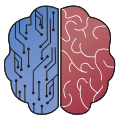
UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud



INGENIERÍA
DE LA SALUD

**TFG del Grado en Ingeniería de la Salud**

# Análisis bioinformático de los genes implicados en el cáncer hepático

Presentado por Ana Paula Cuesta Asín
en Universidad de Burgos

9 de julio de 2024

Tutores: Rubén Ruiz González – Antonia Maiara Marques do Nascimento

UNIVERSIDAD DE BURGOS

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de la Salud

D. Rubén Ruiz González, profesor del departamento de departamento, área de área.

Expone:

Que el alumno D. Ana Paula Cuesta Asín, con DNI 71795586M, ha realizado el Trabajo final de Grado en Ingeniería de la Salud titulado título del trabajo.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 9 de julio de 2024

Vº. Bº. del Tutor:

D. Rubén Ruiz González

Vº. Bº. del Tutor:

D. Antonia Maiara Marques do Nascimento

## Resumen

El carcinoma hepatocelular (CHC) es una de las formas más comunes y letales de cáncer de hígado. Este proyecto se centró en analizar la expresión diferencial de genes en CHC para identificar biomarcadores y dianas terapéuticas. Utilizando datos del conjunto GSE14520 de GEO, se emplearon herramientas bioinformáticas como GEOquery y limma en R. Se identificaron genes sobreexpresados en CHC, incluyendo ATXN7L3B, PRKAB2 y SLC23A2. Además, se analizaron variantes genéticas en CTNNB1 y TP53, utilizando Python y Clustal Omega para los alineamientos de secuencias. Estos hallazgos proporcionan una base sólida para futuras investigaciones y desarrollos clínicos en el tratamiento del CHC.



## Descriptores

Carcinoma Hepatocelular (CHC), Expresión Génica Diferencial, Biomarcadores, Dianas Terapéuticas, Bioinformática, GEOquery, limma, ATXN7L3B, PRKAB2, SLC23A2, CTNNB1, TP53, Alineamiento de Secuencias, Clustal Omega, Análisis de Variantes Genéticas.

## Abstract

Hepatocellular carcinoma (HCC) is one of the most common and lethal forms of liver cancer. This project focused on analyzing the differential gene expression in HCC to identify biomarkers and therapeutic targets. Using data from the GSE14520 dataset from GEO, bioinformatics tools such as GEOquery and limma in R were employed. Genes overexpressed in HCC were identified, including ATXN7L3B, PRKAB2, and SLC23A2. Additionally, genetic variants in CTNNB1 and TP53 were analyzed using Python and Clustal Omega for sequence alignments. These findings provide a solid foundation for future research and clinical developments in the treatment of HCC.



## Keywords

Carcinoma Hepatocelular (CHC), Expresión Génica Diferencial, Biomarcadores, Dianas Terapéuticas, Bioinformática, GEOquery, limma, ATXN7L3B, PRKAB2, SLC23A2, CTNNB1, TP53, Alineamiento de Secuencias, Clustal Omega, Análisis de Variantes Genéticas.

# Índice general

# Índice de figuras

# Índice de tablas

# Introducción

El carcinoma hepatocelular (CHC) es una de las formas más comunes y letales de cáncer de hígado, representando una significativa carga global de morbilidad y mortalidad. A pesar de los avances en las técnicas de diagnóstico y tratamiento, la tasa de supervivencia a largo plazo sigue siendo baja debido a la detección tardía y la naturaleza agresiva de la enfermedad. El desarrollo de estrategias más efectivas para el diagnóstico temprano y el tratamiento personalizado es crucial para mejorar los resultados clínicos en pacientes con CHC.

En este contexto, el análisis de la expresión diferencial de genes emerge como una herramienta poderosa para entender los mecanismos moleculares subyacentes al CHC. La identificación de genes que presentan diferencias significativas en su expresión entre tejidos tumorales y normales puede proporcionar valiosa información sobre los procesos biológicos que impulsan la carcinogénesis hepática. Estos genes diferencialmente expresados pueden servir como biomarcadores potenciales para el diagnóstico temprano, pronóstico y como dianas para nuevas terapias dirigidas.

Este proyecto se centra en el análisis de la expresión génica utilizando datos obtenidos del conjunto de datos GSE14520, disponible en la base de datos Gene Expression Omnibus (GEO). Este conjunto de datos proporciona perfiles de expresión génica de pacientes con CHC y tejidos hepáticos normales, permitiendo una comparación detallada y robusta. La metodología empleada incluye el uso de herramientas bioinformáticas avanzadas para la normalización de datos, análisis de expresión diferencial y anotación de genes.

Los objetivos principales de este estudio son:

Identificación de Biomarcadores: Detectar genes diferencialmente expresados que puedan servir como biomarcadores para el diagnóstico temprano y pronóstico del CHC. Descubrimiento de Nuevos Objetivos Terapéuticos: Identificar genes y vías metabólicas alteradas que puedan ser potenciales dianas para el desarrollo de nuevas terapias. Contribución al Conocimiento Científico: Aumentar la comprensión de los mecanismos moleculares subyacentes al CHC y proporcionar una base sólida para futuras investigaciones. El proyecto también abarca la implementación de metodologías integrativas que combinan diferentes tipos de datos omicos, como la metilación del ADN y la expresión génica, para obtener una visión más completa de los mecanismos moleculares involucrados en el CHC. Además, se presta especial atención a la validación experimental de los genes identificados, utilizando técnicas como la PCR en tiempo real (qPCR) y ensayos funcionales en líneas celulares.

En resumen, este proyecto tiene el potencial de aportar significativamente al campo de la oncología hepática, proporcionando nuevos conocimientos y herramientas para el diagnóstico y tratamiento del carcinoma hepatocelular. La combinación de análisis bioinformático y validación experimental busca asegurar que los resultados obtenidos sean robustos, reproducibles y clínicamente relevantes, contribuyendo al objetivo global de mejorar los resultados clínicos y la calidad de vida de los pacientes con CHC.

# Objetivos

El presente capítulo del trabajo recoge y explica de forma clara y concisa los distintos objetivos que se pretenden alcanzar con el desarrollo del actual proyecto.

## 2.1. Justificación del trabajo

En la actualidad, este tipo de cáncer presenta una elevada incidencia en la población, sobre todo en los países que se encuentran en vías de desarrollo. Este es el motivo por el cual es tan importante comprender los mecanismos moleculares subyacentes que conducen a la aparición de dicha enfermedad, así como el desarrollo de nuevas técnicas de diagnóstico más avanzadas y precisas, y nuevas estrategias de tratamiento más efectivas para la población afectada.

Para lograr sus objetivos, se lleva a cabo un análisis bioinformático de los distintos genes implicados en el cáncer hepático, tanto a través del estudio de las mutaciones más frecuentes que dan lugar a este tipo de patología como al estudio de la expresión diferencial de los distintos genes.

Por lo tanto, este trabajo tiene un gran potencial a la hora de identificar insights valiosos acerca de los mecanismos que presenta dicha enfermedad, a la hora de identificar nuevos biomarcadores y dianas terapéuticas y por último, a la hora de contribuir al desarrollo de tratamientos más efectivos y personalizados para una enfermedad que presenta una elevada incidencia. Pot todos estos motivos, podemos concluir que el proyecto tiene un gran potencial para impactar positivamente tanto la investigación básica como la práctica clínica en oncología hepática.

## 2.2. Objetivo principal del trabajo realizado

El objetivo principal del presente trabajo reside en la obtención de genes implicados en el cáncer hepático (CH) y su análisis bioinformático, para ver como afectan los mismos a la patogénesis de dicha enfermedad.

Sin embargo, para cumplir dicho objetivo principal es necesario dividirlo en objetivos de distinto tipo, que son aquellos que desarrollan a continuación.

### Objetivos marcados por los requisitos del Software/Hardware/Análisis

Entre los objetivos marcados por los requisitos del Software encontramos la utilización tanto de R como de sus paquetes disponibles para la descarga y el análisis de los datos relativos a la expresión génica, así como la utilizacion de Python y Jupyter Notebook para el estudio de las mutaciones que conllevan a la hepaotogénesis así como su procesamiento adicional y visualización de resultados.

Los principales objetivos de Hardware son garantizar la capacidad del equipo empleado para el procesamiento de grandes cantidades de datos sin tener problemas de rendimiento del mismo, así como suficiente almacenamiento y memoria para el manejo de dichos datos y asegurarse de que este Hardware es compatible tanto con el Software como con los sistemas operativos utilizados a lo largo del proyecto.

En cuanto a los objetivos de Análisis, estos son los más importantes y buscan la identificacion de biomarcadores, genes diferencialmente expresados en tejidos cancerígenos de hígado en comparación con los tejidos sanos, ya que pueden ser útiles para el diagnóstico temprano de los pacientes.

