura 14. Estructura de árbol o dendograma. Ejemplo de comparación de genes expresados diferencialmente. Recuperado de (Papa Lucia., 2015)

Uno de los métodos comúnmente utilizado es el mapa de calor el cual consiste en una cuadrícula de colores, donde cada fila representa el comportamiento de un gen, mientras que cada columna representa la muestra de interés. Los gradientes de intensidad de los colores representan los cambios de nivel de expresión génica. En la **Figura 15** se puede ver un ejemplo de un mapa de calor.

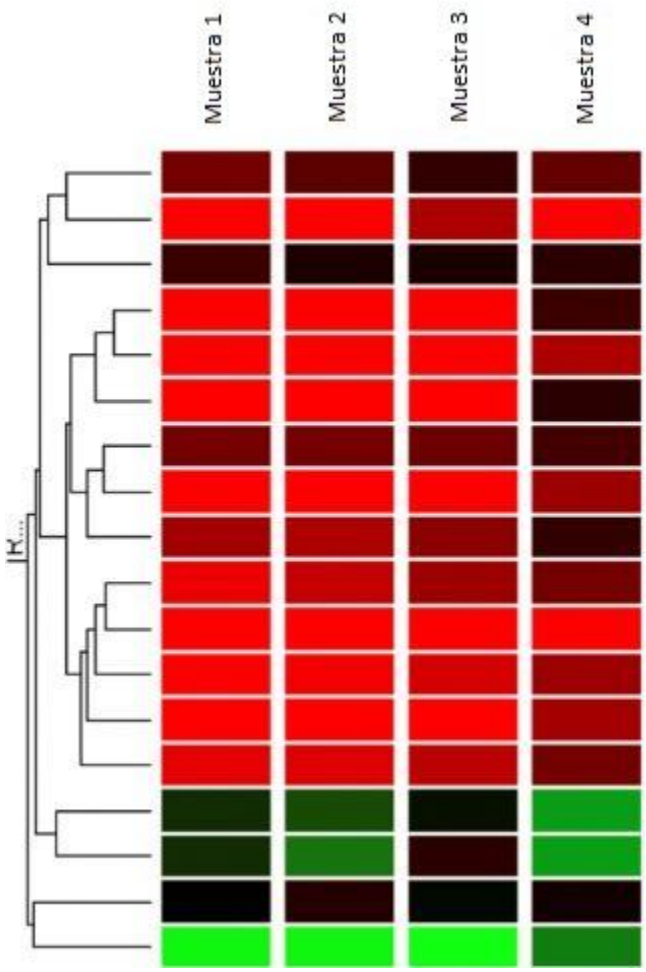

Figura 15. Análisis de microarreglos representado mediante un mapa de calor. El dendograma representa la relación del comportamiento de un grupo de genes con cierta muestra de interés; los gradientes de color rojo representan genes altamente expresados; Los gradientes de color verde representan genes sub-expresados; El color negro representa la expresión de genes que no fueron alterados. Recuperado de (EMBL-EBI, s.f.).

Capítulo 4

Caso de estudio del carcinoma hepatocelular utilizando el lenguaje R y paquete de Bioconductor

Por medio del análisis bioinformático, se puede estudiar los perfiles de expresión diferencial y así, sentar las bases para la interpretación biológica con el objetivo de entender el desarrollo de las enfermedades. Para el presente trabajo, y como se ha mencionado en capítulos anteriores, se utilizaron microarreglos de la empresa Affymetrix, junto con el apoyo de las librerías de Bioconductor, cuyas principales funciones son el análisis y la comprensión de datos genómicos. Adicionalmente, se utilizó el lenguaje de programación R, el cual es el principal intermediario para explotar los conjuntos de datos brindados, así como pintar gráficas de apoyo para el análisis realizado. Finalmente, se muestra una interpretación de la información utilizada; por lo cual se siguió la siguiente metodología:

4.1. Datos de microarreglos

El conjunto de datos analizados proviene del proyecto de investigación del Instituto de Medicina Genómica (INMEGEN), el cual desarrolló un modelo experimental de hepatocarcinogénesis en ratones, inducido químicamente.

La ventaja de los modelos de hepatocarcinogénesis inducida químicamente es que se puede imitar el ciclo de lesión-fibrosis-malignidad vistos en seres humanos. Esto los convierte en los modelos más usados para la investigación del CHC. Se han utilizado principalmente compuestos sintéticos como la dietilnitrosamina, dimetilnitrosamina, etilnitrosourea, 2-acetilaminofluoreno, tetracloruro de carbono y 1,2-dicloroetano (Fuentes Hernández, S., 2018).

4.2. Entorno de trabajo con el Programa R.

R es un lenguaje de comandos de manipulación y análisis estadístico basado en el lenguaje estadístico S, desarrollado por AT&T (Zamora Araya, J.A., 2012). Adicionalmente, es un

lenguaje de código abierto lo que permite que cualquier usuario lo use y lo adapte a sus necesidades.

Además, cuenta con paquetes que extienden sus funciones básicas. Para este proyecto se utilizó el paquete de "Bioconductor", el cual proporciona herramientas para el análisis y la comprensión de datos genómicos de alto rendimiento.

Durante el análisis de datos genómicos se empleó la versión 3.5.3 del lenguaje R y la versión 3.12 del paquete Bioconductor. A continuación, se mencionan los pasos a seguir para trabajar con el paquete Bioconductor.

Iniciar R y ejecutar los comandos mencionados en la **Tabla 1** en el mismo orden, dentro de la interfaz de comando de R para instalar las diferentes funcionalidades requeridas para realizar el análisis de datos genómicos.

Paquetes por instalar.	**Descripción.**
install.packages("dplyr")	Proporcionar herramientas para trabajar con cuadros de datos, tablas de bases de datos, en otras palabras, proporciona los métodos para la explotación de la información.
install.packages("ggplot2")	Permite visualizar la información en gráficas para dar una interpretación de los datos.
install.packages("devtools")	En algunos casos existe la posibilidad, que no se instalen todos los paquetes con sus herramientas requeridas, para ellos es recomendable instalar

	las herramientas de desarrollo del sistema.
install.packages(“plotly”)	Proporciona el manejo de gráficas de tercera dimensión.
install.packages(“BiocManager”)	Gestor de paquetes de Bioconductor, sin dicho paquete no es posible utilizar las diferentes funcionalidades de Bioconductor de igual forma no es posible instalar los demás paquetes.
BiocManager::install(“Biobase”)	Interfaz con los métodos de análisis genómicos, los cuales son implementados en los paquetes subsecuentes.
BiocManager::install(“oligoClasses”)	Contiene definiciones y comprobaciones de validez, también es complemento de los paquetes de oligo y crlmm, por lo que, si se utilizan estos paquetes, oligoClasses debe de estar instalado.
BiocManager::install(“ArrayExpress”)	proporciona los métodos para crear estructuras de datos de Bioconductor.
BiocManager::install(“oligo”)	Permite leer los archivos CEL que son generados por los microarreglos de Affymetrix, dicho de otra forma, el paquete

	de oligo permite analizar las matrices de oligonucleótidos.
BiocManager::install("arrayQualityMetrics")	Permite generar reportes de los datos contenidos en los microarreglos.
BiocManager::install("pd.mogene.2.0.st")	Permite interpretar la información de expresión genética del tejido hepático de ratones.
BiocManager::install("mogene20sttranscriptcluster.db")	Es un objeto de R que contiene asignaciones entre los identificadores de un fabricante y las accesiones de los fabricantes
BiocManager::install("limma")	Permite la aplicación de modelos líneas para el análisis de expresión diferencial.

Tabla 1. Comandos utilizados para el análisis de expresión diferencial.

Estos paquetes fueron necesarios para la realización del presente trabajo en cuestión. Igualmente, en la sección de anexos en el apéndice A, se encuentra el código fuente utilizado para el análisis.

4.3. Extracción de características

Se generó un archivo SDRF (Formato de relación de datos y muestras) el cual describe las características de la muestra y la relación entre muestras, matrices, archivos de datos, etc. La información en el SDRF está organizada de manera que sigue el flujo de un experimento genómico funcional, comienza con la descripción de sus muestras y termina con los nombres de los archivos de datos generados a partir del análisis de los resultados del experimento.

En la **Figura 16** se puede observar al archivo SDRF y el contenido que fue utilizado para el presente análisis.

```
sdrf_location <- "SDRF.file"
SDRF <- read.table(sdrf_location, header=T)
SDRF
#    Source.Name Array.Data.File Factor.Value.phenotype Time.treatment
# 1    PC181.CEL       PC181.CEL                Control        Control
# 2    PC182.CEL       PC182.CEL                Control        Control
# 3    PC183.CEL       PC183.CEL                Control        Control
# 4    PD061.CEL       PD061.CEL                Tratado      6_Semanas
# 5    PD062.CEL       PD062.CEL                Tratado      6_Semanas
# 6    PD063.CEL       PD063.CEL                Tratado      6_Semanas
# 7    PD101.CEL       PD101.CEL                Tratado     10_Semanas
# 8    PD102.CEL       PD102.CEL                Tratado     10_Semanas
# 9    PD103.CEL       PD103.CEL                Tratado     10_Semanas
# 10   PD141.CEL       PD141.CEL                Tratado     14_Semanas
# 11   PD142.CEL       PD142.CEL                Tratado     14_Semanas
# 12   PD143.CEL       PD143.CEL                Tratado     14_Semanas
# 13   PD181.CEL       PD181.CEL                Tratado     18_Semanas
# 14   PD182.CEL       PD182.CEL                Tratado     18_Semanas
# 15   PD183.CEL       PD183.CEL                Tratado     18_Semanas
```

Figura 16. Matriz de información inicial. Generada con los datos de interés, donde se observa la asignación a una variable llamada SDRF, la cual contiene el contenido del archivo.

Con los datos cargados se realizó un mapeo de la información, donde se cambian los nombres de los renglones de la matriz, por los valores de la columna que contiene los nombres de las muestras de interés. En la **Figura 17** se presentan las instrucciones utilizadas para dar este formato.

```
rownames(SDRF) <- SDRF$Array.Data.File
SDRF <- AnnotatedDataFrame(SDRF)
print(SDRF)
##############################################################################
##                                                                          ##
## An object of class 'AnnotatedDataFrame'                                  ##
##   rowNames: PC181.CEL PC182.CEL ... PD183.CEL (15 total)                 ##
##   varLabels: Source.Name Array.Data.File Factor.Value.phenotype Time.treatment ##
##   varMetadata: labelDescription                                          ##
##                                                                          ##
##############################################################################
```

Figura 17. Variable SDRF formateada con el uso del método AnnotatedDataFrame.

Una vez que se tiene el archivo SDRF, se cargó la información de las muestras de interés, como se muestra en la **Figura 18**, se utilizó un método de la librería oligo, la cual permite la lectura de los archivos CEL y mapearlo a un objeto de R.

```
raw_data_dir <- c(getwd())
raw_data <- oligo::read.celfiles(filenames = file.path(raw_data_dir,
                                 SDRF$Array.Data.File),
                                 verbose = FALSE, phenoData = SDRF)

##############################################################################
##                                                                          ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PC181.CEL             ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PC182.CEL             ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PC183.CEL             ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD061.CEL             ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD062.CEL             ##
```

Figura 18. Leyendo archivos CEL del conjunto de datos que fue proporcionado.

Una vez que se extrajo las características de los archivos, se procedió a realizar el control de calidad inicial, en este punto se tenían 2598544 sondas por muestra (se analizaron 15 muestras con este script). En la primera exploración de los datos se puede observar en la **Figura 19** dicha inspección de los mismo, al ser una gran cantidad de sondas se observaron los primeros datos de la matriz, donde se visualizan valores de expresión arriba de 4000 como por debajo de 50.

```
column_sample <- ncol(exprs(raw_data))
row_probes <- nrow(exprs(raw_data))

print(column_sample)
print(c("Número de muestras:",column_sample))
print(c("Número de sondas:",row_probes))

##############################################################################################################
##                                                                                                          ##
##   PC181.CEL PC182.CEL PC183.CEL PD061.CEL PD062.CEL PD063.CEL PD101.CEL PD102.CEL PD103.CEL PD141.CEL PD142.CEL ##
## 1      3108      3698      1976      4092      3112      4392      2415      3499      4518      3718      2506 ##
## 2       107       137        47        98        90        99        81        90       120       123        61 ##
## 3      2893      3630      1980      3925      3025      4605      2372      3128      4661      3255      2353 ##
## 4       101        62        57        83        96       103        87        54       101       137        52 ##
## 5       103       110        72        94        97       147        73        97       131        98        76 ##
## 6        49        63        49       103        54        48        67        68        65        85        61 ##
##   PD143.CEL PD181.CEL PD182.CEL PD183.CEL                                                                  ##
## 1      4129      2980      3031      2790                                                                  ##
## 2        78        60        65        50                                                                  ##
## 3      4158      2838      2854      2969                                                                  ##
## 4        85        57        50        54                                                                  ##
## 5       111        69        82        77                                                                  ##
## 6        68        84        94        59                                                                  ##
##                                                                                                          ##
##############################################################################################################
```

Figura 19. Muestreo inicial de los datos, sin utilizar algún método estadístico.

4.4. Control de calidad

Al tener la información cargada en listas de objetos en R, se inicia la exploración de la información de los datos antes de iniciar el tratamiento de éstos, y así conocer las cantidades de datos iniciales por muestra. En este punto se utilizan los gráficos de cajas y de ACP con el objetivo de reducir el número de variables manteniendo la mayor cantidad de datos posibles. De manera visual, se observa cómo se encuentra distribuida la información del análisis a realizar, dado que datos muy dispersos entre los grupos puede ser un indicador de que la fase experimental podría contener algún error.

4.5. Análisis de componentes principales

Este análisis tiene como objetivo determinar si los datos se agrupan de la manera correcta, debido a las condiciones experimentales. De tal manera, que ACP será de utilidad para encontrar patrones en los datos, de tal manera que éstos puedan comprimirse y reducir el número de muestras.

Se puede realizar un gráfico de los puntos de ACP, utilizando las intensidades logarítmicas de base 2 de los datos, de tal manera que se obtiene un panorama general del muestreo de datos. En este punto se puede utilizar el gráfico de cajas para conocer qué tan alejados de la media se encuentran las varianzas de las muestras; así, al tener varianzas muy diferentes es un indicar que durante la fase experimental sucedió alguna falla; por lo cual, es conveniente verificar el correcto procesamiento de las muestras originales para continuar con los pasos subsecuentes del análisis. En la **Figura 20** se muestran los comandos ejecutados para realizar un ACP.

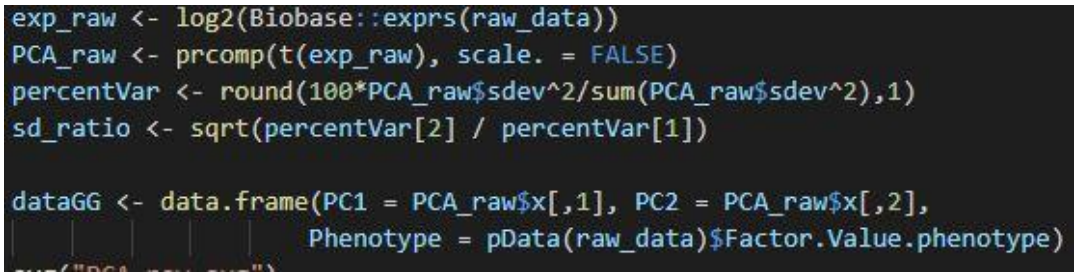

Figura 20. Comandos donde se ejecuta el Análisis de componentes principales.

En la **Figura 21** se presenta de manera gráfica el análisis de componentes principales, donde se utilizaron los identificadores de los datos para representarlos.