También encontramos la necesidad de relacionar cada uno de estos genes diferencialmente expresados con la función que desempeñan a través del empleo de bases de detos y como se puede ver esta afectada, así como la clasificación de los mismos bien como genes subexpresados o sobreexpresados.

Se busca obtener la secuencia de los genes cuyas mutaciones producen un proceso de carcinogénesis con el objetivo de identificar y descubrir nuevas dianas terapéuticas, puesto que los tratamientos actuales a menudo son ineficaces en etapas avanzadas de la enfermedad.

### Objetivos técnicos, relacionados con la calidad de los resultados

Estos objetivos técnicos buscan que los resultados del proyecto obtenidos presentan una alta calidad, que sean reproducibles por otros usuarios, así como que sean de utilidad para la comunidad científica, contribuyendo de manera significativa al conocimiento acerca del cáncer hepático y sus posibles tratamientos.

Para cumplir con estas premisas se pretende identificar con gran exactitud y precisión los genes diferencialmente expresados, lo cual se hará a través de la normalización de los datos, lo cual garantiza su comparabilidad entre muestras, así como a través de la validación de los resultados obtenidos mediante métodos estadísticos y bioinformáticos.

En cuanto a la reproducibilidad, el objetivo principal consiste en la publicación tanto de los scripts de R como de los notebooks de Jupyter para permitir a otros usuarios reproducir los método utilizados.

Otros objetivos técnicos que se persiguen son la paralelización en el uso de los recursos siempre y cuando sea posible para acelerar el análisis de los datos, así como utilizar los recursos tanto de hardware como de software eficientemente para evitar problemas, la representación clara y ordenada de los datos o el almacenaje de estos de una forma segura y confidencial.

Además, para asegurar una mayor calidad en los resultados obtenidos es muy interesante la integración de diferentes tipos de datos relativos al cáncer hepático, como datos de expresión diferencial, datos de mutaciones génicas, etc. para obtener una visión más completa de la enfermedad en el desarrollo del proyecto.

### Objetivos de aprendizaje

Los objetivos de aprendizaje están enfocados a un aprendizaje completo y práctico en el ámbito de la Bioinformática aplicada al estudio del cáncer hepático, por lo que, a la finalización del proyecto se espera haber obtenido un conjunto de habilidades tanto técnicas a la hora de analizar bioinformáticamente los genes implicados en un determinado tipo de cáncer, como teóricas, presentando un mayor grado de conocimiento acerca de esta enfermedad y de sus mecanismos de patogenogénesis, e incluso prácticas, ya

que se podrá poner en práctica el presente proyecto para la investigación del CH tanto en oncología como en medina personalizada y de precisión.

Para ello, es necesario cumplir los siguientes objetivos de aprendizaje a lo largo de la realización del proyecto:
1. Comprensión y entendimiento de los principios básicos y de las aplicaciones de la Bioinformática en el análisis de datos génicos, permitiendo adquirir los conceptos claves de esta disciplina tanto a nivel práctico como teórico.

2. Aprendizaje y desarrollo de habilidades en el campo de la programación. Se espera obtener un mayor manejo de distintos lenguajes de programación, como son Python o R, muy utilizados en esta disciplina, así como de los paquetes disponibles para estos lenguajes, entre los que se encuentan BioConductor y Biopython, que son los más importantes en esta disciplina.

3. Desarrollo del análisis de datos genómicos, aprendiendo a manejar y a trabajar con grandes volúmenes de datos, así como aprendiendo a utilizar herramientas de visualización de dichos datos.

Asimismo, se buscará aprender a enmarcar los datos dentro de su contexto específico, a interpretar los resultados obtenidos, a aplicar herramientas bioinformáticas a los datos con un propósito específico, etc.

# Conceptos teóricos

En el capítulo que se extiende a continuación se explican los conceptos teóricos básicos necesarios para la comprensión y entendimiento del actual proyecto de fin de grado. En este, sobre todo se exponen conceptos del ámbito biológico y sanitario relacionados con el cáncer hepático, puesto que es la temática principal de dicho trabajo.

## 3.1. El cáncer hepático

El cáncer hepático (HC), también conocido como cáncer de hígado o neoplasia hepática, es una enfermedad en la que los principales afectados son los hepatocitos, es decir, las principales células funcionales del hígado.

En este proceso oncológico, los hepatocitos sanos se malignizan, normalmente debido a una alteración genética, dando lugar a células cancerosas. Estas comienzan a multiplicarse y a proliferar sin control, eludiendo así los mecanismos de control del crecimiento y del ciclo celular. Esto, a su vez, conlleva la aparición de masas celulares anormales en alguna de las localizaciones anatómicas de dicho órgano, haciendo que se vea comprometida su función normal, en órganos sanos adyacentes o incluso en otras regiones del organismo, las cuales son alcanzadas por estas células neoplásicas bien a través del torrente sanguíneo o a través del sistema linfático. [National Cancer Institute, 2024]

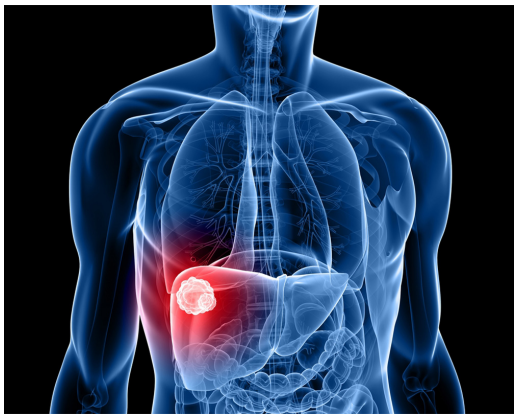


Figura 3.1: Representación de un proceso neoplásico en el hígado.
Fuente:[Fatty Liver Disease, 2024]

## Epidemiología

### Incidencia y prevalencia

El cáncer hepático, principalmente el carcinoma hepatocelular (CHC) ya que es la forma principal de este tipo de cáncer, es uno de los tipos de cáncer más comunes y mortales a nivel mundial. Según las estadísticas globales, el CHC ocupa el sexto lugar en incidencia entre todos los tipos de cáncer, con aproximadamente 905,677 nuevos casos diagnosticados en 2020. La incidencia varía significativamente entre diferentes regiones geográficas, siendo particularmente alta en Asia oriental y en el África subsahariana, debido a la prevalencia de factores de riesgo tales como la hepatitis B y C, y el consumo de alcohol, muy relacionados con el desarrollo de dicha enfermedad. [Srivatanakul P, Sriplung H, Deerasamee S,, 2004]

De hecho, en China se diagnostican el 50 por ciento de los casos a nivel mundial. Esto se debe al gran número de habitantes que presentan infecciones crónicas por el virus de la hepatitis B (VHB). Mientras tanto, en el África Subsahariana, con un elevado número de diagnósticos de CHC, la infección por el VHB es endémica, motivo por el que aumenta considerablemente el número de casos.

Sin embargo, en los países occidentales, las infecciones tanto por el VHB como por el virus de la hepatitis C (VHC) no son tan frecuentes, ya que se toman medidas para evitarlo. Sin embargo, la incidencia de esta enfermedad ha aumentado recientemente debido al incremento de la tasa de obesidad, la

diabetes y a la enfermedad hepática por hígado graso no alcóholico (NAFL), lo cual son todo factores de riesgo de esta neoplasia.

Por su parte, la prevalencia del cáncer hepático o de cualquier otro tipo de cáncer refleja el número total de personas que viven con la enfermedad en un momento dado. Por ejemplo, en el año 2020, se estimó que había aproximadamente 1.5 millones de personas viviendo con cáncer hepático a nivel mundial, lo cual se trata de una cifra muy elevada. [Bosch et al, 2004]

Sin embargo, la tasa de supervivencia de dicha enfermedad sigue siendo muy baja, ya que solo un 18 por ciento de los pacientes que sufren la enfermedad sobreviven a los 5 años siguientes. Esto se debe a que la mayor parte de los casos de esta enfermedad se diagnostican en etapas avanzadas, y el tratamiento disponible es ineficaz, así como la gran incidencia de esta enfermedad en países con un acceso a la sanidad limitado. Por eso, se hace tan importante una combinación del diagnóstico temprano de la misma junto a un tratamiento oportuno, con el objetivo de mejorar las actuales tasas de supervivencia de la enfermedad.

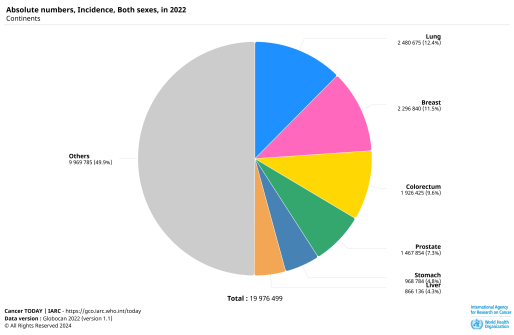



Figura 3.2: Distribución a nivel global de la incidencia de los distintos tipos de cáncer.
Fuente: [International Agency for Research on Cancer, 2024]

### Factores de riesgo

Como ya se ha mencionado en el apartado anterior, los principales factores de riesgo para el desarrollo de esta enfermedad son los siguientes:

- **Hepatitis B y C:** Las infecciones crónicas por los virus de la hepatitis B y C son los principales factores de riesgo para el CHC. La vacunación contra el VHB y el tratamiento antiviral para el VHC han demostrado reducir la incidencia de CHC.

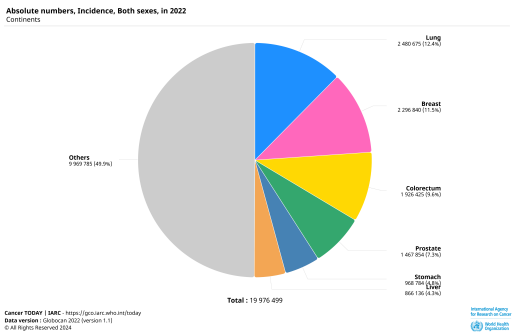



Figura 3.3: Distribución por continentes del cáncer hepático.
Fuente: [International Agency for Research on Cancer, 2024]

- **Consumo de Alcohol:** El consumo excesivo de alcohol puede llevar a cirrosis hepática, que es un importante factor de riesgo para el desarrollo de CHC.

- **Hígado Graso No Alcohólico (NAFLD):** La obesidad y el síndrome metabólico están asociados con NAFLD, que puede progresar a esteatohepatitis no alcohólica (NASH) y aumentar el riesgo de CHC.

- **Aflatoxinas:** La exposición a aflatoxinas, que son toxinas producidas por hongos presentes en alimentos contaminados, es un factor de riesgo significativo en algunas regiones de África y Asia.

## Clasificación

Las células cancerígenas destacan principalmente por su crecimiento descontrolado, por su gran capacidad invasora de tejidos sanos adyacentes, afectando a su funcionalidad, y por su facilidad de metastatizar, es decir, por su facilidad para colonizar localizaciones anatómicas lejanas al origen, llegando hasta ellas a través del sistema linfático y/o circulatorio.

En base a esta capacidad invasora, podemos clasificar las neoplasias hepáticas en dos grupos; los tumores primarios, cuyo origen se encuentra en el propio órgano, y los tumores secundarios o metastásicos. Estos últimos se originan en alguna otra región que no es el hígado, pero acaban alcanzando dicho órgano a través de un proceso metastásico.

Según la clasificación que realiza la Organización Mundial de la Salud (OMS), entre los principales cánceres con origen en las células hepáticas encontramos:

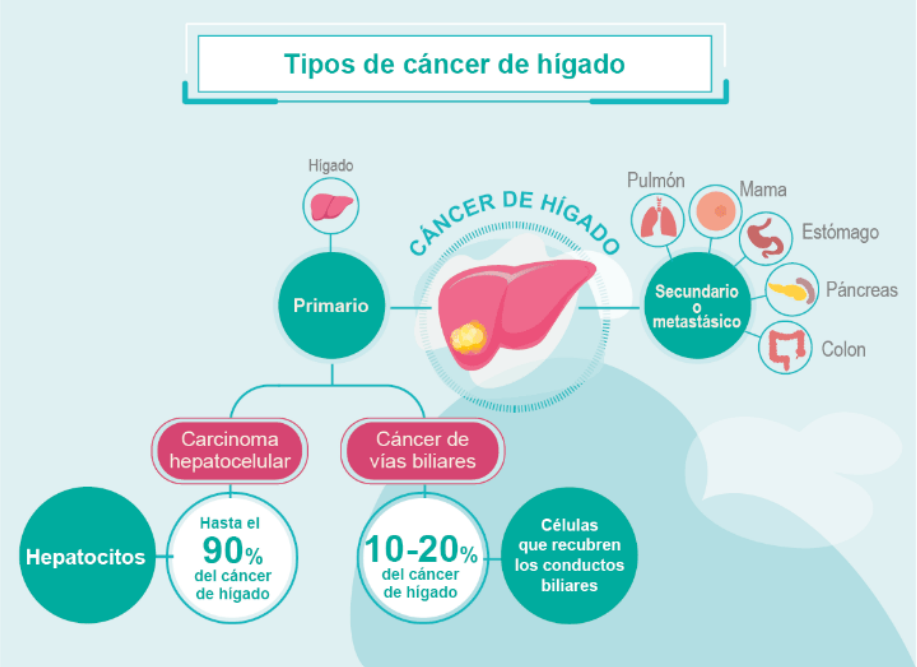



Figura 3.4: Clasificación del cáncer hepático según la OMS.
Fuente: [Roche Pacientes, 2024]

- **Carcinoma Hepatocelular (CHC):** Es el tumor hepático primario más habitual, con una incidencia de entre el 75 y el 85 por ciento de los casos, siendo más habitual en hombres que en mujeres. Este tipo de neoplasia afecta a las principales células funcionales del hígado, es decir, a los hepatocitos. Además, tiene una elevada tasa de mortalidad, ocupando el tercer puesto en cuanto al número de muertes debidas a procesos de índole oncológica. [National Cancer Institute, 2024][Mayo Clinic, 2024]

- **Colangiocarcinoma o cáncer de las vías biliares:** Este tipo de cáncer hepático primario es muy agresivo y se encuentra ocupando el segundo puesto en cuanto a su frecuencia de aparición, con una incidencia de entre el 10 y el 20 por ciento de los casos, siendo más común en mujeres que en hombres, sobre todo en la zona del sudeste asiático. En este proceso neoplásico se ven afectadas las células que conforman los conductos biliares, es decir, las vías que transportan la bilis desde el hígado hacia la vesícula biliar. A su vez, se puede distinguir entre colangiocarcinoma intrahepático y

extrahepático, en función de si los conductos afectados se encuentran en el interior o exterior de dicho órgano, respectivamente. [Burgos San Juan, 2008]

- **Hepatoblastoma:** Este tipo de cáncer tiene su origen en células hepáticas embrionarias, por lo que afecta generalmente a niños de corta edad, hasta los dos o tres años normalmente. De hecho, es el cáncer con mayor tasa de incidencia en la infancia, siendo muy rara su aparición en edades adultas.[Sharma, D., Subbarao, G., and Saxena, R, 2017]

- **Otros tipos:** Aunque estos sean los tumores más comunes en el hígado, existen más tipos con una tasa de incidencia mucho menor. Entre estos se encuentran el angiosarcoma y el hemangiosarcoma, que se originan en los vasos sanguíneos del hígado, o el cistoadenocarcinoma biliar, que también se origina en los conductos biliares, entre otros.

En cuanto a los cánceres secundarios que pueden afectar al hígado, encontramos una amplia multitud, puesto que pueden presentar su origen en casi cualquier región del organismo. Sin embargo, los más comunes son el cáncer de mama, el de pulmón, el colorrectal o el de próstata, debido a su proximidad con el órgano afectado. De hecho, la Sociedad Americana Contra el Cáncer establece que tanto en Europa como en Estados Unidos son más habituales este tipo de neoplasias que las primarias.[Roche Pacientes, 2024]

## Datos de los genes implicados en el cáncer hepático

Hemos concluido que los genes más estrechamente relacionados con el cáncer hepático primario debido a una mutación en su secuencia son los siguientes: - Gen TP53 - Gen CTNNB1 - Gen MUC16 - Gen CSMD3

Estos genes los hemos obtenido a partir de la base de datos TCGA (The Cancer Genome Atlas Program). Esta base de datos nos proporciona una herramienta entre sus funcionalidades, llamada "Mutation Frequency"la cual nos devuelve en forma de gráfico los genes con mayor implicación en un determinado tipo de cáncer, el cual hemos definido previamente. [The Cancer Genome Atlas, 2024]

La gráfica que nos devuelve esta herramienta se muestra a continuación.

Este gráfico ha sido realizado en base a la tasa de mutación de cada uno de los genes en una cohorte de estudio compuesta por 412 pacientes.