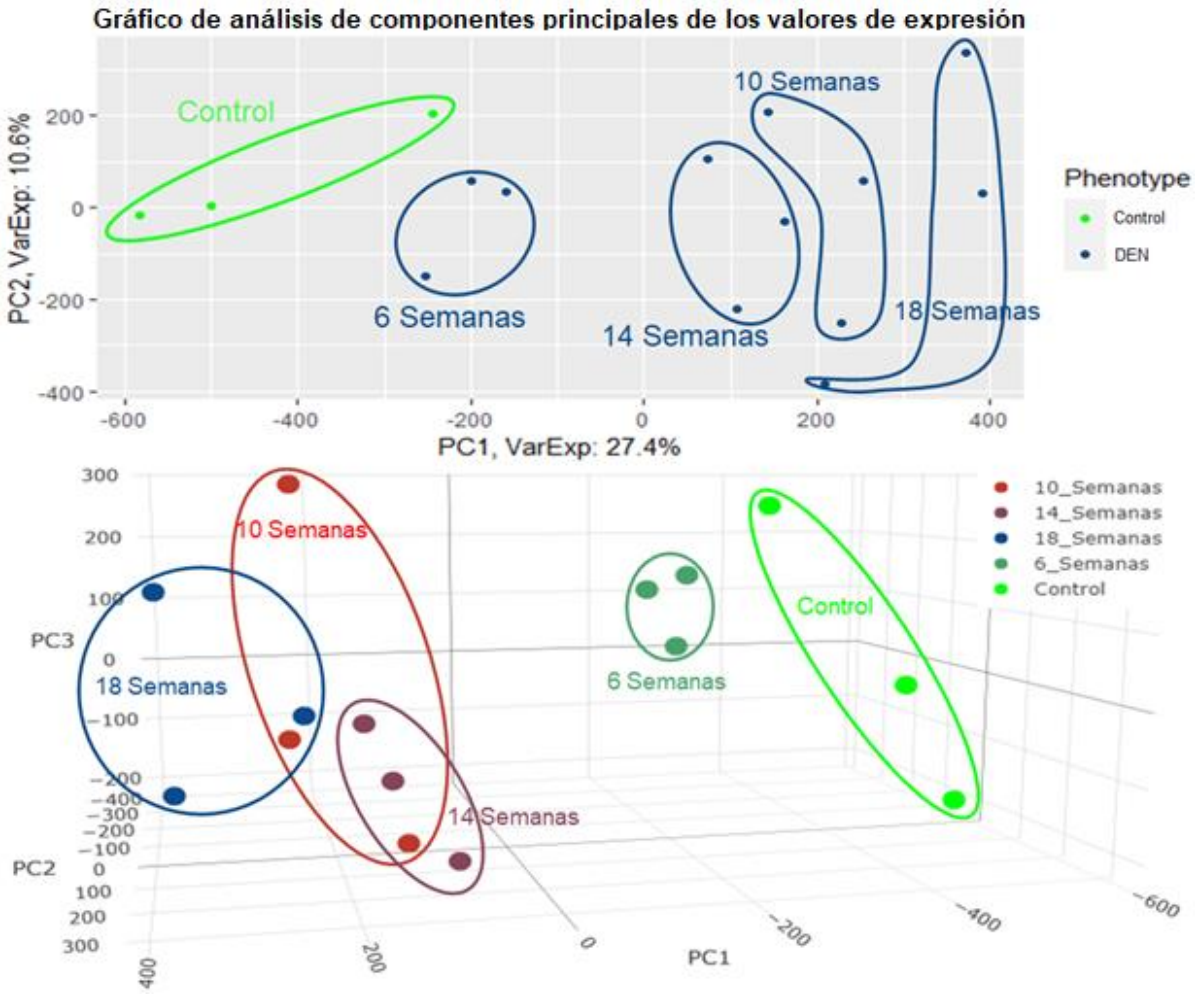

Figura 21. Visualización de la agrupación de la información en 2D y 3D.

Los gráficos muestran que existen valores alejados entre los diferentes grupos experimentales; es decir, entre el grupo control contra el grupo de 18 semanas de tratamiento, existe un 27.4% de diferencias; del mismo modo, en el eje Y se observa un 10.6% de diferencias. El gráfico de los 3 componentes principales de un análisis de ACP, representa que a mayor separación entre cada punto existe una mayor diferencia entre dichos puntos; es decir entre el control contra el grupo de 18 semanas de tratamiento existe un 27.4% de diferencias en sus niveles de expresión genética en el primer componente principal (eje X);

del mismo modo, para el segundo componente principal (eje Y) se observa un 10.6% de diferencia. Al observarse, el control comparado con el grupo de 6 semanas tiene un valor de expresión genética similar, por lo tanto, se puede observar una separación menor entre ellos. Al mismo tiempo, entre los grupos de 10, 14 y 18 semanas los valores de expresión son similares por lo cual se observan con una separación menor entre ellos, sin embargo, existe una mayor diferencia cuando se comparan con el grupo de 6 semanas y el control. Es decir, en etapas tempranas de la enfermedad, se tienen una similitud en la expresión genética comparado contra el grupo de control, mientras que cuando la progresión del carcinoma avanza, se observa una mayor similitud entre los valores de expresión genética en los diferentes tiempos evaluados, pero una mayor separación comparando contra el grupo control.

Los primeros análisis exploratorios permitieron conocer si el conjunto de datos es viable para continuar con el proceso de análisis de microarreglos. Se utilizó el gráfico de cajas como se observa en la **Figura 22**, para conocer si existía una variación entre la varianza de los datos.

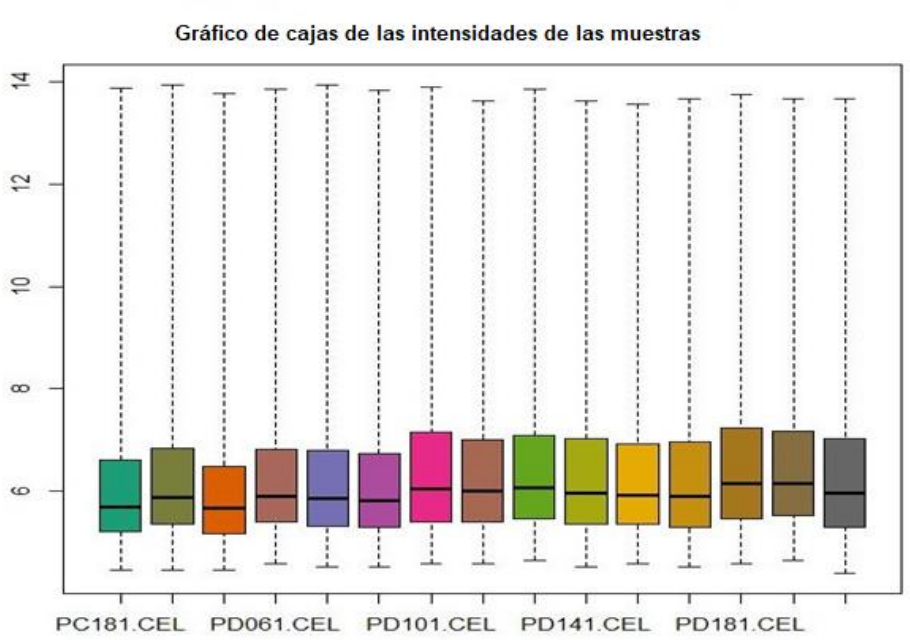

Figura 22. Gráfico de cajas. Este gráfico fue generado por medio del análisis de componentes principales con el objetivo de observar qué tan alejadas se encuentran las varianzas de cada muestra.

En la **Figura 22** se observa una línea de color negro en las diferentes cajas, esta es la varianza de cada muestra, en la cual se puede ver que no hay variaciones significativas entre los datos de los diferentes grupos. Si las variaciones hubieran sido significativas sería un indicador de que el procedimiento experimental fue realizado de manera inadecuada, es decir, si en el gráfico las cajas no estuvieran lo más alineadas posibles, aquellas que se encuentren más alejadas son muestras donde ocurrió algo en la fase experimental. Sin embargo, el análisis muestra que los datos se encuentran de una manera correcta por lo cual se prosigue con las siguientes etapas.

4.6. Normalización

Después de la evaluación inicial de importación y calidad, se procede con la corrección de fondo, dado que una parte de las intensidades de las sondas medidas, se deben a la hibridación no específica y al ruido en el sistema de detección óptica. (Klaus Bernd, 2019). Una vez que se tienen los ajustes a las muestras, se deben de resumir las intensidades normalizadas, así como los ajustes de fondo, en una cantidad proporcional a la cantidad de transcripción del ARNm. Para así mantener el dato independiente del experimento y la tecnología empleada, dado que existen sondas no hibridadas, como lo menciona Klaus Bernd, lo cual genera falsos positivos, y dar un resultado incorrecto en las etapas subsecuentes donde se hacen las comparaciones entre los niveles de expresión.

4.7. Robust multichip average (RMA)

El método para llevar a cabo el análisis de normalización es RMA. El cual consiste en una técnica de estadística con el objetivo de limpiar la información tomando en cuenta la normalización cuantil. Consiste en hacer dos distribuciones idénticas, para realizar esta normalización a una distribución de prueba a una distribución de referencia de la misma longitud, se hace una clasificación. Esto consiste en reorganizar el primer conjunto de datos, cada columna debe ordenarse de menor a mayor valor, después se calcula la media por cada fila para establecer rangos, se toma un orden de clasificación y se reemplazan los valores por la media calculada según sea el caso.

En la **Figura 23** se observa el comando empleado para la realización de este proceso, donde se utilizó el método de RMA.

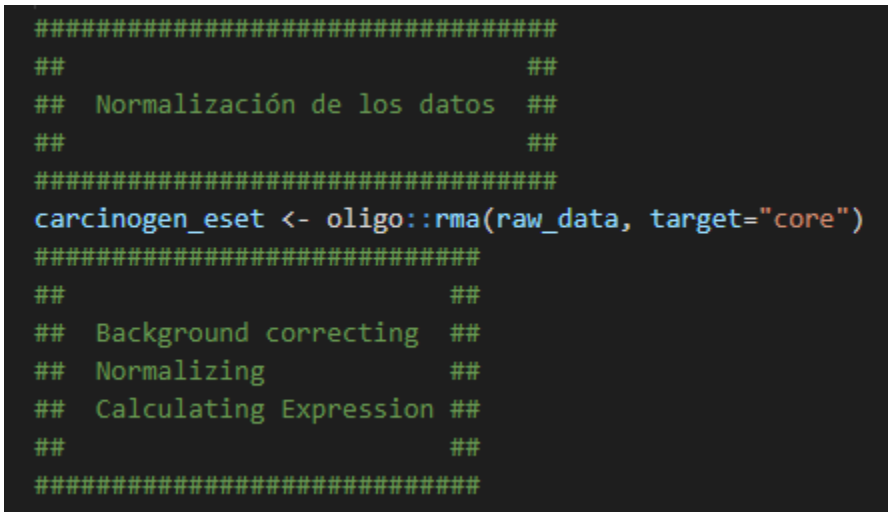

Figura 23. Normalización de la información utilizando el algoritmo de RMA.

Después de realizar dicho proceso se tienen 41345 sondas por cada una de las 15 muestras, con lo cual se redujo en una cantidad significativa el número de sondas. Realizando nuevamente el gráfico de ACP, se observa en la **Figura 24** una mejor agrupación de los datos.

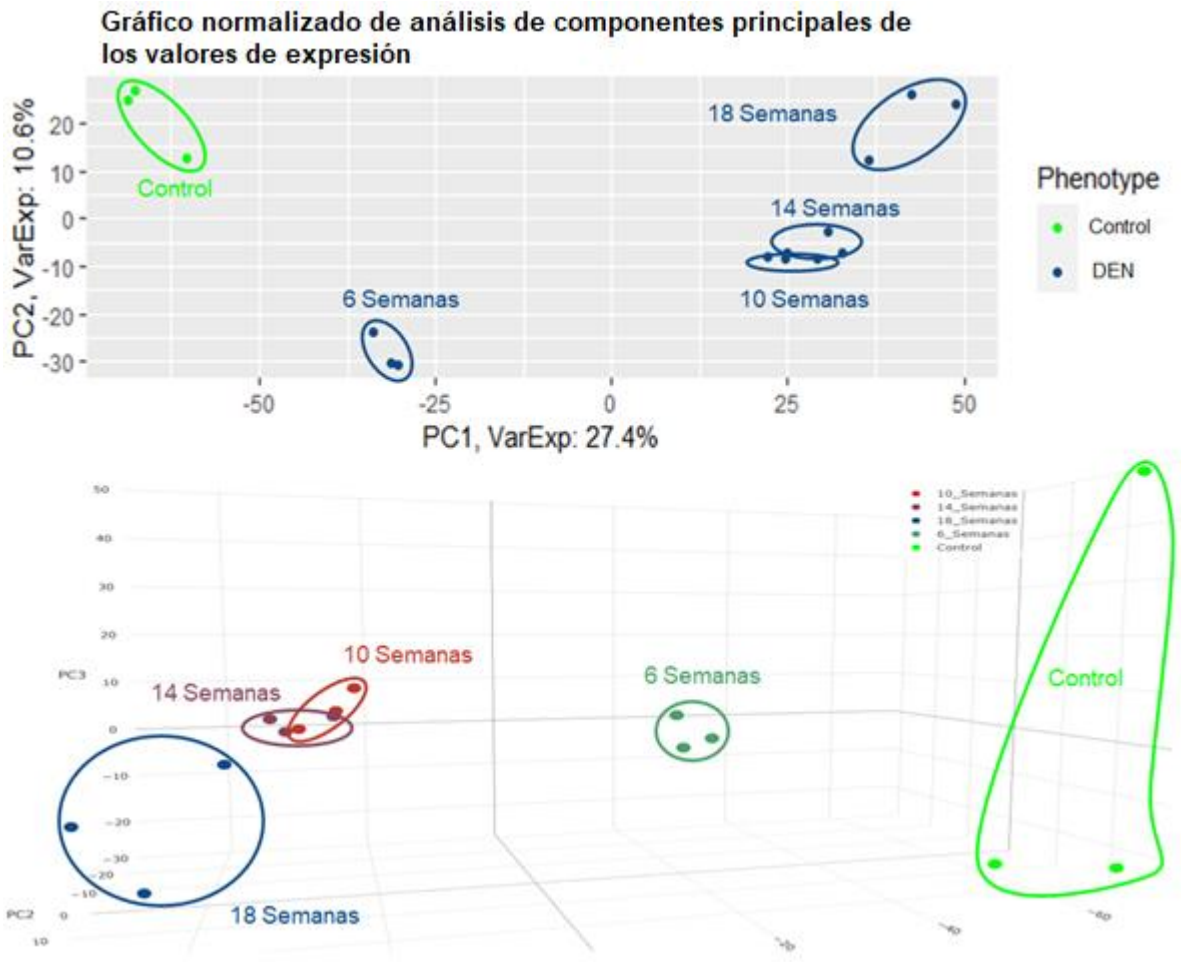

Figura 24. Visualización de la información después del proceso de normalización.

Al generarse el gráfico de componentes principales, después del proceso de normalización, se observó al grupo de 6 semanas separado del control; es decir, su valor de expresión genética ya no es similar al control en comparación al gráfico de la **Figura 18**. Sin embargo, los grupos de 14 y 10 semanas tienen valores similares por lo cual; se observan con una separación menor.

4.8. Mapas de calor

Los mapas de calor son una forma alternativa de representar visualmente la información, trabajan principalmente con *clusters*, lo cual es una técnica de agrupación entre conjuntos de datos que tienen cierta similitud entre ellos, en este caso serán genes que comparten una gran semejanza en el patrón de expresión, y cuyos grupos no se encuentran especificados.

Adicionalmente, la distancia euclidiana calcula la separación que hay entre cada gen, estos valores pueden ser cambiados según sea la necesidad.

Por lo tanto, al reducir la dimensionalidad de los datos con el proceso de normalización, se realizó un mapa de calor, el cual se observa en la **Figura 25**, donde se encuentra la relación de cada grupo con los demás, así como sus diferentes niveles de expresión que estos poseen.