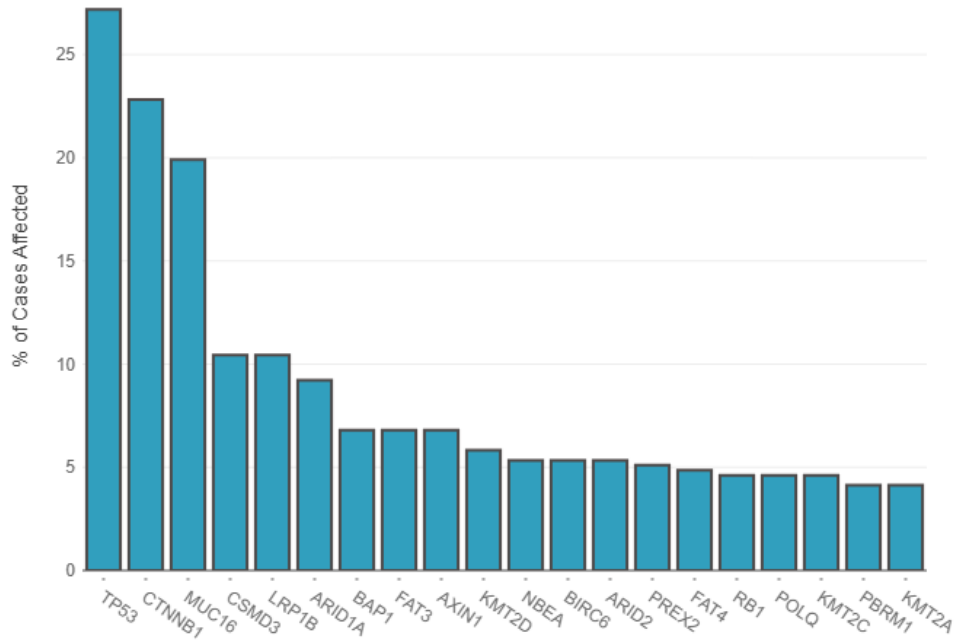



Figura 3.5: Pie de la figura de la figura bla bla bla

## 3.2. Estado del arte y trabajos relacionados.

En esta sección del capítulo vamos a hacer una revisión crítica de la bibliografía y literatura existente acerca de la temática del proyecto, exponiendo así los estudios más relevantes sobre la misma, así como los métodos y técnicas seguidos en estos y los hallazgos y resultados obtenidos tras su realización.

Para llevar a cabo esta revisión, se han buscado artículos relacionados con el cáncer hepático y la expresión diferencial de los genes implicados en este tipo de patología en los principales motores de búsqueda de artículos científicos, entre los que destacamos PubMed y Google Scholar ya que han sido los más utilizados para la obtención de la literatura competente.

En el estudio que describe el artículo *Identification of Hub Genes in Liver Hepatocellular Carcinoma Based on Weighted Gene Co-expression Network Analysis* publicado por la revista *Biochemical Genetics* se utilizan datos relativos a cáncer hepático procedentes tanto de la base de datos TCGA como de la serie de expresion en GEO *(Gene Expression*

*Omnibus)* GSE54236. A partir de estos datos, el artículo explica como se obtuvieron genes centrales (hub genes) en el carcinoma hepatocelular (CHC), es decir, genes que presentan un fuerte grado de asociación con dicha enfermedad. Además, también se analizaron los patrones de expresión de dichos genes, así como la variación de estos genes y de sus productos a lo largo de la progresión de dicha enfermedad y su influencia sobre la respuesta inmune del organismo. Para llevar a cabo dicho análisis se emplea la herramienta WGCNA, de la cual se habla en dicho artículo. [Jiawei Sun and Zizhen Zhang and Jiaru Cai and Xiaoping Li and Xiaoling Xu, 2024]

En el estudio acerca de los subtipos moleculares y los biomarcadores pronósticos que se describe en el artículo *Prognosis-Related Molecular Subtypes and Immune Features Associated with Hepatocellular Carcinoma* se estudian subtipos moleculares del hepatocarcinoma celular (CHC) a partir de datos expresados en cuatro conjuntos de datos bien diferenciados a través de análisis bioinformáticos que se realizaron de dichos datos. En la realización del mismo, se emplearon modelos estadísticos para identificar los genes que presentaban una mayor relación con este tipo de cáncer, pero para poder tener una aplicabilidad clñinica de los biomarcadores identificados es necesario obtener un mayor número de validaciones de los datos empleados en cohortes independientes. [Jiazhou Ye and Yan Lin and Xing Gao and Lu Lu and Xi Huang and Shilin Huang and Tao Bai ar

El artículo denominado *Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers* se realiza un análisis tanto de los patrones de metilación del ADN como de la expresión de los genes en el cáncer hepático a partir de datos que se encontraban almacenados en la base de TCGA con el objetivo de identificar biomarcadores específicos de este tipo de cáncer. En dicho estudio se consiguió la identificación de puntos de metilación diferencial del ADN, así como el hallazgo de genes expresados diferencialmente, aunque aún es necesaria la validación funcional y clínica de dichos biomarcadores para confirmar la utilidad diagnóstica de los mismos.[Hu, Hui and Li, Hong and Jiao, Feng and Han, Tao and Zhuo, Mei and Cui, Jian and Li, Y

En base a estos artículos, podemos concluir que la investigación acerca de la expresión diferencial de los distintos genes en el desarrollo del cáncer hepático ha avanzado significativamente gracias a las técnicas bioinformáticas que se van desarrollando en la actualidad, así como gracias a los grandes conjuntos de datos disponibles en repositorios de Internet. Sin embargo, para que estos hallazgos puedan acabar dando lugar a aplicaciones clínicas

efectivas, es esencial realizar validaciones funcionales y estudios clínicos adicionales.

# Metodología

En este capítulo se describen los datos empleados durante la realización del proyecto, se explican las diferentes técnicas y herramientas utilizadas a lo largo de su desarrollo y se detallan las metodologías seguidas con el objetivo de alcanzar las metas y objetivos propuestos al comienzo de este.

## 4.1. Descripción de los datos.

En esta sección del presente capítulo se incluye información relativa a los datos utilizados a lo largo de la elaboración del proyecto, incluyéndose así los procedimientos utilizados para su correspondiente obtención.

Para el desarrollo del proyecto, dada su índole y ámbito de aplicación, y en vista a cumplir con los objetivos propuestos, ha sido imposible la obtención de nuevas muestras biológicas de pacientes, motivo por el que se ha recurrido a archivos ya publicados en repositorios de acceso público de Internet, los cuales cuentan con datos que han sido previamente analizados y testados en otros laboratorios.

Estos datos los podemos dividir en dos grupos claramente diferenciados en base a su origen y a las herramientas utilizadas para su procesamiento:

| Tipos de datos | Origen (base de datos) | Herramienta utilizada |
|---|---|---|
| **Pacientes reales** | GEO | R |
| **Secuencia variantes** | Ensembl | Biopython |

Tabla 4.1: Clasificación de los datos utilizados en el proyecto.
Fuente: Elaboración propia

## Datos de pacientes reales

Los datos relativos a las muestras elegidas para estudiar las alteraciones en la expresión de ciertos genes, así como su influencia en la aparición y en el desarrollo del cáncer hepático han sido extraídos de la base de datos *Gene Expression Omnibus* (GEO), página web desarrollada y gestionada por el NCBI (*Centro Nacional de Información Biotecnológica de Estados Unidos*). [National Center for Biotechnology Information, 2024a]

Concretamente, se ha elegido la Serie GSE14520 para realizar dicho análisis.
Este conjunto de datos fue depositado en la base GEO a principios del año 2009, sin embargo, ha seguido modificándose hasta septiembre del pasado año 2023 y corresponde a un estudio llevado a cabo por la Unidad de Carcinogéneisis Hepática del Laboratorio de Carcinogénesis Humana del Instituto Nacional del Cáncer de Estados Unidos, situado en la ciudad de Bethedesa, al sur del condado de Montogomery, en el estado de Meryland, Estados Unidos. Los detalles relativos al desarrollo del proyecto se pueden observar en la Figura 4.1. [National Center for Biotechnology Information, 2024b]

| | |
|---|---|
| Submission date | Jan 22, 2009 |
| Last update date | Sep 12, 2023 |
| Contact name | Xin Wei Wang |
| E-mail(s) | xw3u@nih.gov |
| Phone | 240-760-6858 |
| Organization name | National Cancer Institute |
| Department | Laboratory of Human Carcinogenesis |
| Lab | Liver Carcinogenesis Unit |
| Street address | 37 Convent Drive |
| City | Bethesda |
| State/province | MD |
| ZIP/Postal code | 20892-4255 |
| Country | USA |

Figura 4.1: Resumen de los principales datos del proyecto utilizado.
Fuente: [National Center for Biotechnology Information, 2024b]

Esta serie recibe el nombre 'Gene expression data of human hepatocellular carcinoma (HCC)' y contiene información relativa a los mecanismos y frecuencia de expresión de los genes con mayor implicación en el cáncer de hígado, las funcionalidades que desarrollan los productos para los que codifican dichos genes, los patrones de metilaciones de estos, etc.

Estos datos e información relativa a las muestras se resume en matrices que se almacenan en archivos de texto en función de la plataforma o array

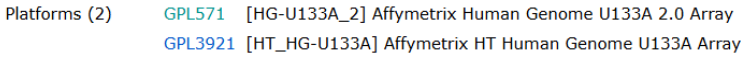



Figura 4.2: Plataformas utilizadas en el experimento de la Serie GSE14520 Fuente: [National Center for Biotechnology Information, 2024b]

empleado durante la realización del análisis. Concretamente, en este experimento se utilizan dos plataformas, GPL571 y GPL3921, cuyos resultados se almacenan en dos archivos de texto, "GSE14520GPL571seriesmatrix.txt.gzz "GSE14520GPL3921seriesmatrix.txt.gz", respectivamente.