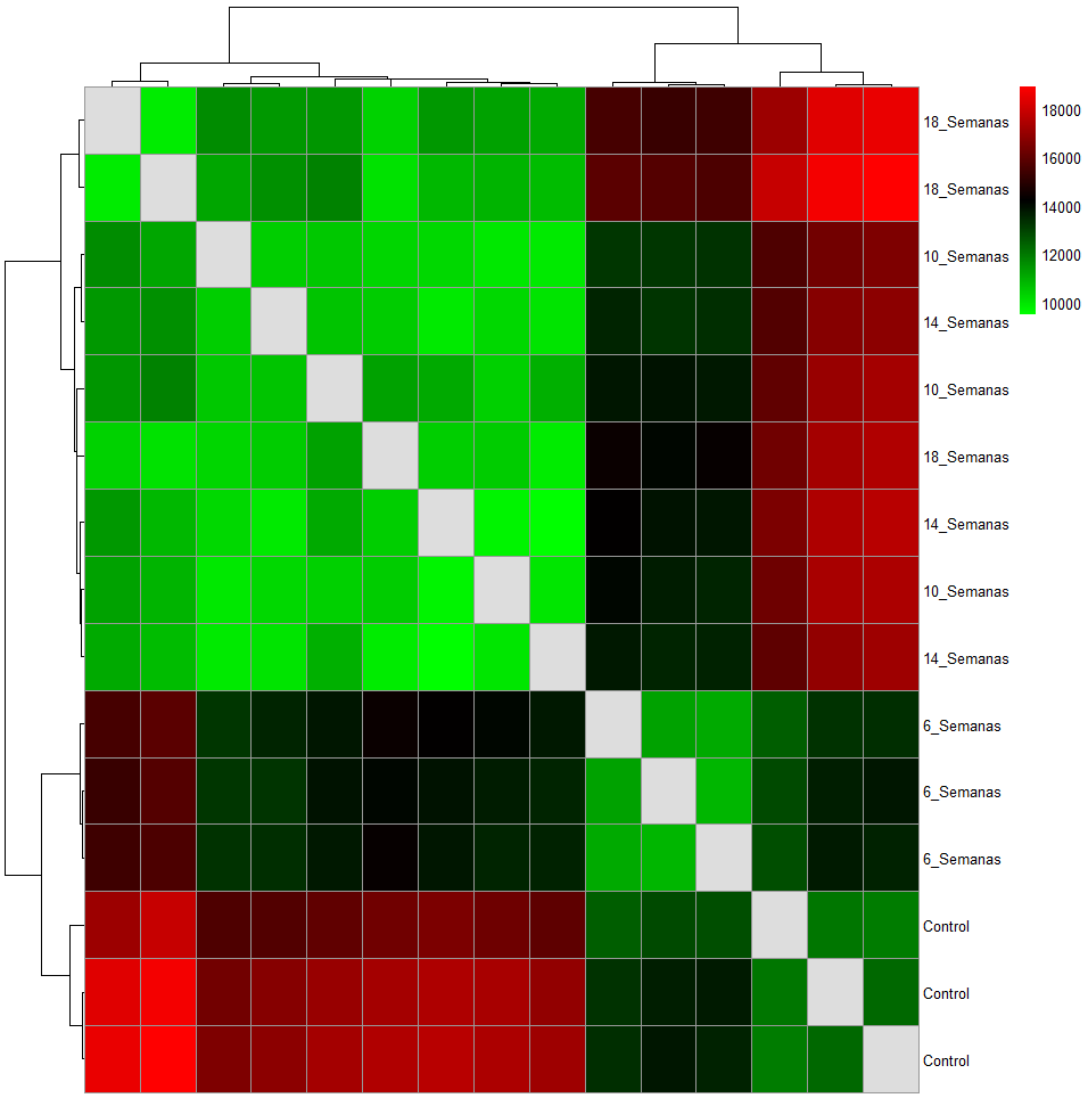

Figura 25. Mapa de calor mostrando la relación de cada muestra. Los gradientes de color rojo cuya tonalidad es más fuerte, representan los genes más expresados; Los gradientes de color negro representan genes con una expresión basal; Los gradientes de color verde representan genes sub-expresados.

En el mapa de calor se observa la relación que existe entre los grupos, es decir, las similitudes que hay entre un grupo control, comparado con alguno de los que se les administró DEN en las diferentes semanas. Adicionalmente, la diagonal principal se encuentra de color gris, dado que es una comparación del grupo consigo mismo, lo cual no muestra ningún cambio.

4.9. Filtrado de genes

Los microarreglos comúnmente muestran una gran cantidad de sondas, entre las cuales pueden estar genes poco expresados. Las librerías de Bioconductor tienen la capacidad de realizar este filtrado, utilizando el paquete de limma, entre sus parámetros se recomienda utilizar una intensidad "suave", con lo cual se puede obtener histograma de las intensidades medias. Se debe ajustar a una distribución normal, para después utilizar el cuantil del 5% de dicha distribución como umbral. De esta forma, se conservarán aquellos genes que muestran una expresión más alta que el límite establecido.

Al aplicar el filtrado de datos se omitieron aquellas sondas poco expresadas del experimento, con lo cual se obtuvo la cantidad final de 39724 sondas por muestra. Para ello se utilizaron los siguientes comandos. Primero se añadieron anotaciones para tener un identificador del gen, el nombre de éste y una descripción, como se observa en la **Figura 26**.

```
carcinogen_filtered_anno <- AnnotationDbi::select(mogene20sttranscriptcluster.db,
    keys=(featureNames(carcinogen_filtered)),
    columns = c("SYMBOL", "GENENAME"),
    keytype="PROBEID")
########################################################################
##                                                                    ##
##  'select()' returned 1:many mapping between keys and columns      ##
##                                                                    ##
########################################################################
```

Figura 26. Comando para generar las anotaciones de los genes.

Con ello se tiene identificado al gen y se podrá hacer una investigación individual del mismo para conocer si puede establecer las bases del CHC.

4.10. Análisis de expresión diferencial

Se realiza una comparativa con los grupos del conjunto datos, de tal manera que se generan tablas de genes, cuyos valores de expresión fueron los más significativos.

Se debe de generar una matriz de contraste, para proceder a realizar la comparación de grupos y así obtener el listado de genes que tuvieron los valores más expresados comparando contra los demás. Para realizar el análisis de expresión diferencial, se creó una nueva matriz como se observa en la **Figura 27**.

```
########################################
##                                    ##
##  Análisis de expresión diferencial  ##
##                                    ##
########################################
carcinogen_exprs_final = exprs(carcinogen_final)
head(carcinogen_exprs_final)
fac_int <- pData(carcinogen_final)$Factor.Value.phenotype
design <- model.matrix(~0 + fac_int)
print(design)
########################################
##                                    ##
##     fac_intControl fac_intTratado  ##
##  1               1              0  ##
##  2               1              0  ##
##  3               1              0  ##
##  4               0              1  ##
##  5               0              1  ##
##  6               0              1  ##
##  7               0              1  ##
##  8               0              1  ##
##  9               0              1  ##
##  10              0              1  ##
##  11              0              1  ##
##  12              0              1  ##
##  13              0              1  ##
##  14              0              1  ##
##  15              0              1  ##
##  attr(,"assign")                   ##
##  [1] 1 1                           ##
##  attr(,"contrasts")                ##
##  attr(,"contrasts")$fac_int        ##
##  [1] "contr.treatment"             ##
##                                    ##
########################################
```

Figura 27. Matriz simple para realizar una comparativa de los grupos de genes.

Con este diseño se generó una nueva matriz de contraste, como se puede ver en la **Figura 28**.

```
cont_tratado <- makeContrasts(Control-Tratado, levels = design)
print(cont_tratado)
###################################
##                                ##
##          Contrasts             ##
## Levels    Control - Tratado    ##
##   Control                  1    ##
##   Tratado                 -1    ##
##                                ##
###################################
```

Figura 28. Matriz de contraste utilizada para generar la tabla de resultados, donde se indican los grupos a comprar "Control-tratado".

4.11. Resultados

Finalmente se generó un mapa de calor por cada cuartil de distribución con la finalidad de separar de mejor manera los datos. Se generaron gráficos de volcán con un valor de cambio diferente, para conocer en número de genes con condiciones de expresión diferencial específica. En la **Figura 29** se observa el primer gráfico de volcán generado para realizar los diferentes mapas de calor, donde se visualizan 39724 genes, pero solo se tomaron en consideración aquellos con un valor de ($p \leq 0.05$) con sus respectivos valores de cambio para cada caso.

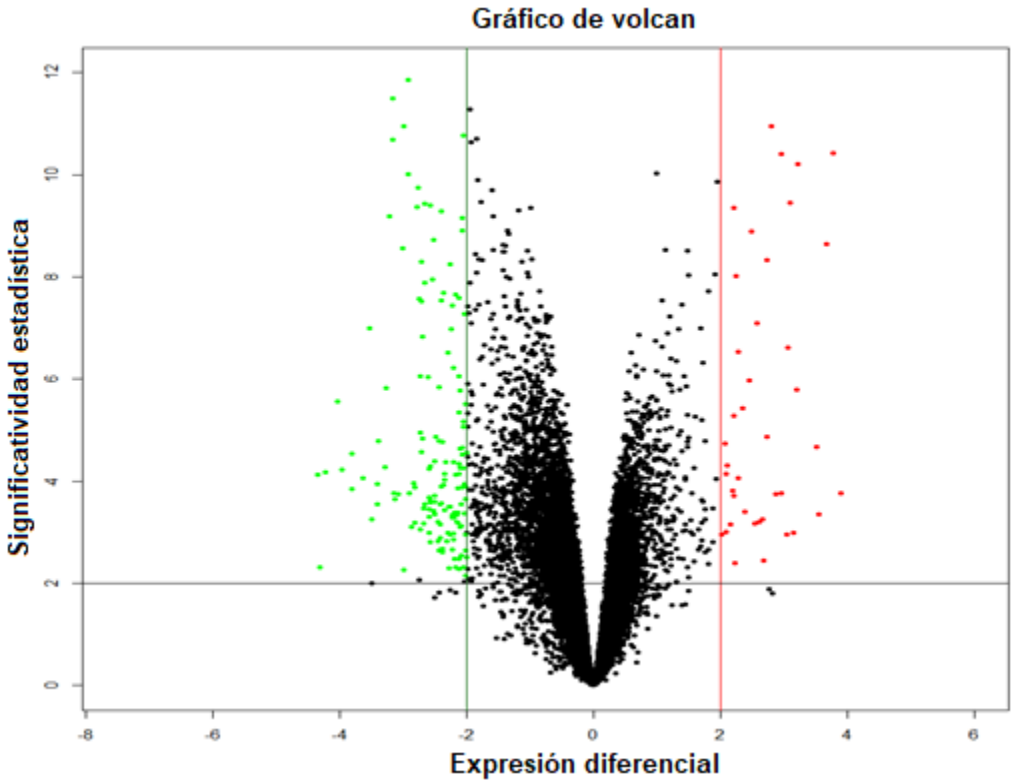

Figura 29. Gráfico de volcán. El gráfico muestra los genes sub-expresados y sobre-expresados en color verde y rojo, respectivamente. Los genes que se analizaron subsecuentemente fueron aquellos sobre-expresados con un valor de cambio mayor a 2 veces respecto al grupo control.

Posteriormente, se generaron los diferentes mapas de calor por cada cuartil. Se generaron 4 mapas de calor con un valor de cambio de 2. Adicionalmente, se muestra el mapa de calor general con todos los genes, el cual está representado en la **Figura 30**. En la **Figura 31** se observa el mapa de calor en el primer cuartil, en la **Figura 32** se muestra el segundo cuartil y finalmente en la **Figura 33** el tercer cuartil. En estos mapas de calor se observa la relación de cada gen con su respectivo tratamiento; así como también, su expresión secuencial a lo largo del periodo en el que se estudió la progresión del CHC.

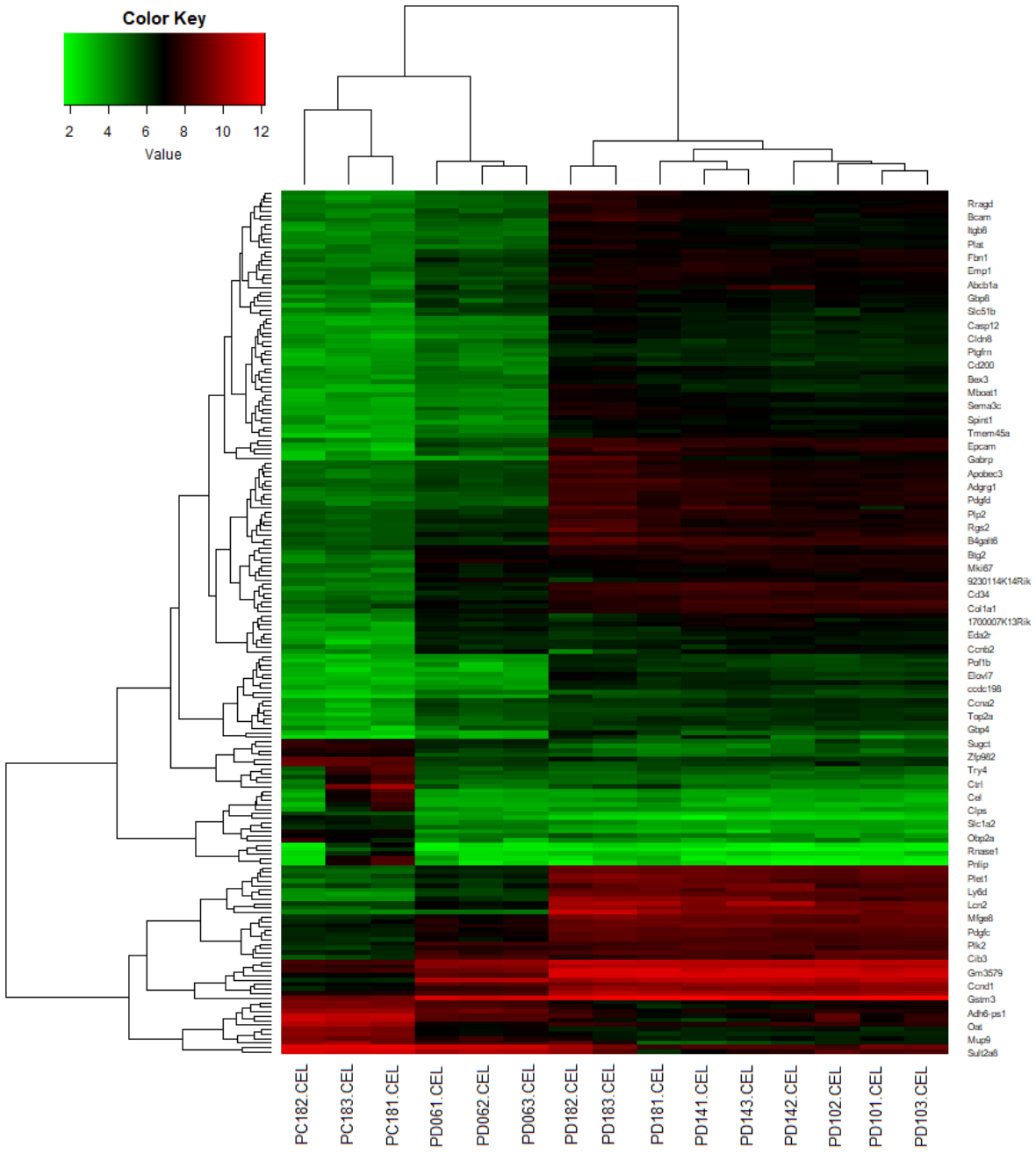

Figura 30. Mapa de calor con todos los genes a una proporción de cambio de 2 (192 genes). Donde se observa cada gen en específico, comparado contra los respectivos tratamientos, en el eje Y se observan los genes, mientras que en el eje X se encuentran las muestras con las que se realizó la comparación.

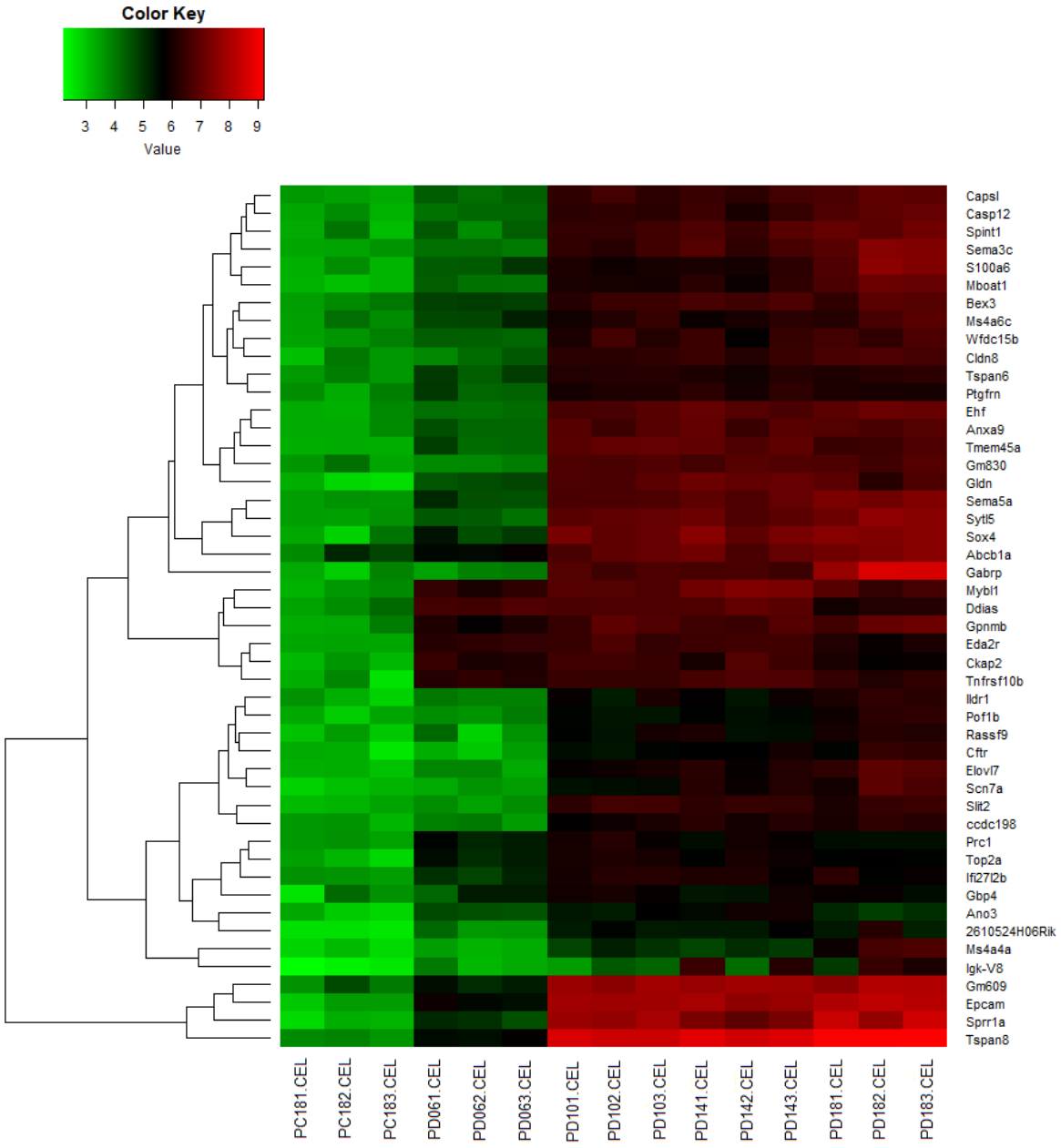

Figura 31. Mapa de calor del primer cuartil con una proporción de cambio de 2 (48 genes). Donde se puede observar cada gen en específico, comparado contra los respectivos tratamientos, en el eje Y se observan los genes, mientras que en el eje X se encuentran las muestras con las que se realizó la comparación.