Esta división de las 488 muestras del estudio en plataformas se realiza en base a las bibliotecas empleadas para la codificación de las sondas utilizadas durante el transcurso del experimento, es decir, la forma en que se codifican los nombres de los genes estudiados durante su realización.

Las plataformas utilizadas se pueden observar en la Figura 4.2 del proyecto.

### Plataforma GPL571

Esta plataforma, que recibe el nombre de microarray .Affymetrix Human Genome U133A Array 2.0", utiliza la biblioteca "HG-U133A2", la cual contiene la codificación de las sondas de oligonucleótidos utilizadas para la expresión de los genes y el estudio de su influencia en el cáncer hepático.

En este array se han analizado un total de 43 muestras que se reparten en 19 muestras procedentes de tejido normal hepático, 22 muestras procedentes de tejidos hepáticos cancerígenos y 2 muestras de tejido hepático procedentes de individuos sanos. Sin embargo, como solo contamos con dos muestras de este tipo, los resultados que obtengamos acerca de los genes expresados en este caso no van a ser representativos de este tipo de neoplasia, por lo que no los vamos a utilizar en el desarrollo del proyecto.

Una pequeña representación de estas muestras, junto a sus valores de expresión se puede observar en la Figura 4.3.

### Plataforma GPL3921

La plataforma identificada como GPL3921 de forma única en la base de datos GEO hace referencia al microarray denominado 'Affymetrix HT

```
> resultados_GPL571
                logFC  AveExpr         t      P.Value   adj.P.Val          B
222358_x_at  0.16467793 2.673542  6.117073 2.331156e-07 0.005191019  6.69200317
219224_x_at  0.11468197 2.716689  5.111134 6.782941e-06 0.075521268  3.67890940
204630_s_at  0.09116538 3.033408  4.647179 3.100474e-05 0.230137828  2.31752652
207283_at    0.17370077 2.319945  4.513388 4.773313e-05 0.231309364  1.93124456
208741_at    0.14630270 2.498121  4.432847 6.178385e-05 0.231309364  1.70037647
211387_x_at  0.13878681 2.653036  4.383618 7.228851e-05 0.231309364  1.55992885
205583_s_at  0.13348253 2.644939  4.369351 7.564653e-05 0.231309364  1.51932411
204181_s_at  0.16538665 2.468969  4.304092 9.305361e-05 0.231309364  1.33417303
206317_s_at -0.12410346 2.097020 -4.276667 1.014851e-04 0.231309364  1.25665857
213850_s_at  0.07479374 3.070251  4.259233 1.072282e-04 0.231309364  1.20747300
```

Figura 4.3: Representación de los valores de expresión de los genes analizados con el array GPL571. Fuente: Código realizado en R

```
> resultados_GPL3921
                   logFC  AveExpr          t      P.Value    adj.P.Val        B
212952_at        0.14829347 3.075325  12.041052 4.046552e-29 6.040511e-25 55.32281
215259_s_at     -0.12271637 1.770177 -11.970273 7.728971e-29 6.040511e-25 54.68675
208461_at       -0.10707347 1.744662 -11.964624 8.137926e-29 6.040511e-25 54.63607
208522_s_at     -0.10336801 1.748434 -11.830054 2.769561e-28 1.541815e-24 53.43233
217422_s_at     -0.09811215 1.817244 -11.726259 7.088784e-28 3.157061e-24 52.50869
208495_at       -0.09553109 1.751604 -11.545482 3.605726e-27 1.338205e-23 50.91027
207592_s_at     -0.11434359 1.852693 -11.470811 7.032354e-27 2.237092e-23 50.25391
AFFX-TrpnX-M_at -0.07817878 1.709391 -11.399389 1.329378e-26 3.700047e-23 49.62826
207588_at       -0.07274004 1.648659 -11.386160 1.495438e-26 3.700047e-23 49.51261
```

Figura 4.4: Representación de los valores de expresión de los genes analizados con el array GPL3921. Fuente: Código realizado en R

Human Genome U133A Array' emplea la biblioteca "HTHGU133A"para la codificación de las sondas de oligonucleótidos que han sido utilizadas para el estudio de la expresión de los distintos genes tanto en tejidos sanos como en tejidos con cáncer hepático.

El archivo que contiene las muestras analizadas en esta plataforma presenta un total de 445 muestras, las cuales se dividen en un total de 220 muestras de tejido hepático sano y 225 muestras de tejido cancerígeno. Por lo tanto, como tenemos un gran número de muestras que han sido analizadas con esta plataforma, los resultados que obtengamos si que pueden tener una cierta validez clínica a la hora de establecer los genes que se expresan diferencialmente en el cáncer de hígado, ya que utiliza una cantidad de pacientes bastante representativa de la población.

Una pequeña representación de estas muestras, junto a sus valores de expresión se puede observar en la Figura 4.4.

### Datos de las variantes

Además de la expresión diferencial de los genes, en el presente proyecto también se van a crear una serie de Notebooks de Python que van a contener las funciones necearias para a partir del identificador de un gen concreto en la base de datos de Ensembl, poder obtener sus variantes tanto patogénicas como benignas, acompañadas de la secuencia de estas. Además, se incluye la función que realiza el alineamiento de las mismas a través de la herramienta Clustal Omega y una función que representa dicho alineamiento para poder identificar regiones con un mayor número de variaciones y regiones mucho más conservadas. [Ensembl Genome Browser, 2024]

## 4.2. Herramientas

A continuación, en este apartado del capítulo "Metodología"vamos a detallar brevemente cuales han sido las herramientas o recursos elegidos para la realización del proyecto, así como el motivo para su empleo en el mismo.

### Lenguajes de programación

#### Programación en Python

Python se trata de un lenguaje de programación que fue desarrollado a principios de los años 90 por Guido van Rossum. De hecho, su primera versión fue lanzada al mercado en el año 1991.


Figura 4.5: Lenguaje de programación Python.
Fuente:[Learners' Galaxy, 2024]

Este lenguaje se caracteriza principalmente por tratarse de un lenguaje de alto nivel, ya que presenta una sintaxis sencilla y fácilmente legible, la

cual se acerca considerablemente al lenguaje humano natural. Asimismo, presenta un tipado de los datos tanto fuerte como dinámico, lo que confiere al usuario una mayor flexibilidad a la hora de realizar la asignación de los tipos a los datos. Cabe destacar que se trata de un lenguaje orientado a objetos, por lo que en su núcleo de funcionamiento se encuentra la creación de clases y objetos para alcanzar las metas deseadas. Sin embargo, también soporta paradigmas de programación tanto imperativa como funcional, siendo esta última la más ampliamente utilizada. Además, se trata de un lenguaje interpretado, cuyo código se ejecuta línea por línea, lo que facilita la detección y correción de posibles errores. [Python Software Foundation, 2024b]

En los años transcurridos desde su primer lanzamiento se han ido obteniendo nuevas versiones de dicho lenguaje, incluyéndose mejoras y nuevas funcionalidades en cada una de ellas. Entre estas se encuentran la versión Python 2 y la versión Python 3, esta última es la que se encuentra actualmente en uso. El paso de Python 2 a Python 3 fue un antes y un después en la utilización de este lenguaje de programación, puesto que se vio ampliamente mejorada con su aparición tanto la consistencia, como la eficacia de los productos de programación que utilizaban este lenguaje, hecho que condujo al fin de la vida útil de la versión Python 2 en enero de 2020. Las diferencias entre ambas versiones se pueden ver claramente en la Figura 4.3.

Las características que ya hemos mencionado, junto a la gran cantidad de documentación e información que encontramos para este lenguaje en foros de discusión, tutoriales, etc. y a la gran variedad de librerías de las que dispone, hicieron que se posicionará como el lenguaje de programación más utilizado a nivel mundial por el Índice de Popularidad de Lenguajes de Programación (PYPL) en el año 2019, tal y como podemos ver en la Figura 4.4. [Statista, 2024]

Todo lo que se ha expuesto en referencia a dicho lenguaje de programación ha sido decisivo para la elección de su utilización en el desarrollo del presente proyecto a la hora de obtener las secuencias de las variantes de los genes implicados en el cáncer hepático. Este lenguaje de programación ha sido utilizado a través de la plataforma de Anaconda, mediante la realización de funciones en cuadernos de Jupyter Notebook utilizando las bibliotecas disponibles en este lenguaje.

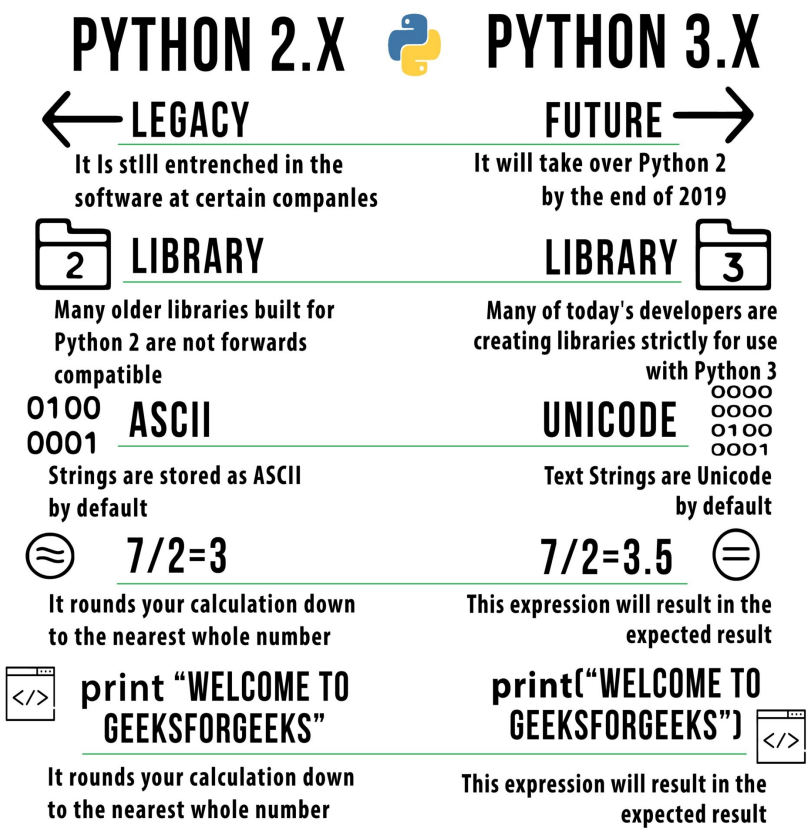



Figura 4.6: Principales diferencias entre Python 2 y Python 3.
Fuente: [Códigos Python, 2024]

## Programación en R

El lenguaje de programación R, al igual que Python, fue desarrollado a principios de los años 90. Este lenguaje, concretamente, fue desarrollado en los laboratorios Bell a cargo de Ross Ihaka y Robert Gentleman a partir de un lenguaje ya existente en esa época, conocido con el nombre de lenguaje S.

Este lenguaje se caracteriza tanto por su utilidad en programación, como por tratarse de un software dedicado al análisis estadístico y gráfico de datos, gracias a la gran variedad de herramientas de visualización con las que cuenta. Estos aspectos le han hecho ganar mucha popularidad desde su aparición, aspecto que conlleva que sea utilizado en una gran multitud de campos,


Figura 4.7: Lenguajes de programación más utilizados en 2019 según PYPL. Fuente: [Statista, 2024]

Figura 4.8: Lenguaje de programación R. Fuente: [UNIR, 2024]

entre los que destaca su aplicación a la investigación.[R Project, 2024]

En cuanto a su funcionamiento a nivel programático, cabe destacar que se trata de un lenguaje orientado a objetos, por lo que basa su funcionamiento en la creación de clases y objetos, el tipado dinámico que presenta, facilitando la asignación de tipos a los datos que se procesan utilizando dicho lenguaje. También es importante añadir que se trata de un lenguaje de alto nivel, presentando un bajo grado de abstracción, lo cual le hace fácilmente comprensible ante los usuarios de dicho lenguaje o el público en general.

A estas particularidades se le suma la gran comunidad de usuarios que utiliza dicho lenguaje de programación, la cual proporciona una gran cantidad de recursos referidos a este lenguaje, como documentación, tutoriales, soporte técnico, etc. Además, cuenta con una gran cantidad de bibliotecas y funcionalidades, que permiten hacer desde representaciones o análisis estadísticos sencillos, hasta procesos de evaluación y visualización de gran complejidad.[RDocumentation, 2024]

Por lo tanto, podemos concluir que se trata de un lenguaje de programación con un potencial enorme y una amplia gama de funcionalidades y bibliotecas, motivo por el cual ha sido elegido para la realización del presente proyecto, concretamente en el análisis de la expresión diferencial de los diferentes genes en el cáncer hepático.

### Entorno de desarrollo. RStudio

RStudio es un entorno integrado de desarrollo (IDE) que sirve para programar con el lenguaje de R, aunque también permite la programación con otros lenguajes como son Python, SQL o Stan. [RStudio, PBC, 2024a]


Figura 4.9: RStudio. Entorno de desarrollo del lenguaje de programación R. Fuente: [Hertie Coding Club, 2024]

Este, cuenta con una interfaz amigable y poderosa, puesto que cuenta con una gran cantidad de herramientas que permiten tanto la personalización del código desarrollado, como una mayor facilidad a la hora de leer y escribir este, ya que cuenta con un ayudante de código que te ayuda a programar de una forma más rápida y efectiva, evitando el cometer errores topográficos.[RStudio, PBC, 2024b]

Además, también cabe destacar el gran potencial que tiene como herramienta para la visualización de datos y la gestión de proyectos desarrollados con R.

Este IDE cuenta con una interfaz de usuario muy intuitiva, en la que se combinan varios paneles que representan de forma separada la consola de comandos, el script de código, los gráficos y productos generados y el espacio de trabajo, lo cual permite mantener el espacio de trabajo organizado, así como facilitar su uso. [RStudio, PBC, 2024c]

Asimismo también destaca por la gran cantidad de herramientas de visualización de datos que alberga, que permite a los usuarios de este IDE la realización de gráficos y visualizaciones de los datos interactivas. Por ejemplo, las bibliotecas más utilizadas de representación de datos son *ggplot2* o *plotly*, así como herramientas para crear aplicaciones web interactivas, como es *shiny*.

Estas características hacen que RStudio sea una herramienta que presenta un valor invaluable en el campo de la bioinformática. Esto se debe a la gran capacidad que presenta este IDE para manejar y analizar volúmenes de datos biológicos o genómicos de gran tamaño de una manera muy eficiente y reproducible, así como la visualización y representación o presentación de los mismos; motivo por el que hemos decidido realizar nuestro proyecto utilizando dicha herramienta.[RStudio, PBC, 2024d]

## Bibliotecas

Las bibliotecas, también conocidas como 'librerías' o 'módulos' son fragmentos de código organizados en forma de funciones ya predifinidas que aumentan las capacidades del lenguaje base de programación.

Cada biblioteca contiene funciones que realizan acciones interrelacionadas entre sí, puesto que el objetivo de su diseño es la ejecución de tareas específicas, tales como pueden ser operaciones matemáticas avanzadas, interacción con bases de datos, manipulación de archivos, manejo de secuencias biológicas, etc. [Abalozz, 2024]

De esta forma, se ahorra tiempo a los programadores desarrollando nuevas funciones, pudiendo ejecutar tareas complejas de una forma mucho más eficiente. En lugar de eso, se utilizan estas bibliotecas que hemos mencionado, las cuales contienen funciones predifinidas por otros desarrolladores.

| Lenguaje R | Lenguaje Python |
|---|---|
| Biopython | GEOquery |
| requests | limma |
| json | 'hgu133a2.db' |
| sys | 'hgu133a.db' |

Tabla 4.2: Bibliotecas utilizadas en el proyecto.
Fuente: Elaboración propia

Asimismo, aumenta la modularidad del código, ya que permite la división de los programas de mayor complejidad en pequeñas partes de código, facilitando su organización y aumentando la claridad del mismo.

Por último, cabe destacar que para la utilización de dichas bibliotecas en un programa es necesaria su importación dentro del código fuente de este. La forma de importación varía en función del lenguaje de programación. Por ejemplo, en R se utiliza el comando ïnstall"mientras que en Python se utiliza ïmport".

En la Tabla 4.2 aparecen las bibliotecas que se han empleado durante la realización del proyecto.