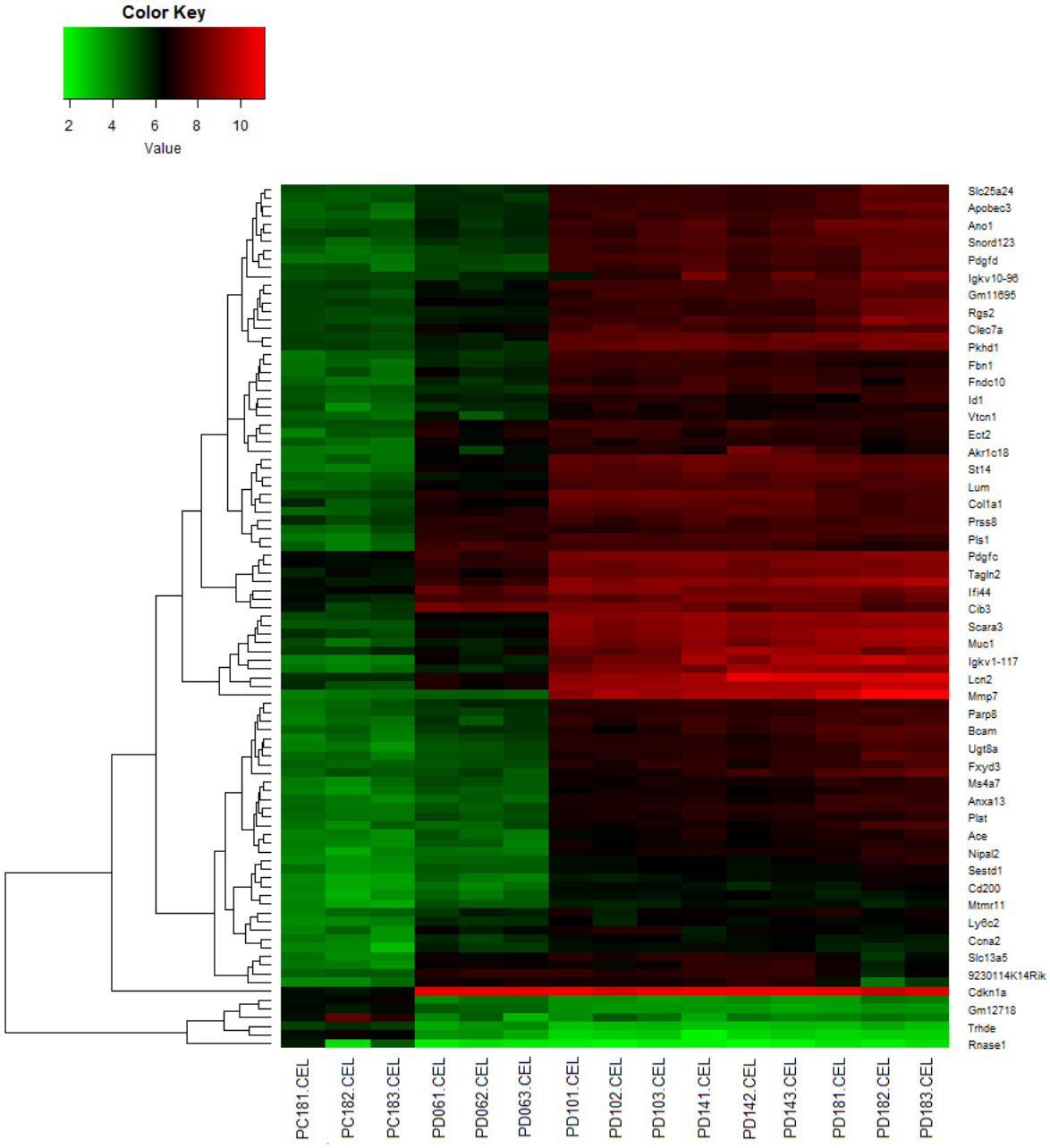

Figura 32. Mapa de calor del segundo cuartil con una proporción de cambio de 2 (99 genes). Donde se puede observar cada gen en específico, comparado contra los respectivos tratamientos, en el eje Y se observan los genes, mientras que en el eje X se encuentran las muestras con las que se realizó la comparación.

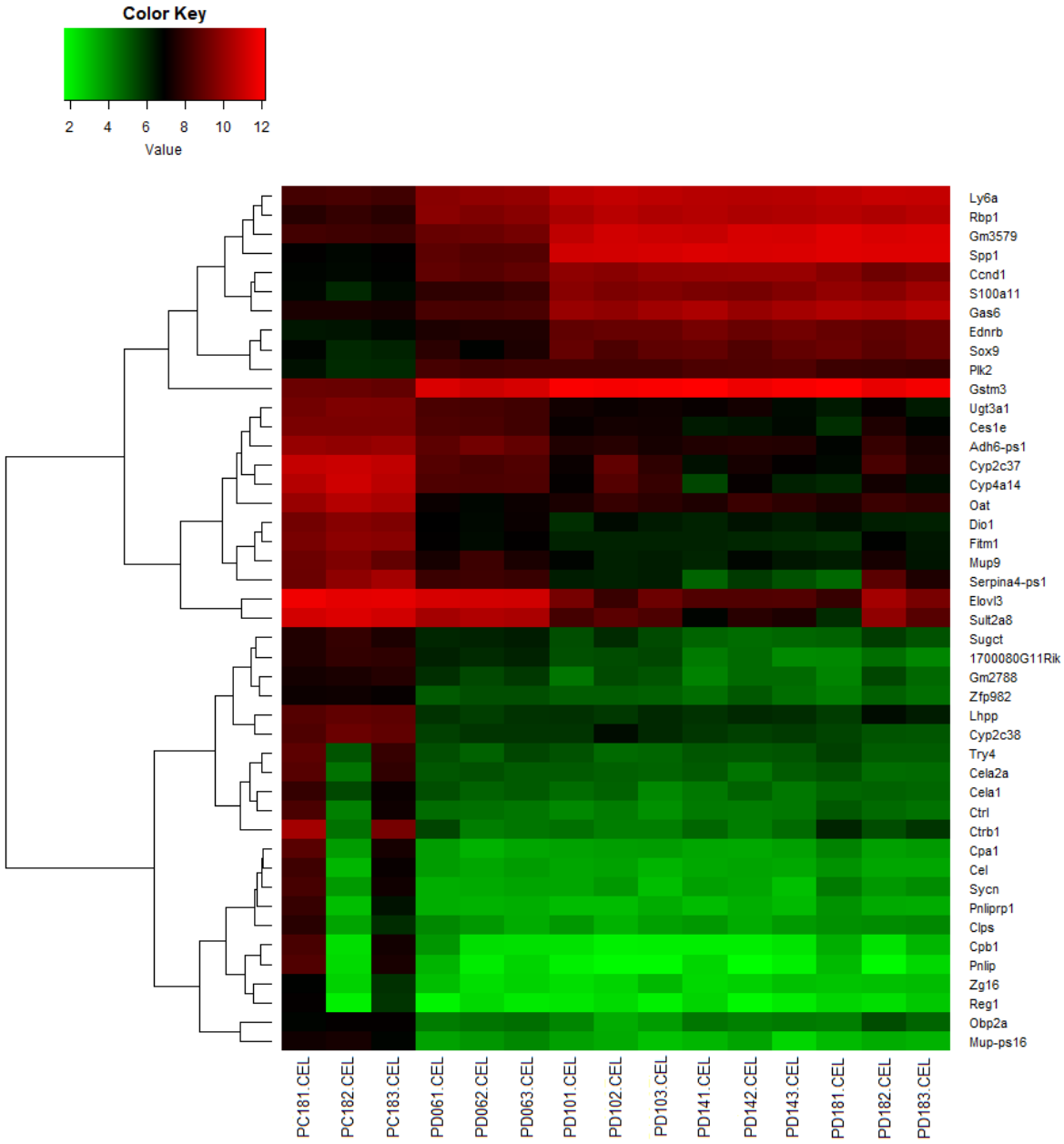

Figura 33. Mapa de calor del tercer cuartil con una proporción de cambio de 2 (45 genes). Donde se puede observar cada gen en específico, comparado contra los respectivos tratamientos, en el eje Y se observan los genes, mientras que en el eje X se encuentran las muestras con las que se realizó la comparación.

Posteriormente, se buscaron aquellos genes con un valor de cambio de 3 veces respecto del grupo control, con lo cual se obtuvo un gráfico de volcán como se observa en la **Figura 34**. El gráfico muestra una reducción de los genes de interés, con la nueva condición de un valor de cambio de 3.

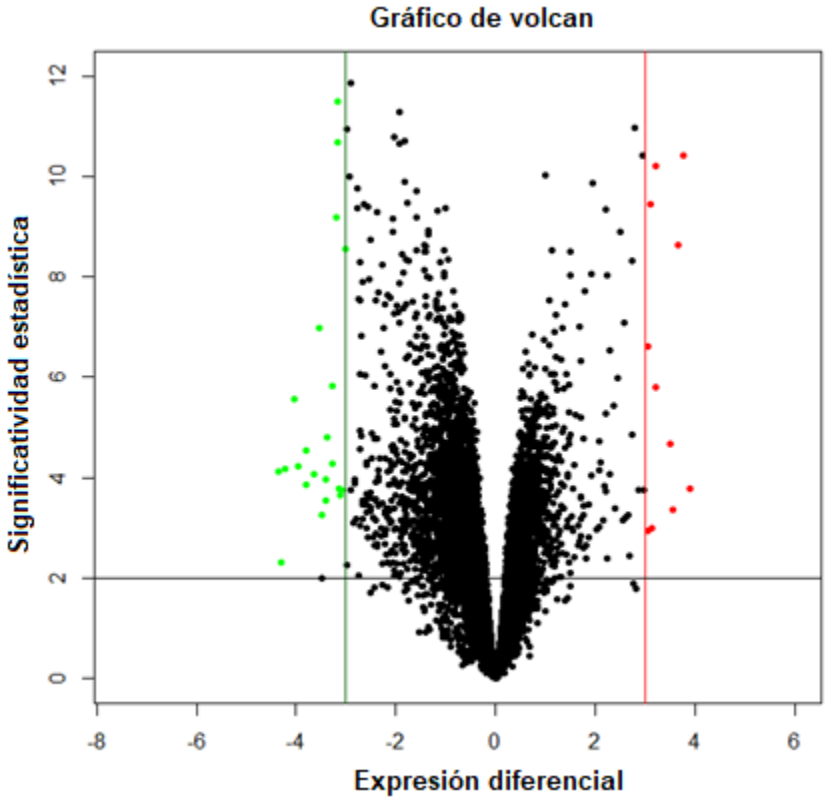

Figura 34. Gráfico de volcán. El gráfico muestra los genes sub-expresados y sobre-expresados en color verde y rojo, respectivamente. Los genes que se analizaron subsecuentemente fueron aquellos sobre-expresados con un valor de cambio mayor a 3 veces respecto al grupo control.

Por lo tanto, al realizar los diferentes mapas de calor, donde se nota una reducción en el tamaño de la matriz dado que la cantidad de genes que cumplen la condición de un valor de cambio de 3 es menor, como se observa en la **Figura 35**, con el mapa de calor general se observa una reducción en la dimensionalidad de los datos, con lo cual la **Figura 36** del primer

cuartil, la **Figura 37** del segundo cuartil y la **Figura 38** con el tercer cuartil, presenta una cantidad menor de genes.

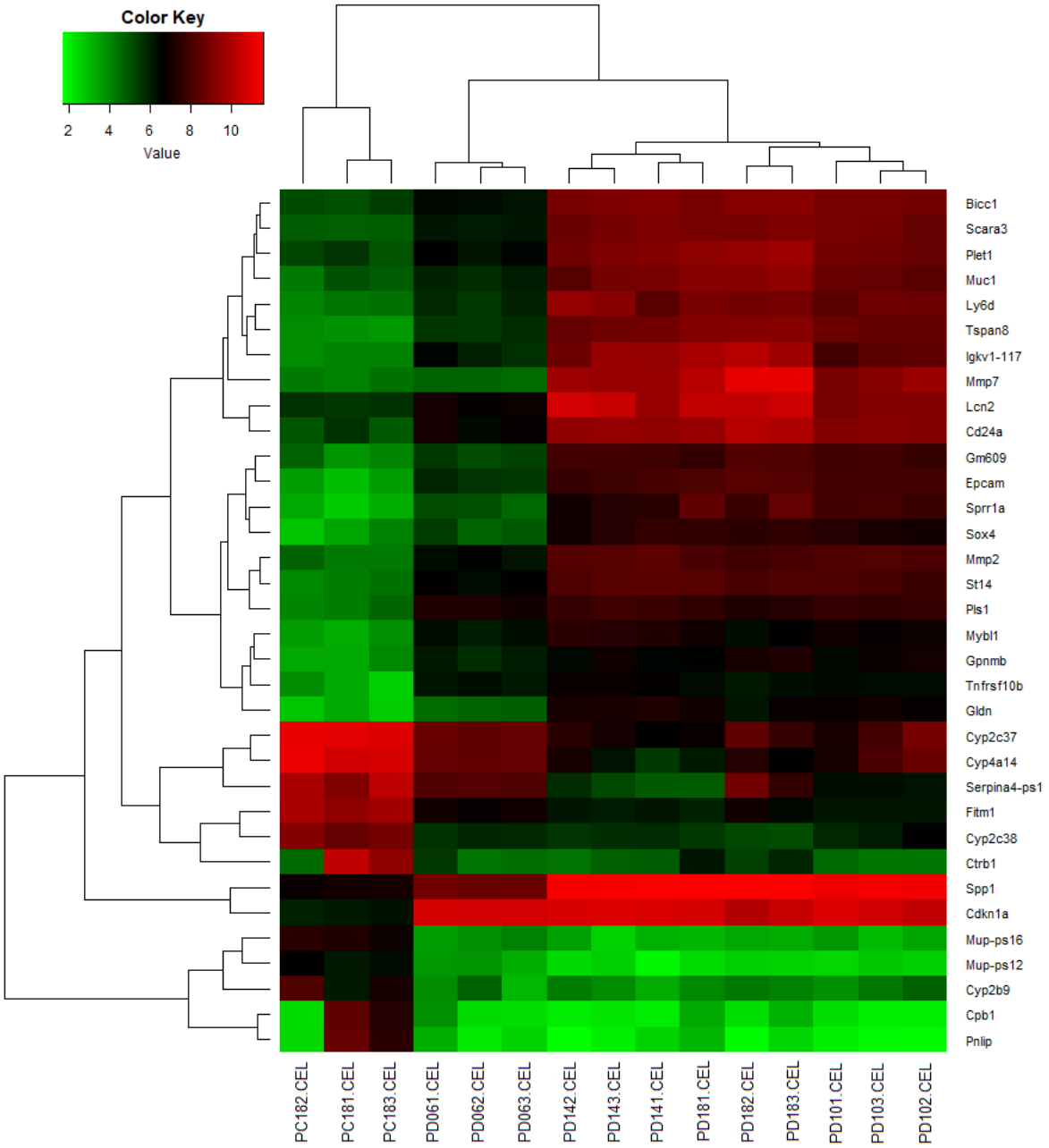

Figura 35. Mapa de calor con todos los genes a una proporción de cambio de 3 (34 genes). Donde se puede observar cada gen en específico, comparado contra los respectivos tratamientos, en el eje Y se observan los genes, mientras que en el eje X se encuentran las muestras con las que se realizó la comparación.

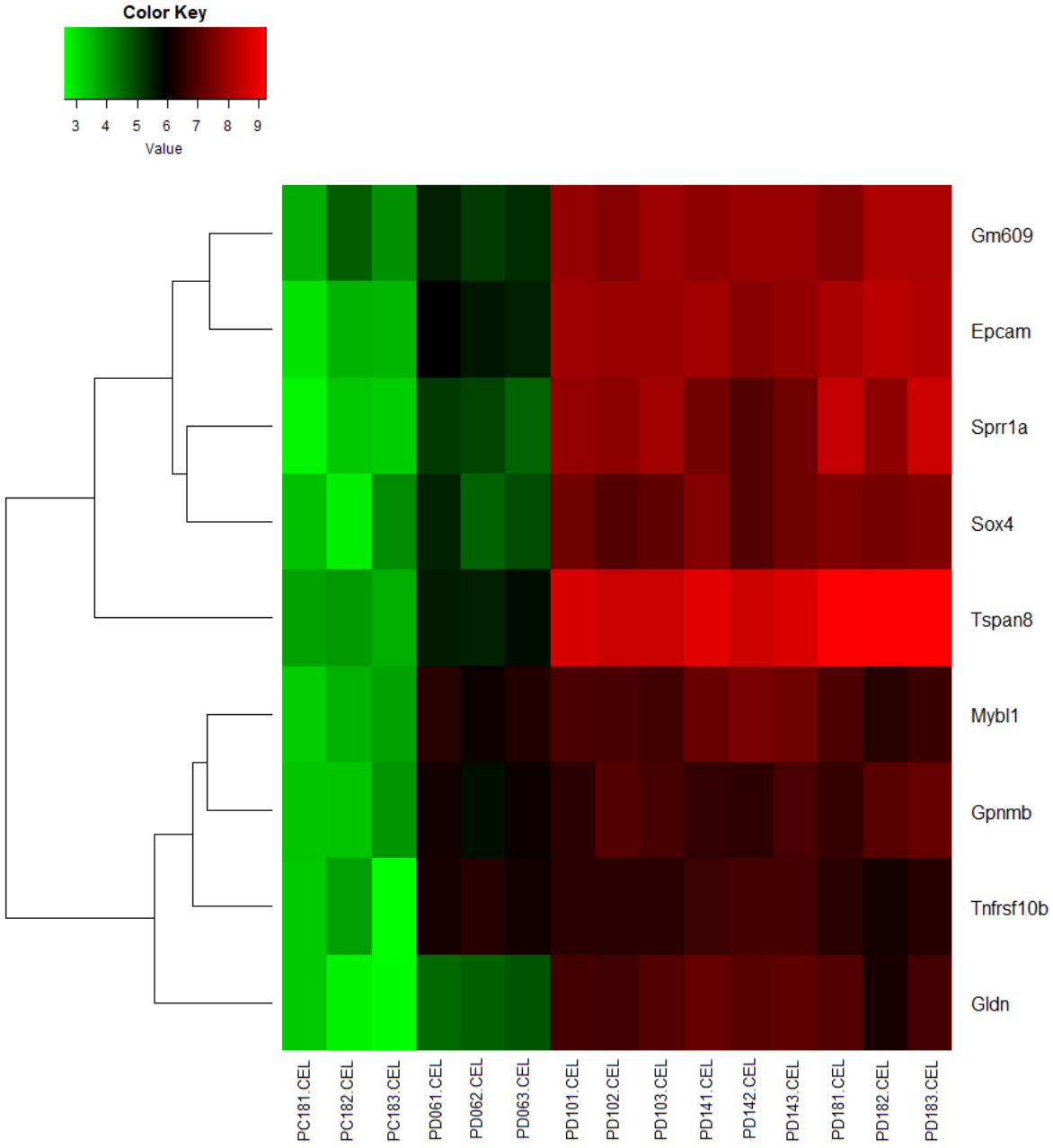

Figura 36. Mapa de calor del primer cuartil con una proporción de cambio de 3 (9 genes). Donde se puede observar cada gen en específico, comparado contra los respectivos tratamientos, en el eje Y se observan los genes, mientras que en el eje X se encuentran las muestras con las que se realizó la comparación.

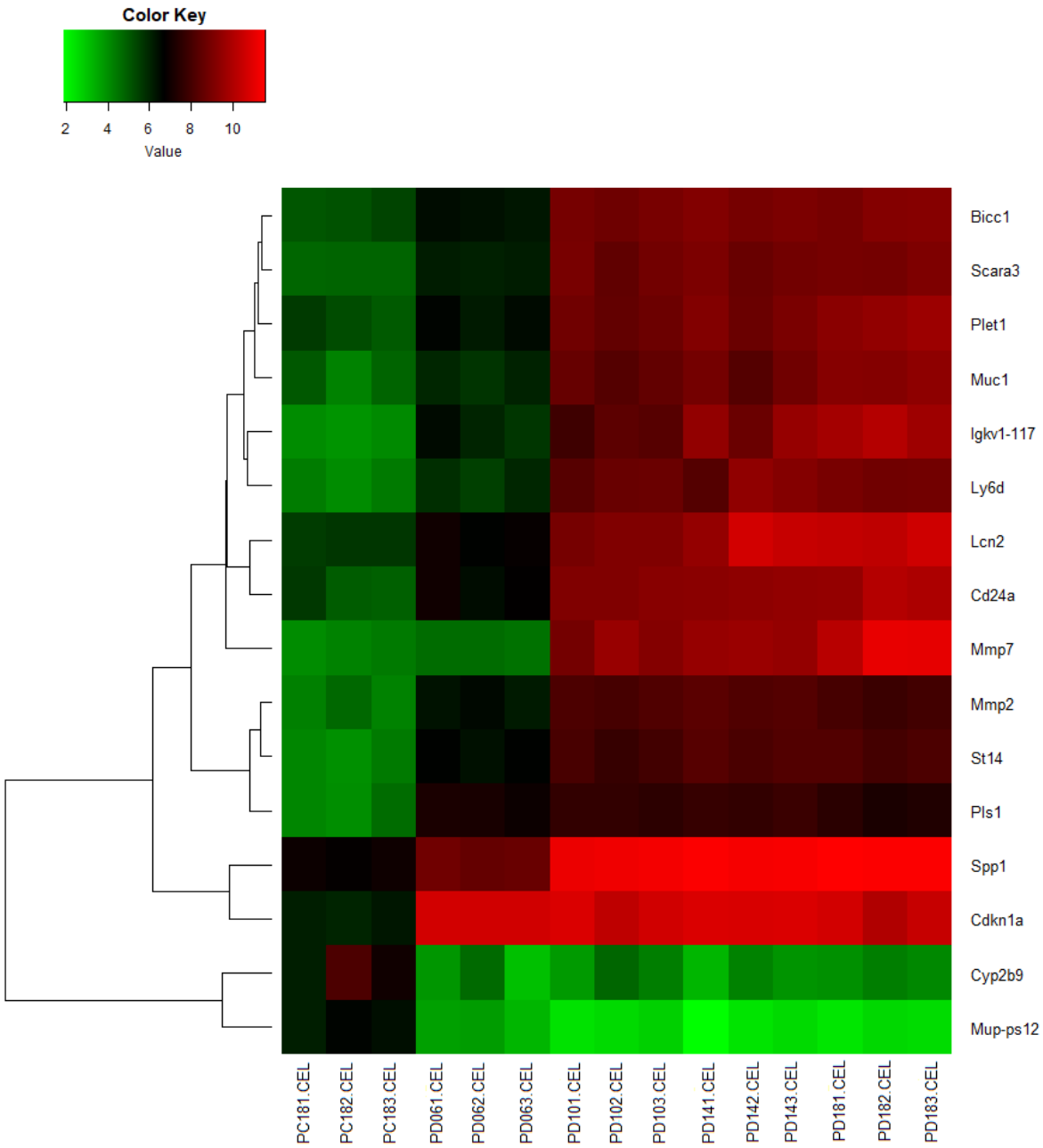

Figura 37. Mapa de calor del segundo cuartil con una proporción de cambio de 3 (16 genes). Donde se puede observar cada gen en específico, comparado contra los respectivos tratamientos, en el eje Y se observan los genes, mientras que en el eje X se encuentran las muestras con las que se realizó la comparación.

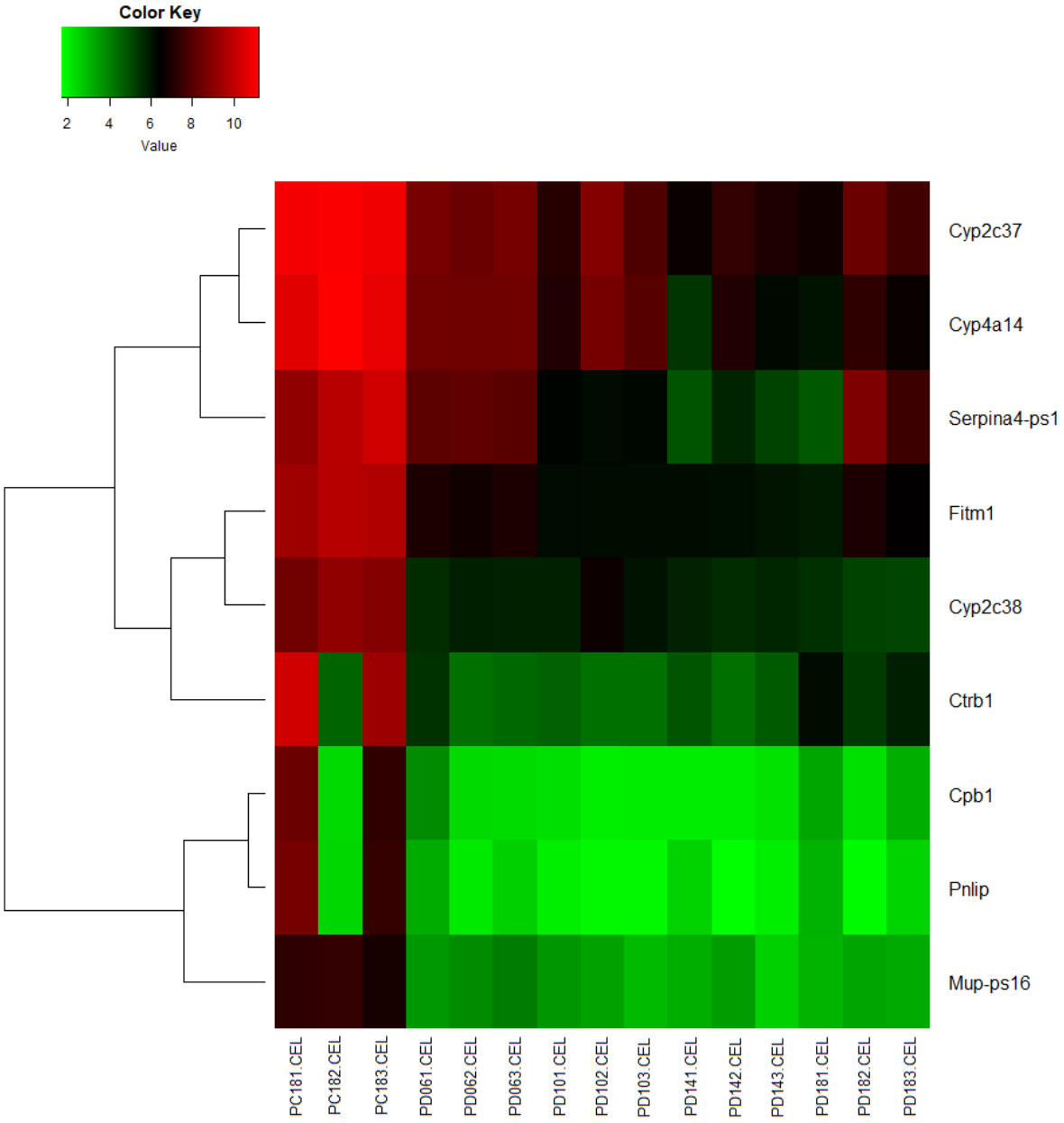

Figura 38. Mapa de calor del tercer cuartil con una proporción de cambio de 3 (9 genes). Donde se puede observar cada gen en específico, comparado contra los respectivos tratamientos, en el eje Y se observan los genes, mientras que en el eje X se encuentran las muestras con las que se realizó la comparación.

De esta manera, se encuentran ciertos genes que estuvieron en los mapas de calor con un valor de cambio de 2 se mantuvieron en los mapas de calor con un valor de cambio de 3, los cuales se observan en la **Tabla 2**. Por lo tanto, los genes que se mantuvieron a un valor de cambio de 3 son los más sobre-expresados, dichos genes se observan en las tablas de resultados finales. Los genes con un valor de cambio mayor a 2 y 3 se enlistan en la **Tabla 2**, la cual muestra los 40 genes sobre-expresados obtenidos después de realizar el análisis. Mientras que en la **Tabla 3** se muestran solo 11 genes sobre-expresados, respectivamente.

Control-DEN		
Lhpp	*Mup9*	*Reg1*
Mup-ps16	*Sugct*	*Cel*
Dio1	*1700080G11Rik*	*Pnliprp1*
Fitm1	*Cela1*	*Ctrl*
Cyp2c38	*Cyp4a14*	*Cela2a*
Zfp982	*Ugt3a1*	*Ctrb1*
Slc1a2	*Adh6-ps1*	*Clps*
Mup-ps12	*Rnase1*	*Serpina4-ps1*
Oat	*Ces1e*	*Sult2a8*
Trhde	*Pnlip*	*Elovl3*
Obp2a	*Cpa1*	
Cyp2b9	*Sycn*	
Gm12718	*Try4*	
Gm2788	*Zg16*	
Cyp2c37	*Cpb1*	

Tabla 2. Lista de genes sobre-expresados. Con una proporción de cambio de 2, y un valor *($p \leq 0.05$)*.

Control-DEN	
Mup-ps16	*Cyp4a14*
Fitm1	*Pnlip*
Cyp2c38	*Cpb1*
Mup-ps12	*Ctrb1*
Cyp2b9	*Serpina4-ps1*
Cyp2c37	

Tabla 3. Lista de genes sobre-expresados. Con una proporción de cambio de 3, y un valor *(p≤0.05)*.

Finalmente, de un análisis inicial de 2598544 sondas por cada muestra se llegó a un resultado de 40 posibles genes comparando el grupo control con los grupos tratados con DEN, con la condicional de un valor de cambio de 2 y un valor *(p≤0.05)*, adicionalmente, se encontraron 11 genes con el mismo valor de *p*, pero con un valor de cambio de 3. Los cuales podrían jugar un papel central en la progresión del CHC.

Trabajo futuro

El análisis del diseño experimental implementado por el laboratorio del INMEGEN, el cual consiste de un grupo de control (sano) y grupos a los que se les indujo el CHC con el carcinógeno DEN a una dosis de 20 mg/kg durante las semanas 6, 10, 14 y 18; se examinó la afectación de DEN en el perfil de expresión, con la finalidad de dar una explicación a la progresión de enfermedades hepáticas crónicas.

Por lo que, el análisis presentado se basa en la comparación de grupos. Para este estudio se empleó la prueba *t*-test, ya que se busca los genes cuyos perfiles de expresión se diferencien de manera significativa entre las condiciones experimentales.

Como perspectivas de este estudio, se requiere llevar a cabo un análisis que permita registrar la anotación de conjuntos de genes en bases de datos para vías metabólicas, procesos biológicos, función molecular, componente celular, ortología, entre otros.

Conclusiones

La tecnología de microarreglos ha permitido la realización de análisis transcriptómicas en el desarrollo del cáncer. Permitiendo que se presenten nuevas ideas y caminos biológicos que permitan comprender la progresión del cáncer. En el modelo de cáncer de hígado de ratón, la expresión de genes se ha utilizado para comparar las alteraciones histopatológicas y moleculares a través de análisis bioinformático. Se han empleado técnicas de ACP, que son utilizadas para encontrar similitudes entre los grupos de genes con respecto al tiempo que duró el tratamiento de cada grupo experimental. El objetivo del método es capturar las varianzas de cada gen, se obtuvo un bosquejo de la información donde se encontraban genes más alejados que otros, debido a que sus varianzas eran más altas que el valor promedio, por lo cual no fue necesario tomar esos valores en etapas posteriores. Es recomendable utilizar la técnica de ACP para conocer cómo se encuentran los datos, y así poder analizar si el modelo en cuestión es adecuado. Además, el gráfico de cajas es una herramienta útil para revisar que los grupos de datos no se encuentren tan dispersos y así complementar la información obtenida por medio de la técnica de ACP. Adicionalmente, el método de RMA redujo el tamaño de la muestra, el cual se basa en la normalización por cuantil y sirve para generar un mapa de calor con los genes y su respectivo tratamiento, para conocer qué tan expresados se encuentran dichos grupos contra los demás. Posteriormente se filtró la información con respecto de una distribución normal, utilizando el cuantil del 5%, para tener definido un umbral de aprobación, y mantener los genes más expresados. Finalmente, se seleccionaron aquellos genes que tengan un valor p menor a 0.05, y cuyo valor absoluto de cambio fue mayor a 2, con lo cual se obtuvo un listado de 40 genes, mientras que con un valor de cambio mayor a 3 se obtuvieron 11 genes que fueron los más expresados durante el análisis, y podrían tener un papel importante en la progresión del CHC, con lo cual se cumplió el objetivo de identificar genes expresados diferencialmente durante la progresión del carcinoma hepatocelular.