### Biopython

Biopython, aunque se presente como una biblioteca, realmente es un proyecto surgido en 1990 para mediante la colaboración de distintos programadores, obtener un conjunto de herramientas bioinformáticas de código abierto en Python que permitan realizar multitud de acciones relacionadas con el campo de la Bioinformática, que actualmente se encuentra en auge.[Chapman, Brad and Chang, Jeffrey, 2000]

Esta herramienta bioinformática contiene una gran cantidad de módulos que la permiten presentar un variado rango de funcionalidades en el campo de la bioinformática, tales como la creación de secuencias de ADN, ARN o proteínas, el análisis de estas secuencias creadas o de otras importadas desde algún formato de archivo aceptado, alineamiento tanto pareado como múltiple de secuencias, trabajar con la estructura de proteínas o ácidos nucleicos, acceder de una forma sencilla a los datos contenidos en bases de datos tales como el NCBI, PubMed, Genbank, etc. [Biopython Contributors, 2024a]


Figura 4.10: Biblioteca Biopython. Manejo de secuencias biológicas.
Fuente: [Wikipedia, 2024]

Entre la amplia gama de herramientas que contiene Biopython, nosotros hemos utilizado las que se detallan a continuación, encontrándose la mayoría de ellas en Bio, el paquete principal de la herramienta Biopython:

- **Módulo "Seq":** Es el módulo principal y más utilizado en Biopython para trabajar con secuencias de carácter biológico. Las principales acciones que se pueden realizar con este módulo son la lectura, escritura, manipulación y análisis de estas secuencias. Estas pueden ser de distinta naturaleza, y no hace falta especificar el tipo, sino que el intérprete detecta automáticamente si se trata de ADN, ARN, proteínas, o incluso ácidos nucleicos. [Biopython Contributors, 2024b]

- **Módulo "SeqIO":** Es un módulo ampliamente utilizado en Biopython. Este sirve tanto para la lectura, como para la escritura de secuencias en archivos de texto soportados para este tipo de secuencias, como son FASTA o GenBank, entre otros. [Biopython Contributors, 2024d]

- **Módulo "SeqRecord":** Este módulo permite tanto la representación como el manejo de registros de secuencias biológicas.En cada uno de estos registros encontramos una secuencia biológica de tipo Seq, junto a los metadatos relativos a dicha secuencia tales como el identificador (id), la descripción (description), anotaciones relativas a la secuencia (annotations) o incluso información adicional para el manejo de dicha secuencia (features). [Biopython Contributors, 2024c]

En la Figura 4.11 podemos observar la forma de importación de estos módulos en el entorno de Biopython de nuestro proyecto para el manejo de las secuencias biológicas con las que vamos a trabajar, concretamente las variantes de ciertos genes.

```
In [ ]: from Bio import SeqIO
        from Bio.Seq import Seq
        from Bio.SeqRecord import SeqRecord
```

Figura 4.11: Importación de los módulos de Biopython. Fuente: Código de Python

**requests**

Esta biblioteca de Python sirve para la realización de solicitudes HTTP a bases de datos. Estas solicitudes se utilizan para comunicarse con las APIs de las páginas web y la obtención de información o datos a partir de estas, o bien para interactuar para los recursos disponibles en esta. [Python Software Foundation, 2024c]

Esta librería ha sido elegida para utilizarla en el proyecto, en comparación con otras librerías estándar de Python como puede ser *urllib* debido a su simplicidad y facilidad de uso. Concretamente, se pone en uso para la realización de solicitudes GET a la base de datos Ensembl con el objetivo de obtener información y datos relativos a las variantes que incluyamos en la solicitud en forma de parámetros. Así como para el manejo de las solicitudes POST que nos devuelve dicha página web en forma de archivos de texto (.txt) o de archivos json.[Requests: HTTP for Humans, 2024]

**json**

Esta biblioteca ha sido creada con el propósito de trabajar con datos que se encuentren representados en formato JSON (JavaScript Object Notation). Este formato de datos es de alto nivel, siendo fácilmente accesible, puesto que su forma de lectura y escritura es bastante sencilla.[Python Software Foundation, 2024a]

Estas características hacen que sea un formato idóneo para su utilización en aplicaciones web, así como para recabar información y datos a partir de las APIs de bases de datos.

Además, es capaz de manejar excepciones tales como JSONDecodeError o JSONEncodeError a la hora de trabajar con la información almacenada en la base de datos.

Por lo tanto, este formato de archivos ha sido elegido en la realización del presente proyecto como herramienta utilizada a la hora de recabar información almacenada en distintas bases de datos.

#### sys

La biblioteca sys en Python se trata de una biblioteca que almacena funciones y variables que permiten acceder de una forma fácil y directa tanto al intérprete de Python como al entorno del sistema operativo subyacente.[Python Software Foundation, 2024d].

Por lo tanto, es esencial utilizar esta biblioteca con sumo cuidado, ya que si cometemos algún fallo a la hora de emplear dicha biblioteca podemos ver alterado bien el comportamiento del programa, o incluso del sistema subyacente.

### GEOquery

GEOquery se trata de una biblioteca creada como parte del proyecto BioConductor, iniciado por Gentleman y otra serie de desarrolladores en el año 2004. Dicho proyecto se desarrolló utilizando el lenguaje de programación R con el objetivo de permitir a los usuarios de este lenguaje analizar y comprender la gran cantidad de datos genómicos que se obtienen en la actualidad, así como el análisis de los microarrays que se llevan a cabo con los datos de esta índole. Además, dicho proyecto se caracteriza por ser tanto de código como de desarrollo abierto, lo cual permite la colaboración de diferentes programadores y desarrolladores.

Por otro lado, tenemos la base de datos GEO, que pertenece al NCBI, en la que se almacenan datos experimentales obtenidos a partir de una amplia variedad de organismos de todo tipo, tejidos o incluso diferentes estados de una enfermedad obtenidos a partir de aproximadamente 140.000 experimentos en los que se analiza la expresión de diferentes genes. Estos datos se almacenan en archivos de distinto tipo: datos de expresión génica (GSE), muestras individuales (GSM) y plataformas (GPL).

Por lo tanto, como parte de este proyecto BioConductor se ha desarrollado la biblioteca GEOquery, una herramienta que permite el acceso a la

información relativa a los análisis de expresión que se almacenan en la base de datos GEO directamente desde un entorno de desarrollo que use R como lenguaje de programación y que tenga el paquete BioConductor instalado. Es decir, esta biblioteca establece un enlace muy eficaz entre GEO y Bioconductor con el objetivo de analizar y meta-analizar datos de carácter genómico y sus perfiles de expresión. [Davis, Sean and S. Meltzer, Paul , 2007]

Esta biblioteca, por consiguiente, nos facilita considerablemente los siguientes aspectos a la hora de acceder a los datos de la plataforma GEO a través de BioConductor.

- A la hora de realizar la descarga de los datos se realiza de una manera mucho más sencilla, necesitando solo el identificador único del conjunto de datos que queremos obtener. De esta forma, evitamos la complejidad de descargar los datos directamente desde la base de datos de GEO, así como su posterior carga en el IDE de R. [Davis, Sean and S. Meltzer, Paul , 2007]

- El manejo de los datos obtenidos de GEO, ya que cuando se realiza la descarga de los mismos, estos se guardan en formatos fáciles de manipular y analizar dentro del entorno de R. Estos datos suelen encontrarse mayoritariamente en forma de DataFrames o ExpressionSet.

Por lo tanto, en el desarrollo del proyecto, esta biblioteca nos permite enfocarnos de una forma más directa en el análisis de los datos, ahorrando tiempo en lo relativo a su obtención, formateo y preparación. Dicho motivo es por el que se ha elegido la utilización de la biblioteca en el proyecto.

#### limma

La palabra limma es la abreviatura de Linear Models for Microarrat Data y corresponde a una biblioteca de BioConductor que se utiliza muy a menudo en el entorno de R para el análisis de datos procedentes de la expresión de distintos genes, puesto que fue creada con el objetivo de manejar y analizar los experimentos realizados con microarrays y secuenciación de RNA. [Gordon K. Smyth and others, 2024]

Entre sus funcionalidades destacan la normalización de los datos mediante modelos lineales para que los resultados y comparaciones entre las distintas muestras sean de gran precisión. También realiza análisis estadístico de estos a través de métodos robustos que identifican genes diferencialmente expresados entre dos condiciones diferenciadas utilizando pruebas de carácter

estadístico como el test de moderación de T de Bayes y la visualización de los resultados obtenidos a partir de estos, en forma de gráficos o volcanos principalmente. [Smith G. K., Data]

Por lo tanto, este paquete ha sido de gran utilidad en el actual proyecto, ya que su principal objetivo residía en la identificación de genes diferencialmente expresados entre dos condiciones, el tejido hepático sano y el tejido hepático canceroso.

### 'hgu133a2.db'

La biblioteca 'hgu133a2.db' se encuentra almacenada en el paquete *AnnotationDbi*, el cual se encuentra formando parte del ecosistema Bioconductor de R.

Esta biblioteca recopila las anotaciones y metadatos relacionados con el microarray Affymetrix HGU133, array con el que se realiza el análisis de la expresión génica de las muestras presentes en nuestro estudio. Entre la información que se encuentra almacenada en dicha plataforma encontramos detalles acerca de los genes, las sondas u otras características relativas a dicha plataforma. Esto incluye, entre otras cosas, los identificadores de los genes, la descripción funcional de estos y sus localizaciones genómicas.[National Center for Biotechnology Information, 2024b]

En el presente proyecto esta biblioteca es la encargada de codificar las sondas que se analizan en el microarray Affimetrix HGU133, el cual se identifica en el experimento concreto como la plataforma GPL571.