Bibliografía

[1] Acevedo, R., C., G., Álvarez, E., Zafra, R., G., et al. (2007). Microarreglos de ADN y cáncer cervicouterino: identificación de marcadores tumorales. 206

[2] American Cancer Society, Recuperado de https://www.cancer.org/es/cancer/aspectos-basicos-sobre-el-cancer/que-es-el-cancer.html

[3] Benítez, Bribiesca, L., (2004). Los microarreglos de DNA y su aplicación clínica. 125

[4] Betancor, L., Gadea, M., Flores, K., (s.f.). Genética bacteriana. 65

[5] BioEnciclopedia (2015). La célula. Recuperado de https://www.bioenciclopedia.com/la-celula/

[6] BIOINNOV, (2016). Expresión Génica. Facultad de Ciencias, UNED, Madrid. Recuperado de http://www.innovabiologia.com/biodiversidad/diversidad-animal/expresion-genica/

[7] Castagnino, J., M., (2006). Nanotecnología, microchips y microarreglos. Federación Bioquímica de la Provincia de Buenos Aires. 162

[8] Celaya, X., (2014). Cáncer de hígado: un mal prevenible Recuperado de https://www.jornada.com.mx/2014/10/02/ls-lacontra.html

[9] Chapman & Hall, (2012). Statistics and Data Analysis for Microarrays Using R and Bioconductor., 2,16,23-25,30-33,715

[10] Chaves, F., J., Martínez-Hervás, S., García García, A., B., (2007). ¿Cómo diseñar un estudio genético? Extracción de ADN y ARN. Avances en Diabetología revista oficial de la sociedad española de diabetes.

[11] Corella. D., Ordovas. M., J., (2017). Conceptos básicos en biología molecular relacionados con la genética y la epigenética. 746

[12] Dangoor Education, Cancer Research UK (2020). Recuperado de https://www.cancerresearchuk.org/what-is-cancer/how-cancer-starts/types-of-cancer

[13] Douglas Hanahan, Weinbger, R., (2011). Hallmarks of Cancer: The next Generation.

[14] EMBL-EBI, Recuperado de https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays/analysis-of-microarray-data/

[15] Equipo de expertos de la Universidad Internacional de Valencia. Recuperado de https://www.universidadviu.com/adn-arn/

[16] Equipo de redactores y equipo de editores médicos de la Sociedad Americana Contra El Cáncer, (2019) ¿Qué es el cáncer de hígado? Recuperado de https://www.cancer.org/es/cancer/cancer-de-higado/acerca/que-es-cancer-de-higado.html

[17] Equipo de redactores y equipo de editores médicos de la Sociedad Americana Contra El Cáncer., (2019). ¿Se puede descubrir el cáncer de hígado en sus comienzos? Recuperado de https://www.cancer.org/es/cancer/cancer-de-higado/deteccion-diagnostico-clasificacion-por-etapas/deteccion.html

[18] Estela Raffino, M., (2020). Célula. Recuperado de: https://concepto.de/celula-2/ Última edición: 10 de agosto de 2020.

[19] Fuentes Hernández, S., (2018). Caracterización de un modelo de hepatocarcinogénesis en el ratón. México CDMX, Ciudad Universitaria., 13, 14, 17

[20] García Flores, M. T., Santillana Hernández. S. P., Galván Oseguera, H., Pérez Rodríguez, G., Martínez Chapa, H. D., (2017). Diagnóstico situacional de la atención oncológica en el instituto Mexicano del Seguro Social. Recuperado de https://www.redalyc.org/jatsRepo/4577/457753492003/html/index.html

[21] García Olmedo, F., (s.f.). La Tercera Revolución Verde Avances en nutrición y alimentación animal. Departamento de Biotecnología, Universidad Politécnica de Madrid. 1

[22] Gémez Mayen, A. P., Corral Guillé, G., (2005). genArise: Herramienta de software para análisis de microarreglos Ciudad Universitaria, Facultad de ciencias

[23] González Aguirre, A. J., Casanova Sánchez, I. E., Vilatobá Chapa, M., Contreras Saldivar, A., Castro Narro, G., García Juárez, I., Huitzil Meléndez, F. D., (2013). Carcinoma hepatocelular: Diagnóstico y tratamiento, Gaceta Mexicana de Oncología., 335

[24] González Torres, L., (2007). Las proteínas en la nutrición. Revista salud pública y nutrición. 1

[25] Gonzalo, R. & Sánchez, A., (2018). Introduction to Microarrays Technologu and Data Analysis, 22

[26] Imaz Rosshandler, I., (2012). Análisis comparativo de algoritmos de normalización en datos de microarreglos de alta densidad Facultad de estudios superiores Acatlán, Naucalpan, Edo de México.

[27] Instituto Nacional del Cáncer (NIH). (Sin fecha). Recuperado de https://www.cancer.gov/espanol/publicaciones/diccionario/

[28] Jiménez García, L., F., Larios Merchant. H., (2003). Biología celular y molecular. 5, 9, 13, 23, 25,26

[29] Jiménez García, L. F. y Merchant Larios, H., (2003). Biología celular y molecular, Edo de México, Naucalpan de Juárez: Pearson Education.

[30] Jollige, I., T., (2002). Principal Component Analysis. 1

[31] Kern Parma, (2018). El hígado, órgano esencial del cuerpo humano. Recuperado de https://www.kernpharma.com/es/blog/el-higado-organo-esencial-del-cuerpo-humano

[32] Khanacademy, Recuperado de https://es.khanacademy.org/science/biology/gene-regulation/gene- regulation-in-eukaryotes/a/overview-of-eukaryotic-gene-regulatio

[33] Klaus Bernd, (2019). An end to end workflow for differential gene expresión using Affymetrix microarrays. F10000Research

[34] Lee, J.S., Chu, I.S., Mikaelyan, A., Calvisi, D.F., Heo, J., Reddy, J.K., Thorgeirsson, S.S., 2004. Application of comparative functional genomics to identify best-fit mouse models to study human cancer. Nat. Genet.

[35] Lewin, B., (2008). Genes IX, Eds. Mc Graw Hill. México. 2, 3, 10, 25

[36] Madan Babu, M. An Introductión to Microarray Data Analysis

[37] Martínez-Hervás, S., García-García, A., B., Chaves, F., J., (2007). Técnicas para el estudio del ADN y el ARN.

[38] Mateos García, D., (s.f.). Sistemas Regulatorios de la Expresión Génica. Departamento de Lenguajes y sistemas informáticos. Universidad de Sevilla. 8, 10, 11, 13

[39] México INFOCÁNCER (2019). El cáncer en el mundo y México. Recuperado de https://www.infocancer.org.mx/?c=conocer-el-cancera=estadisticas-mundiales- y-locales

[40] Meza-Junco, J., Montaño-Loza, A., Aguayo-González, A., (2006). Bases moleculares del cáncer. Revista de investigación clínica, 56, 58, 69

[41] Ming-an Sun, Xiaijian Shao, Yejung Wang., (2018). Microarray Data Analysis for Transcriptome Profiling

[42] Miranda, J., Bringas, R., (2008). Análisis de datos de microarreglos de ADN. **Parte I**: Antecedentes de la tecnología y diseño experimental. Centro de ingeniería genética y biotecnología. 83, 86

[43] Miranda, J., Bringas, R., (2008). Análisis de datos de microarreglos de ADN. **Parte II**: Cuantificación y análisis de la expresión génica. Centro de ingeniería genética y biotecnología. 2, 5, 8

[44] Mora, Díaz, L., Mesa, Irizar, M., (2019). Software de procesamiento de imágenes de microarreglos para diagnósticos. Revista cubana de ciencias informáticas.

[45] Morales General, I., (2017). Creación de una herramienta web para el análisis de datos ómicos, España: Universitat Oberta de Catalunya., 21, 22

[46] National Human Genome Research Institute. Recuperado de https://www.genome.gov/es/genetics-glossary

[47] National Institutes of health, (2002). Cirrosis del Hígado. New York, American Liver Foundation, 3

[48] Ochoa Carrillo, F. J., Cervantes Sánchez, G., Fuentes Alburo, A., (2012). Guía Mexicana de tratamiento del hepatocarcinoma avanzada, Gaceta Mexicana de Oncología, 4

[49] Oliva, R., Ballesta, F., Oriola, J., Clária, J., (2004). Genética Médica. Universitat de Barcelona. 41, 43

[50] OMS, Recuperado de https://www.who.int/es/news-room/fact-sheets/detail/cancer

[51] Papa Lucia, Quaglino, Marta , B., Dianda, Daniela., F., Orellano, Elena., G., Daurelio, Lucas., D., (2015). Identificación de patrones de expresión génica en plantas rutáceas bajo estrés biótico mediante análisis de conglomerados. Universidad Nacional de Rosario., 14

[52] Pevsner, J., (2009). Bioinformatics and Functional Genomics.

[53] Pierce Benjamin, A., (2010). Genética un enfoque conceptual. Southwestern University. 426

[54] Ramírez, J., Chávez, L., Santillán, J., L., Guzmán, Simón., (2003). Microarreglos de DNA. Fac. Medicina, Universidad Nacional Autónoma de México. 97, 105

[55] Reynoso Noverón, N. & Torres Domínguez. J., A., (2017). Epidemiologia del cáncer en México: Carga global y proyecciones 2000-2020, Revista Latinoamericana de Medicina Conductual.

[56] Rinflerch, A., (2008). El ARN: ¿Origen del origen y de la diversidad? Instituto de Ciencias Básicas y Medicina Experimental. 51

[57] Roger D, P., (2019). R Programming for Data Science., 7, 8

[58] Rojas Lemusa, M., Milán Chpavezb, R., Delgado Medinac, A., Bizarro Nevareza, P., Cano Gutiérreza, G., Cafaggi Padilla, D., Cervantes Yépezd, S., (2017). El hepatocito como

un ejemplo de la interacción entre la biología celular y las rutas metabólicas México CDMX, Ciudad Universitaria/Facultad de Medicina, 53, 57, 58

[59] Salcedo, M., Vázquez, G., Hidalgo, A., Pérez, C., Piña, P., Santillán, K., Alatorre, B., Arreola, H., López, R., Montoya, C., Navarro, B., Cerón, T., (2003). MicroArreglos en oncología. Revista especializada en ciencias de la salud, 20

[60] Segovia, J., C., (2013). El gen como medicamento: terapia génica. SEBBM Divulgación la ciencia al alcance de la mano. 1

[61] Soberón. X., Bolivar Zapata. F., (1999). Gen y Genoma. 10, 13, 18

[62] Sociedad Española de oncología médica, (2019) ¿Qué es el cáncer y cómo se desarrolla? Recuperado de https://seom.org/informacion-sobre-el-cancer/que-es-el-cancer-y-como-se-desarrolla

[63] Soto Cruz, I., (2003) Transducción de señales y cáncer. Revista Especializada de la Salud. 46

[64] Tovar, V., Villanueva, A., & Llovet, J. M. (2007). Biología celular y genética en el cáncer de hígado. Gastroenterología y hepatología, 360

[65] Uribe Esquivel, M., García Sáens de Silicia, M., Chávez Tapia, N., Román Sandoval, J., J., (2010). Carcinoma Hepatocelular, Revista de Gastroenterología de México., 175

[66] Valladares Salgado, A., (2004). Alteraciones cromosómicas y perfiles de expresión génica en mujeres mexicanas con cáncer de mama. México D.F., Escuela Nacional de Ciencias Biológicas.

[67] Vallin Plous, C., (2007). Microarreglos de ADN y sus aplicaciones en investigaciones biomédicas. Laboratorio de Genética, Centro de Química Farmacéutica. 1

[68] Vallin Plous, C., (2007). Microarreglos de ADN y sus aplicaciones en investigaciones biomédicas, Ciudad de La Habana, Revista CENIC Ciencias Biológicas

[69] Walter and Eliza Hall Institute of Medical Research, (2020) A guide to creating design matrices for gene expression experiments. Recuperado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7873980/

[70] Zamora Araya, J., A., Vallejos Brenes, R., T., Fernandez Acuña, L., M., (2012). Aprendiendo estadística con R., 1

Anexos

Apéndice A.

Script de Análisis de Datos Genómicos en R

En este apartado se incluye el script generado para la realización del análisis de datos de 2 grupos de información. A continuación, se muestra el código desarrollado para el análisis de datos para la identificación de genes expresados diferencialmente en la progresión del carcinoma hepatocelular. Este código se encuentra disponible para su descarga al público interesado, en un repositorio de GitHub con el siguiente enlace: https://github.com/Noxus14/CHC_analysis

```
############################
##                        ##
##  Librerías a utilizar  ##
##                        ##
############################
library(dplyr)
library(Biobase)
library(oligoClasses)
library(ArrayExpress)
library(oligo)
library(arrayQualityMetrics)
library(pd.mogene.2.0.st)
library(mogene20sttranscriptcluster.db)
library(limma)
library(ggplot2)
library(calibrate)
library(plotly)
library(pheatmap)
library(RColorBrewer)
library(mvtnorm)
library(gplots)
library(DAAG)

########################################
##                                    ##
##  Definiendo ruta de la información ##
##                                    ##
########################################
setwd("C:/../../../DataGenomic/analysisPC-PD")
getwd()

#####################################################################
##                                                                 ##
##  Se carga archivo SDRF con los datos fenotipicos del raton.     ##
##                                                                 ##
#####################################################################
sdrf_location <- "SDRF.file"
SDRF <- read.table(sdrf_location, header=T)
print(SDRF)
#############################################################################
##                                                                         ##
##    Source.Name Array.Data.File Factor.Value.phenotype Time.treatment    ##
## 1    PC181.CEL       PC181.CEL                Control        Control     ##
## 2    PC182.CEL       PC182.CEL                Control        Control     ##
## 3    PC183.CEL       PC183.CEL                Control        Control     ##
## 4    PD061.CEL       PD061.CEL                Tratado      6_Semanas     ##
## 5    PD062.CEL       PD062.CEL                Tratado      6_Semanas     ##
## 6    PD063.CEL       PD063.CEL                Tratado      6_Semanas     ##
## 7    PD101.CEL       PD101.CEL                Tratado     10_Semanas     ##
## 8    PD102.CEL       PD102.CEL                Tratado     10_Semanas     ##
```

```
## 9    PD103.CEL       PD103.CEL                Tratado     10_Semanas  ##
## 10   PD141.CEL       PD141.CEL                Tratado     14_Semanas  ##
## 11   PD142.CEL       PD142.CEL                Tratado     14_Semanas  ##
## 12   PD143.CEL       PD143.CEL                Tratado     14_Semanas  ##
## 13   PD181.CEL       PD181.CEL                Tratado     18_Semanas  ##
## 14   PD182.CEL       PD182.CEL                Tratado     18_Semanas  ##
## 15   PD183.CEL       PD183.CEL                Tratado     18_Semanas  ##
##                                                                       ##
###########################################################################

#############################################################
##                                                         ##
##  Se les asigna nombre a los renglones del archivo,      ##
##  tomando los valores de la columna Array.Data.File      ##
##                                                         ##
#############################################################
rownames(SDRF) <- SDRF$Array.Data.File
SDRF <- AnnotatedDataFrame(SDRF)
print(SDRF)

###############################################################################################
##                                                                                           ##
## An object of class 'AnnotatedDataFrame'                                                   ##
##   rowNames: PC181.CEL PC182.CEL ... PD183.CEL (15 total)                                  ##
##   varLabels: Source.Name Array.Data.File Factor.Value.phenotype Time.treatment            ##
##   varMetadata: labelDescription                                                           ##
##                                                                                           ##
###############################################################################################

#####################################
##                                 ##
##  Extracción de características  ##
##                                 ##
#####################################
raw_data_dir <- c(getwd())
raw_data <- oligo::read.celfiles(filenames = file.path(raw_data_dir,
                                 SDRF$Array.Data.File),
                                 verbose = FALSE, phenoData = SDRF)
#########################################################################
##                                                                     ##
##  Colocar los archivos .CEL en la misma ruta del SDRF.               ##
##  raw_Data contiene los archivos .CEL, los datos de expresiones      ##
##  y los datos fenotipicos en un mismo archivo.                       ##
##                                                                     ##
#########################################################################

###############################################################################
##                                                                           ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PC181.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PC182.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PC183.CEL              ##
```

```
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD061.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD062.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD063.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD101.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD102.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD103.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD141.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD142.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD143.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD181.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD182.CEL              ##
## Reading in : C:/../../../DataGenomic/analysisPC-PD/PD183.CEL              ##
##                                                                           ##
###############################################################################

#####################################################
##                                                 ##
##  Se detiene si no es un objeto valido de R.     ##
##                                                 ##
#####################################################
stopifnot(validObject(raw_data))

#################################################
##                                             ##
##  Head solo da los primeros 5 datos.         ##
##                                             ##
#################################################
head(Biobase::pData(raw_data))

###############################################################################################
##                                                                                           ##
##            Source.Name Array.Data.File Factor.Value.phenotype Time.treatment             ##
##  PC181.CEL   PC181.CEL       PC181.CEL                Control        Control             ##
##  PC182.CEL   PC182.CEL       PC182.CEL                Control        Control             ##
##  PC183.CEL   PC183.CEL       PC183.CEL                Control        Control             ##
##  PD061.CEL   PD061.CEL       PD061.CEL                Tratado      6_Semanas             ##
##  PD062.CEL   PD062.CEL       PD062.CEL                Tratado      6_Semanas             ##
##  PD063.CEL   PD063.CEL       PD063.CEL                Tratado      6_Semanas             ##
##                                                                                           ##
###############################################################################################

#####################################################################
##                                                                 ##
##  Conocer el número de muestra, el cual debe de coincidir        ##
##  con el número de muestras del SDRF.                            ##
##                                                                 ##
#####################################################################
column_sample <- ncol(exprs(raw_data))
row_probes <- nrow(exprs(raw_data))

print(column_sample)
print(c("Número de muestras:",column_sample))
```

```
print(c("Número de sondas:",row_probes))
```

```
##############################################################################################################
##                                                                                                          ##
##   PC181.CEL PC182.CEL PC183.CEL PD061.CEL PD062.CEL PD063.CEL PD101.CEL PD102.CEL PD103.CEL PD141.CEL PD142.CEL ##
## 1      3108      3698      1976      4092      3112      4392      2415      3499      4518      3718      2506 ##
## 2       107       137        47        98        90        99        81        90       120       123        61 ##
## 3      2893      3630      1980      3925      3025      4605      2372      3128      4661      3255      2353 ##
## 4       101        62        57        83        96       103        87        54       101       137        52 ##
## 5       103       110        72        94        97       147        73        97       131        98        76 ##
## 6        49        63        49       103        54        48        67        68        65        85        61 ##
##   PD143.CEL PD181.CEL PD182.CEL PD183.CEL                                                                  ##
## 1      4129      2980      3031      2790                                                                  ##
## 2        78        60        65        50                                                                  ##
## 3      4158      2838      2854      2969                                                                  ##
## 4        85        57        50        54                                                                  ##
## 5       111        69        82        77                                                                  ##
## 6        68        84        94        59                                                                  ##
##                                                                                                          ##
##############################################################################################################
```

```
###############################################
##                                           ##
##  dim(exprs(raw_data))                     ##
##   renglones - 2598544     Columnas - 15   ##
##                                           ##
###############################################
```

```
########################################
##                                    ##
##  Inicio del control de calidad.    ##
##                                    ##
########################################
```

```
############################################################
##                                                        ##
##  Biobase::exprs(raw_data)[1:5, 1:5] - (Opcional)       ##
##                                                        ##
############################################################
```

```
#############################################################
##                                                         ##
##   PC181.CEL PC182.CEL PC183.CEL PD061.CEL PD062.CEL     ##
## 1      3108      3698      1976      4092      3112     ##
## 2       107       137        47        98        90     ##
## 3      2893      3630      1980      3925      3025     ##
## 4       101        62        57        83        96     ##
## 5       103       110        72        94        97     ##
##                                                         ##
#############################################################
exp_raw <- log2(Biobase::exprs(raw_data))
PCA_raw <- prcomp(t(exp_raw), scale. = FALSE)
percentVar <- round(100*PCA_raw$sdev^2/sum(PCA_raw$sdev^2),1)
sd_ratio <- sqrt(percentVar[2] / percentVar[1])

dataGG <- data.frame(PC1 = PCA_raw$x[,1], PC2 = PCA_raw$x[,2],
                     Phenotype = pData(raw_data)$Factor.Value.phenotype)
```

```
#####################################################################
##                                                                 ##
##  Gráfico de Análisis de componentes principales.                ##
##  Agregar más colores en caso de tener más de 2 fenotipos.       ##
##                                                                 ##
#####################################################################
jpeg("PCA_raw.jpg")
ggplot(dataGG, aes(PC1, PC2)) + geom_point(aes(colour = Phenotype)) +
       ggtitle("PCA plot of the log-transformed raw expression data") +
       xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
       ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
       theme(plot.title = element_text(hjust = 0.5)) +
       coord_fixed(ratio = sd_ratio) +
       scale_shape_manual(values = c(4,15)) +
       scale_color_manual(values = c("darkorange2", "dodgerblue4")) +
       theme_bw()
dev.off()

########################
##                    ##
##  Gráfico de cajas  ##
##                    ##
########################
svg("Boxplot_raw.svg")
oligo::boxplot(raw_data, target = "core",
               main = "Gráfico de cajas del log2 de las intensidades")
dev.off()

########################################
##                                    ##
##  Generar gráfico en 3D (Opcional)  ##
##                                    ##
########################################
#############################################################################################
##                                                                                         ##
##  dataGG <- data.frame(PC1 = PCA_raw$x[,1], PC2 = PCA_raw$x[,2], PC3 = PCA_raw$x[,3],    ##
##                   Phenotype = pData(raw_data)$Factor.Value.phenotype,                   ##
##                   Time = pData(raw_data)$Time.treatment)                                ##
##                                                                                         ##
##  p <- plot_ly(dataGG, x = ~PC1, y = ~PC2, z = ~PC3,                                     ##
##           color = ~Time, colors = c('#BF382A', '#0C4B8E','#D0F9B1'))                    ##
##                                                                                         ##
##  plot_ly(dataGG, x = ~PC1, y = ~PC2, z = ~PC3,                                          ##
##       color = ~Time, colors = c('#BF382A', '#0C4B8E','green'))                          ##
##                                                                                         ##
#############################################################################################

arrayQualityMetrics(expressionset = raw_data,
                    outdir = "Reporte_Sin_Normalizar",
                    force = TRUE , do.logtransform = TRUE,
                    intgroup = c("Factor.Value.phenotype", "Time.treatment")
)
```

```
#################################
##                             ##
##  Normalización de los datos  ##
##                             ##
#################################
carcinogen_eset <- oligo::rma(raw_data, target="core")
############################
##                        ##
##  Background correcting  ##
##  Normalizing            ##
##  Calculating Expression ##
##                        ##
############################

############################
##                        ##
##  dim(carcinogen_eset)  ##
##  Features  Samples     ##
##    41345       15      ##
##                        ##
############################

exp_carcinogen_eset <- exprs(carcinogen_eset)

#############################################################
##                                                         ##
##  Indica número de renglones y columnas,                 ##
##  llevar control del universo de probes que tenemos.     ##
##                                                         ##
#############################################################
dim(exp_carcinogen_eset)
PCA_eset <- prcomp(t(exprs(carcinogen_eset)), scale = FALSE)

dataGG <- data.frame(PC1 = PCA_eset$x[,1], PC2 = PCA_eset$x[,2],
                     Phenotype = pData(carcinogen_eset)$Factor.Value.phenotyp
e)

#####################################################################
##                                                                 ##
##  Gráfico de PCA Normalizado                                     ##
##  Agregar más colores en caso de tener más de 2 fenotipos.       ##
##                                                                 ##
#####################################################################
jpeg("PCA_raw_Normalized.jpg")
ggplot(dataGG, aes(PC1, PC2)) + geom_point(aes(colour = Phenotype)) +
        ggtitle("PCA plot of the log-transformed raw expression data") +
        xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
        ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
```

```
        theme(plot.title = element_text(hjust = 0.5)) +
        coord_fixed(ratio = sd_ratio) +
        scale_shape_manual(values = c(4,15)) +
        scale_color_manual(values = c("green", "dodgerblue4")) +
        theme_bw()
dev.off()

##########################################
##                                      ##
##  Generar gráfico en 3D (Opcional)  ##
##                                      ##
##########################################
#############################################################################################
##                                                                                         ##
##  dataGG <- data.frame(PC1 = PCA_eset$x[,1], PC2 = PCA_eset$x[,2], PC3 = PCA_eset$x[,3],  ##
##                  Phenotype = pData(carcinogen_eset)$Factor.Value.phenotype,             ##
##                  Time = pData(carcinogen_eset)$Time.treatment)                          ##
##                                                                                         ##
##  p <- plot_ly(dataGG, x = ~PC1, y = ~PC2, z = ~PC3,                                     ##
##           color = ~Time, colors = c('#BF382A', '#0C4B8E','#D0F9B1'))                    ##
##                                                                                         ##
##  plot_ly(dataGG, x = ~PC1, y = ~PC2, z = ~PC3,                                          ##
##       color = ~Time, colors = c('#BF382A', '#0C4B8E','green'))                          ##
##                                                                                         ##
#############################################################################################

###############################################
##                                           ##
##       Generar Mapa de Calor               ##
##  de las relaciones entre las muestras     ##
##                                           ##
###############################################

dists <- as.matrix(dist(t(exp_carcinogen_eset), method = "manhattan"))
colnames(dists) <- NULL
diag(dists) <- NA
rownames(dists) <- pData(carcinogen_eset)$Time.treatment
hmcol <- colorRampPalette(c("red","black","green"))(255)

png("Heatmap.png")
pheatmap(dists, col = rev(hmcol), clustering_distance_rows = "manhattan",
        clustering_distance_cols = "manhattan")
dev.off()

############################
##                        ##
##  Se genera Histograma  ##
##                        ##
############################

carcinogen_medians <- rowMedians(exprs(carcinogen_eset))
svg("hist_normal.svg")
```

```
hist_res <- hist(carcinogen_medians, 100, freq=FALSE)
dev.off()

#########################
##                     ##
##  Filtrado de Genes  ##
##                     ##
#########################

emp_mu <- hist_res$breaks[which.max(hist_res$density)]
emp_sd <- mad(carcinogen_medians)/2
prop_cental <- 0.50 ##  Estandar de la distribucion
cut_val <- 0.05 / prop_cental ## el 5%  de la distribucion
thresh_median <- qnorm(0.05 / prop_cental, emp_mu, emp_sd)
no_of_samples <- table(paste0(pData(carcinogen_eset)$Factor.Value.disease.,
"_",
                      pData(carcinogen_eset)$Factor.Value.phenotype.))
samples_cutoff <- min(no_of_samples)
idx_thresh_median <- apply(exprs(carcinogen_eset), 1, function(x){
sum(x > thresh_median) >= samples_cutoff } )

table(idx_thresh_median)
#########################
##                     ##
##  idx_thresh_median  ##
##      TRUE           ##
##      41345          ##
##                     ##
#########################

carcinogen_filtered <- subset(carcinogen_eset, idx_thresh_median)

carcinogen_filtered_anno <- AnnotationDbi::select(mogene20sttranscriptcluste
r.db,
    keys=(featureNames(carcinogen_filtered)),
    columns = c("SYMBOL", "GENENAME"),
    keytype="PROBEID")
#####################################################################
##                                                                 ##
##  'select()' returned 1:many mapping between keys and columns  ##
##                                                                 ##
#####################################################################

carcinogen_filtered_anno <- subset(carcinogen_filtered_anno, !is.na(SYMBOL))
carcinogen_filtered_anno_grouped <- group_by(carcinogen_filtered_anno, PROBE
ID)
head(carcinogen_filtered_anno_grouped )
```

```
################################################################################
##                                                                            ##
##  A tibble: 6 x 3                                                           ##
##  Groups:   PROBEID [6]                                                     ##
##    PROBEID  SYMBOL  GENENAME                                              ##
##    <chr>    <chr>   <chr>                                                  ##
##  1 17210855 Lypla1  lysophospholipase 1                                    ##
##  2 17210869 Tcea1   transcription elongation factor A (SII) 1             ##
##  3 17210883 Gm16041 predicted gene 16041                                   ##
##  4 17210887 Atp6v1h ATPase, H+ transporting, lysosomal V1 subunit H        ##
##  5 17210904 Oprk1   opioid receptor, kappa 1                              ##
##  6 17210912 Rb1cc1  RB1-inducible coiled-coil 1                           ##
##                                                                            ##
################################################################################

dim(carcinogen_filtered_anno_grouped)
########################
##                    ##
##  [1] 36392     3  ##
##                    ##
########################

anno_summarized <- dplyr::summarize(carcinogen_filtered_anno_grouped,
                                    no_of_matches = n_distinct(SYMBOL))
anno_filtered <- filter(anno_summarized, no_of_matches > 1)

ids_to_exlude <- ((featureNames(carcinogen_filtered) %in% anno_filtered$PROB
EID) |
                    featureNames(carcinogen_filtered) %in% subset(carcinogen_
filtered_anno,
                    is.na(SYMBOL))$PROBEID)

carcinogen_final <- subset(carcinogen_filtered, !ids_to_exlude)
############################
##                        ##
## dim(carcinogen_final)  ##
## Features  Samples      ##
##   39724       15       ##
##                        ##
############################
fData(carcinogen_final)$PROBEID <- rownames(fData(carcinogen_final))
fData(carcinogen_final) <- left_join(fData(carcinogen_final), carcinogen_fil
tered_anno)
############################
##                        ##
##  Joining, by = PROBEID  ##
##                        ##
############################
```

```
rownames(fData(carcinogen_final)) <-fData(carcinogen_final)$PROBEID

########################################
##                                    ##
##  Análisis de expresión diferencial  ##
##                                    ##
########################################
carcinogen_exprs_final = exprs(carcinogen_final)
head(carcinogen_exprs_final)
fac_int <- pData(carcinogen_final)$Factor.Value.phenotype
design <- model.matrix(~0 + fac_int) ## Solamente funciona para un fenotipo
binario, si se tienen más fenotipos, cambiar los contrastes
print(design)
########################################
##                                    ##
##     fac_intControl fac_intTratado  ##
##  1               1              0  ##
##  2               1              0  ##
##  3               1              0  ##
##  4               0              1  ##
##  5               0              1  ##
##  6               0              1  ##
##  7               0              1  ##
##  8               0              1  ##
##  9               0              1  ##
##  10              0              1  ##
##  11              0              1  ##
##  12              0              1  ##
##  13              0              1  ##
##  14              0              1  ##
##  15              0              1  ##
##  attr(,"assign")                   ##
##  [1] 1 1                           ##
##  attr(,"contrasts")                ##
##  attr(,"contrasts")$fac_int        ##
##  [1] "contr.treatment"             ##
##                                    ##
########################################

colnames(design) <- unique(factor(fac_int))
######################################################
##                                                  ##
##  Se definen los grupos de acuerdo al fenotipo.   ##
##                                                  ##
######################################################
print(design)
###################################
##                               ##
##    Control Tratado            ##
##  1        1       0          ##
```

```
##  2        1       0          ##
##  3        1       0          ##
##  4        0       1          ##
##  5        0       1          ##
##  6        0       1          ##
##  7        0       1          ##
##  8        0       1          ##
##  9        0       1          ##
##  10       0       1          ##
##  11       0       1          ##
##  12       0       1          ##
##  13       0       1          ##
##  14       0       1          ##
##  15       0       1          ##
##  attr(,"assign")             ##
##  [1] 1 1                     ##
##  attr(,"contrasts")          ##
##  attr(,"contrasts")$fac_int  ##
##  [1] "contr.treatment"       ##
##                              ##
##################################
```

```
lmfit <- lmFit(carcinogen_final, design) ##Establecemos modelos lineales,
correlaciones lineales entre los contrastes
######################################################################
##                                                                  ##
##  Cambiar aquí los contrastes, si se tienen más fenotipos.  ##
##                                                                  ##
######################################################################
cont_tratado <- makeContrasts(Control-Tratado, levels = design)
print(cont_tratado)
##################################
##                              ##
##          Contrasts           ##
## Levels    Control - Tratado  ##
##   Control                1  ##
##   Tratado               -1  ##
##                              ##
##################################
```

```
lmfit.cont <- contrasts.fit(lmfit, cont_tratado)
lmfit.cont.ebayes <- eBayes(lmfit.cont)

print(topTable(lmfit.cont.ebayes))
#########################################################################################################################
##                                                                                                                     ##
##           PROBEID    SYMBOL                                                          GENENAME     logFC  ##
##  17335467 17335467    Cdkn1a                       cyclin-dependent kinase inhibitor 1A (P21) -4.471969  ##
##  17543396 17543396     Eda2r                                    ectodysplasin A2 receptor -2.911571  ##
##  17529764 17529764      Pls1                                        plastin 1 (I-isoform) -3.161245  ##
##  17261865 17261865     Ccng1                                                cyclin G1 -1.936880  ##
##  17484068 17484068      Lhpp phospholysine phosphohistidine inorganic pyrophosphate phosphatase  2.801670  ##
##  17534385 17534385     Gria3                  glutamate receptor, ionotropic, AMPA3 (alpha 3) -2.981105  ##
```

```
##  17289794 17289794      Plk2                                                  polo like kinase 2 -2.044308  ##
##  17462437 17462437     Usp18                                  ubiquitin specific peptidase 18 -1.835164  ##
##  17301697 17301697 Tnfrsf10b            tumor necrosis factor receptor superfamily, member 10b -3.165906  ##
##  17394538 17394538     Sulf2                                                 sulfatase 2 -1.933202  ##
##             AveExpr         t      P.Value    adj.P.Val        B                                          ##
##  17335467  9.736155 -30.75620 1.229233e-15 4.881530e-11 22.83271                                          ##
##  17543396  5.782974 -19.57152 1.394459e-12 2.768838e-08 18.07583                                          ##
##  17529764  6.825638 -18.53864 3.201208e-12 4.237545e-08 17.42081                                          ##
##  17261865 10.216770 -17.92845 5.336955e-12 5.298529e-08 17.00953                                          ##
##  17484068  6.565091  17.09573 1.100483e-11 7.467486e-08 16.41676                                          ##
##  17534385  6.806257 -17.06770 1.128246e-11 7.467486e-08 16.39614                                          ##
##  17289794  7.906152 -16.63134 1.670781e-11 8.986179e-08 16.06939                                          ##
##  17462437  7.291381 -16.45519 1.962887e-11 8.986179e-08 15.93433                                          ##
##  17301697  5.808938 -16.38732 2.089462e-11 8.986179e-08 15.88180                                          ##
##  17394538  7.426638 -16.30111 2.262837e-11 8.986179e-08 15.81467                                          ##
##                                                                                                         ##
#############################################################################################################
```

```
lmfit.cont.ebayes.table <- topTable(lmfit.cont.ebayes, number = Inf)
lmfit.cont.ebayes.table2 = subset(lmfit.cont.ebayes.table,
                            lmfit.cont.ebayes.table$GENENAME != "NA")

#####################################################################
##                                                                 ##
##  Genera el resumen de la expresion diferencial de los genes,  ##
##  ordenado del mas significativo al menos significativo        ##
##                                                                 ##
#####################################################################
Diff_expresed_genes = topTable(lmfit.cont.ebayes, num = Inf)
print(head(Diff_expresed_genes))
print(str(Diff_expresed_genes))
print(str(topTable(lmfit.cont.ebayes, num = Inf)))

print(str(Diff_expresed_genes))
#####################################################################
##                                                                 ##
##  Filtramos los mas significativos, se puede modificar el logFC,  ##
##  Ejemplo: 3, 4 y el adj.P.value (5% de significancia)           ##
##                                                                 ##
#####################################################################
Log_FC_Diff_expresed_genes = subset(Diff_expresed_genes, abs(logFC) > 2 & ad
j.P.Val < 0.05)

#####################################################################
##                                                                 ##
##  Hacemos un dataframe de los valores de expresión del carcinoma  ##
##                                                                 ##
#####################################################################
carcinogen_exprs_final_df = as.data.frame(carcinogen_exprs_final)

print(str(carcinogen_exprs_final_df))
#####################################################################
##                                                                 ##
##  Filtramos los datos de expresión con los genes más significativos,  ##
##  basado en el código de la sonda.                               ##
```

```
##                                                                    ##
########################################################################
carcinogen_exprs_final_dif_expres = carcinogen_exprs_final_df[
        rownames(carcinogen_exprs_final_df) %in% Log_FC_Diff_expresed_genes$
PROBEID,
        ]

################################################################
##                                                            ##
##  Reducimos el objeto a solo código de nombres y sondas.  ##
##                                                            ##
################################################################
Log_FC_Diff_expresed_genes_names = Log_FC_Diff_expresed_genes[, c(1:2)]

##############################################################################
##                                                                          ##
##  Se genera el mapa de calor de todos los genes, con el FC de interes.  ##
##                                                                          ##
##############################################################################
carcinogen_exprs_final_dif_expres_ord = subset(carcinogen_exprs_final_dif_ex
pres)
carcinogen_exprs_final_dif_expres_ord = carcinogen_exprs_final_dif_expres_or
d[
        rownames(carcinogen_exprs_final_dif_expres_ord) %in% Log_FC_Diff_exp
resed_genes$PROBEID,
        ]
carcinogen_exprs_final_dif_expres_ord$PROBEID = rownames(carcinogen_exprs_fi
nal_dif_expres_ord)

Log_FC_Diff_expresed_genes_merge_ord = merge(carcinogen_exprs_final_dif_expr
es_ord, Log_FC_Diff_expresed_genes_names)
Log_FC_Diff_expresed_genes_merge_final_ord = subset( Log_FC_Diff_expresed_ge
nes_merge_ord,
                                                     Log_FC_Diff_expresed_ge
nes_merge_ord$SYMBOL != "NA")
Log_FC_Diff_expresed_genes_merge_final2_ord = Log_FC_Diff_expresed_genes_mer
ge_final_ord[
        !duplicated(Log_FC_Diff_expresed_genes_merge_final_ord$SYMBOL),]

rownames(Log_FC_Diff_expresed_genes_merge_final2_ord) = Log_FC_Diff_expresed
_genes_merge_final2_ord$SYMBOL

Log_FC_Diff_expresed_genes_merge_final2_ord$PROBEID = NULL
Log_FC_Diff_expresed_genes_merge_final2_ord$SYMBOL = NULL

diff_matrix2_ord = as.matrix(Log_FC_Diff_expresed_genes_merge_final2_ord)

png("HeatmapFC_AllGenes.png")
```

```
heatmap.2(diff_matrix2_ord, trace= "none", col = colorRampPalette(c("green",
"black","red"))(255),
         margins = c(5,10), lwid = c(5,15), lhei = c(3,15), density.info = "
none" )
dev.off()

##############################################################################
##                                                                          ##
##  Calculamos la distribución de los cuartiles en el primer control,       ##
##  Nota: modificar el objeto al nombre de sus controles                    ##
##  Modificar los valores de corte dependiendo de los cuartiles de salida   ##
##                                                                          ##
##############################################################################
print(summary(carcinogen_exprs_final_dif_expres$PC181.CEL))
##############################################################################
##                                                                          ##
##  Se genera el mapa de calor del primer cuartil                           ##
##  Modificar los valores de corte dependiendo de los cuartiles de salida   ##
##                                                                          ##
##############################################################################
carcinogen_exprs_final_dif_expres_ord_1st = subset(carcinogen_exprs_final_di
f_expres,PC181.CEL <= 3.828)
carcinogen_exprs_final_dif_expres_ord_1st2 = carcinogen_exprs_final_dif_expr
es_ord_1st[
        rownames(carcinogen_exprs_final_dif_expres_ord_1st) %in% Log_FC_Diff
_expresed_genes$PROBEID,
        ]
carcinogen_exprs_final_dif_expres_ord_1st2$PROBEID = rownames(carcinogen_exp
rs_final_dif_expres_ord_1st2)

Log_FC_Diff_expresed_genes_merge_ord_1st = merge(carcinogen_exprs_final_dif_
expres_ord_1st2, Log_FC_Diff_expresed_genes_names)
Log_FC_Diff_expresed_genes_merge_final_ord_1st = subset( Log_FC_Diff_exprese
d_genes_merge_ord_1st,
                                                         Log_FC_Diff_exprese
d_genes_merge_ord_1st$SYMBOL != "NA")
Log_FC_Diff_expresed_genes_merge_final2_ord_1st = Log_FC_Diff_expresed_genes
_merge_final_ord_1st[
        !duplicated(Log_FC_Diff_expresed_genes_merge_final_ord_1st$SYMBOL),]

rownames(Log_FC_Diff_expresed_genes_merge_final2_ord_1st) = Log_FC_Diff_expr
esed_genes_merge_final2_ord_1st$SYMBOL

Log_FC_Diff_expresed_genes_merge_final2_ord_1st$PROBEID = NULL
Log_FC_Diff_expresed_genes_merge_final2_ord_1st$SYMBOL = NULL

diff_matrix2_ord_1st = as.matrix(Log_FC_Diff_expresed_genes_merge_final2_ord
_1st)

png("Heatmap_1st_cuartil.png")
```

```
heatmap.2(diff_matrix2_ord_1st, trace= "none", col = colorRampPalette(c("gre
en","black","red"))(255),
         margins = c(5,10), lwid = c(5,15), lhei = c(3,15), density.info = "
none"
       , Colv=FALSE )
dev.off()
##############################################################################
##                                                                          ##
##  Se genera el mapa de calor del segundo  cuartil                         ##
##  Modificar los valores de corte dependiendo de los cuartiles de salida.  ##
##                                                                          ##
##############################################################################

carcinogen_exprs_final_dif_expres_ord_2st = subset(carcinogen_exprs_final_di
f_expres,PC181.CEL > 3.828 & PC181.CEL < 6.423)
carcinogen_exprs_final_dif_expres_ord_2st2 = carcinogen_exprs_final_dif_expr
es_ord_2st[
        rownames(carcinogen_exprs_final_dif_expres_ord_2st) %in% Log_FC_Diff
_expresed_genes$PROBEID,
        ]
carcinogen_exprs_final_dif_expres_ord_2st2$PROBEID = rownames(carcinogen_exp
rs_final_dif_expres_ord_2st2)

Log_FC_Diff_expresed_genes_merge_ord_2st = merge(carcinogen_exprs_final_dif_
expres_ord_2st2, Log_FC_Diff_expresed_genes_names)
Log_FC_Diff_expresed_genes_merge_final_ord_2st = subset( Log_FC_Diff_exprese
d_genes_merge_ord_2st,
                                                         Log_FC_Diff_exprese
d_genes_merge_ord_2st$SYMBOL != "NA")
Log_FC_Diff_expresed_genes_merge_final2_ord_2st = Log_FC_Diff_expresed_genes
_merge_final_ord_2st[
        !duplicated(Log_FC_Diff_expresed_genes_merge_final_ord_2st$SYMBOL),]

rownames(Log_FC_Diff_expresed_genes_merge_final2_ord_2st) = Log_FC_Diff_expr
esed_genes_merge_final2_ord_2st$SYMBOL

Log_FC_Diff_expresed_genes_merge_final2_ord_2st$PROBEID = NULL
Log_FC_Diff_expresed_genes_merge_final2_ord_2st$SYMBOL = NULL

diff_matrix2_ord_2st = as.matrix(Log_FC_Diff_expresed_genes_merge_final2_ord
_2st)

png("Heatmap_2nd_cuartil.png")
heatmap.2(diff_matrix2_ord_2st, trace= "none", col = colorRampPalette(c("gre
en","black","red"))(255),
         margins = c(5,10), lwid = c(5,15), lhei = c(3,15), density.info = "
none"
       , Colv=FALSE )
dev.off()
```

```
#############################################################################
##                                                                         ##
##  Se genera el mapa de calor del tercer cuartil                          ##
##  Modificar los valores de corte dependiendo de los cuartiles de salida.  ##
##                                                                         ##
#############################################################################

carcinogen_exprs_final_dif_expres_ord_3st = subset(carcinogen_exprs_final_di
f_expres, PC181.CEL >= 6.423)
carcinogen_exprs_final_dif_expres_ord_3st2 = carcinogen_exprs_final_dif_expr
es_ord_3st[
        rownames(carcinogen_exprs_final_dif_expres_ord_3st) %in% Log_FC_Diff
_expresed_genes$PROBEID,
        ]
carcinogen_exprs_final_dif_expres_ord_3st2$PROBEID = rownames(carcinogen_exp
rs_final_dif_expres_ord_3st2)

Log_FC_Diff_expresed_genes_merge_ord_3st = merge(carcinogen_exprs_final_dif_
expres_ord_3st2, Log_FC_Diff_expresed_genes_names)
Log_FC_Diff_expresed_genes_merge_final_ord_3st = subset( Log_FC_Diff_exprese
d_genes_merge_ord_3st,
                                                         Log_FC_Diff_exprese
d_genes_merge_ord_3st$SYMBOL != "NA")
Log_FC_Diff_expresed_genes_merge_final2_ord_3st = Log_FC_Diff_expresed_genes
_merge_final_ord_3st[
        !duplicated(Log_FC_Diff_expresed_genes_merge_final_ord_3st$SYMBOL),]

rownames(Log_FC_Diff_expresed_genes_merge_final2_ord_3st) = Log_FC_Diff_expr
esed_genes_merge_final2_ord_3st$SYMBOL

Log_FC_Diff_expresed_genes_merge_final2_ord_3st$PROBEID = NULL
Log_FC_Diff_expresed_genes_merge_final2_ord_3st$SYMBOL = NULL

diff_matrix2_ord_3st = as.matrix(Log_FC_Diff_expresed_genes_merge_final2_ord
_3st)

png("Heatmap_3rd_cuartil.png")
heatmap.2(diff_matrix2_ord_3st, trace= "none", col = colorRampPalette(c("gre
en","black","red"))(255),
         margins = c(5,10), lwid = c(5,15), lhei = c(3,15), density.info = "
none"
        , Colv=FALSE )
dev.off()

#########################
##                     ##
##  Gráfico de Volcan  ##
##                     ##
#########################
```

```
png("volcano_plot.png")
with(lmfit.cont.ebayes.table2, plot(logFC, -log10(P.Value),
    pch=20, main="Volcano plot", xlim=c(-7.5,6), ylim=c(0,12)))
with(subset(lmfit.cont.ebayes.table2, adj.P.Val<.05 & abs(logFC)>2),
     points(logFC, -log10(P.Value), pch=20, col="red"))
with(subset(lmfit.cont.ebayes.table2, adj.P.Val<.05 & (logFC)< -2 ),
     points(logFC, -log10(P.Value), pch=20, col="green"), pCutoff = 10e-32)
abline(v = 2, col ="red")
abline(v = -2, col ="darkgreen")
abline(h = 2, col ="black")

with(subset(lmfit.cont.ebayes.table2, adj.P.Val<.05 & abs(logFC)>3),
    textxy(logFC, -log10(P.Value), labs=SYMBOL, cex=0.3, offset=0.5))

dev.off()

results = decideTests(lmfit.cont.ebayes)
#####################################
##                                 ##
##  TestResults matrix             ##
##             Contrasts           ##
##              Control - Tratado  ##
##    17200001                  0  ##
##    17200003                  0  ##
##    17200005                  0  ##
##    17200007                  0  ##
##    17200009                  0  ##
##  39707 more rows ...            ##
##                                 ##
#####################################

DE_genes_php_pc <- subset(lmfit.cont.ebayes.table, adj.P.Val < 0.05)$PROBEID

back_genes_idx <- genefilter::genefinder(carcinogen_final,
                  as.character(DE_genes_php_pc), method = "manhattan", scale
 = "none")

back_genes_idx <- sapply(back_genes_idx, function(x)x$indices)
back_genes <- featureNames(carcinogen_final)[back_genes_idx]
intersect(back_genes, DE_genes_php_pc)

gene_IDs <- rownames(lmfit.cont.ebayes.table)
in_universe <- gene_IDs %in% c(DE_genes_php_pc, back_genes)
in_selection <- gene_IDs %in% DE_genes_php_pc
all_genes <- in_selection[in_universe]
all_genes <- factor(as.integer(in_selection[in_universe]) == 1)
names(all_genes) <- gene_IDs[in_universe]
differentially_espresed_IDs = all_genes[all_genes == TRUE]
differentially_expresss_genes_expression = exprs(carcinogen_final)[
```

```
    rownames(exprs(carcinogen_final)) %in% c(names(differentially_espresed_I
Ds)),]

##########################################
##                                      ##
##  Se generan los diferentes archivos.  ##
##                                      ##
##########################################

write.table(file="DE_express_control_tratados.txt",
            differentially_expresss_genes_expression,
            quote=F, sep="\t")

write.table(file="TopTable_genes.txt",
            lmfit.cont.ebayes.table2,
            quote=F, sep="\t")

write.table(file="TopTable_genes_underexpressed.txt",
            subset(lmfit.cont.ebayes.table2$SYMBOL,
            lmfit.cont.ebayes.table2$adj.P.Val < 0.05 &
            lmfit.cont.ebayes.table2$logFC <= 2),
            quote=F)
write.table(file="TopTable_genes_overexpressed.txt",
            subset(lmfit.cont.ebayes.table2$SYMBOL,
            lmfit.cont.ebayes.table2$adj.P.Val < 0.05 &
            lmfit.cont.ebayes.table2$logFC >= 2),
            quote=F)
```