### 'hgu133a.db'

Al igual que la biblioteca 'hgu133a2.db' que acabamos de describir, esta biblioteca se encuentra almacenada en *AnnotationDbi*, paquete que integra el ecosistema Bioconductor de R.

En esta biblioteca también se almacenan los identificadores de los genes, la localización de estos y las funciones que desempeñan los genes de los que forman parte las sondas de oligonucleótidos que se utilizan en los experimentos llevados a cabo en el microarray Affymetrix HT Human Genome U133A Array. Además, también podemos encontrar información acerca de

los metadatos y las anotaciones relativas a la utilización de dicho array. [National Center for Biotechnology Information, 2024b]

En el proyecto desarrollado, esta biblioteca es la que se utiliza para codificar las sondas de la plataforma GPL3921, que corresponde al microarray HT HGU133.

### Clustal Omega

Clustal Omega se trata de una herramienta bioinformática ampliamente utilizada para realizar alineaciones de secuencias múltiples (MSA, en inglés).

Esta herramienta esta diseñada con el objetivo de alinear múltiples secuencias procedentes bien de secuencias de ADN, ARN o proteínas. Tal y como ya hemos indicado sirve para identificar regiones de similitud entre las secuencias pasadas como argumento, por lo tanto, podemos identificar relaciones bien funcionales, estructurales o evolutivas entre las distintas secuencias empleadas en el alineamiento. [Sievers, Fabian and Wilm, Andreas and Dineen, David and Gibson,

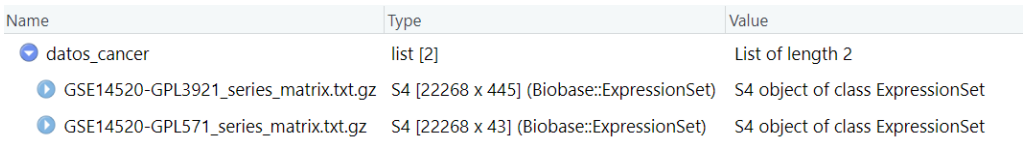
## 4.3. Metodología y procedimiento seguido

### Datos de pacientes de la plataforma GEO

#### Obtención y organización de los datos

Los datos que recogen la información relativa a la expresión de los distintos genes en el cáncer hepático, así como la relativa a las variaciones en dicha expresión, tal y como ya se ha indicado, son los que se van a utilizar para la realización del proyecto. Concretamente, los datos almacenados en la "Serie GSE14520", la cual contiene datos de expresión de genes tanto en tejidos sanos como en tejidos patogénicos, es decir, tejidos hepáticos que han sufrido un proceso neoplásico. Este puede ser primario, cuando se origina en el propio órgano, o bien secundario, cuando se produce la malignización a través de una metástasis de un tumor originado en otra localización corporal. [National Center for Biotechnology Information, 2024b]

Para la obtención de dichos datos se ha utilizado la biblioteca "GEOquery", disponible en R y cuyo principal propósito es la obtención de datos a partir de la plataforma GEO (*Gene Expression Omnibus*).



| Name | Type | Value |
|---|---|---|
| datos_cancer | list [2] | List of length 2 |
| GSE14520-GPL3921_series_matrix.txt.gz | S4 [22268 x 445] (Biobase::ExpressionSet) | S4 object of class ExpressionSet |
| GSE14520-GPL571_series_matrix.txt.gz | S4 [22268 x 43] (Biobase::ExpressionSet) | S4 object of class ExpressionSet |

Figura 4.12: Objeto que contiene los datos de expresión de los distintos tipos de cáncer.
Fuente: Código del proyecto

Una vez obtenido el conjunto de datos de la serie GSE14520, se ha almacenado en un objeto al que hemos denominado 'datoscancer' y que es de tipo *ExpressionSet*. El contenido de dicho objeto se puede observar en la Figura 4.12.

La totalidad de los datos no fue analizada utilizando la misma plataforma. En los datos que hemos descargado desde GEO y almacenado en el objeto 'datoscancer' encontramos dos objetos jerárquicamente organizados en función del array utilizado para su análisis.

De hecho fueron utilizadas dos plataformas distintas para la realización del análisis, el array GPL571 y el array GPL3921, utilizándose una biblioteca distinta en cada una de ellas para la codificación del gen al que hace referencia cada una de las sondas utilizadas en el array del experimento.

Por lo tanto, vamos a separar los datos en función de los arrays que se utilizaron para su procesamiento. Los datos de la plataforma GPL571 los almacenamos en el objeto 'arrayGPL571' cuyo contenido lo podemos observar en la Figura 4.13. Por otra parte, los datos que han sido analizados utilizando la plataforma GPL3921 los vamos a guardar en un objeto que denominamos .ªrrayGPL3921ƨ cuyo contenido se muestra en la Figura 4.14.

Como podemos observar tanto en la Figura 4.13 como en la Figura 4.14, en ambos objetos de las plataformas, se encuentra la información relativa a la expresión de los genes que nos interesan para el proyecto. Pero para poder obtener esta información y trabajar con ella es necesario acceder a la lista 'assayData', que tal y como su nombre indica, contiene los datos de expresión de los genes y, dentro de ella, es necesario acceder a la tabla 'exprs' en la que se representan los datos en crudo de cada una de las sondas que se emplean en el array del experimento para cada una de las plataformas,

```
> array_GPL571
$`GSE14520-GPL571_series_matrix.txt.gz`
ExpressionSet (storageMode: lockedEnvironment)
assayData: 22268 features, 43 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM362947 GSM362948 ... GSM363451 (43 total)
  varLabels: title geo_accession ... Tissue:ch1 (43 total)
  varMetadata: labelDescription
featureData
  featureNames: 1007_s_at 1053_at ... AFFX-TrpnX-M_at (22268 total)
  fvarLabels: ID GB_ACC ... Gene Ontology Molecular Function (16 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 21159642
```

Figura 4.13: Representación del objeto'array$_GPL571, enelquesealmacenanlosdatosdelaplataf$
*Fuente* : *Códigodelproyecto*

```
> array_GPL571
$`GSE14520-GPL571_series_matrix.txt.gz`
ExpressionSet (storageMode: lockedEnvironment)
assayData: 22268 features, 43 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM362947 GSM362948 ... GSM363451 (43 total)
  varLabels: title geo_accession ... Tissue:ch1 (43 total)
  varMetadata: labelDescription
featureData
  featureNames: 1007_s_at 1053_at ... AFFX-TrpnX-M_at (22268 total)
  fvarLabels: ID GB_ACC ... Gene Ontology Molecular Function (16 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 21159642
```

Figura 4.14: Representación del objeto'array$_GPL3921, enelquesealmacenanlosdatosdelaplata$
*Fuente* : *Códigodelproyecto*

GPL571 y GPL3921 respectivamente.

### Procesamiento de los datos en crudo

Como los datos que hemos almacenado en los objetos que aparecen en las Figuras 4.13 y 4.14 respectivamente contienen los datos de expresión génica en crudo, vamos a normalizar dichos datos a través de una transformación logarítmica en base dos para poder realizar las comparaciones relativas al nivel de expresión entre los diferentes genes, ya que a través de esta normalización facilitamos tanto este proceso como el posterior análisis de los datos.

Estos datos normalizados los guardamos en los objetos 'exparrayGPL571' y 'exparrayGPL3921' en función de la plataforma utilizada durante el análisis, Estos datos de expresión normalizados los podemos visualizar en las

imágenes 4.15 y 4.16 respectivamente.

```
> exp_array_GPL571[1:10,]
          GSM362947 GSM362948 GSM362949 GSM362950 GSM362951 GSM362952 GSM362953 GSM362954
1007_s_at  2.693989  2.920674  3.078951  2.884403  2.765747  2.861360  2.784923  2.906313
1053_at    2.341986  2.168963  2.159306  2.114034  2.083724  2.042644  1.791606  1.800330
117_at     1.981487  2.211324  1.967906  1.997473  1.973060  2.008272  1.982583  2.035624
121_at     2.742222  2.712816  2.732052  2.723777  2.783876  2.777788  2.895109  2.777157
1255_g_at  1.718088  1.613060  1.595981  1.621993  1.677170  1.635058  1.722029  1.650765
1294_at    2.616593  2.587605  2.822526  2.645702  2.682573  2.417920  2.659240  2.779680
1316_at    1.976730  1.890641  2.012569  1.930170  2.054154  2.085085  2.282736  2.101650
1320_at    1.901108  1.873813  1.817214  1.849599  1.853197  1.815575  1.876173  1.898789
1405_i_at  2.063848  2.175684  2.312375  2.308011  2.392317  2.064538  2.341417  2.251871
1431_at    1.897628  3.064366  3.435229  2.877744  3.716881  3.437627  3.610700  2.669027
```

Figura 4.15: Datos de expresión del array GPL571 normalizados.
Fuente: Código del proyecto

```
> exp_array_GPL3921 [1:10, ]
          GSM362958 GSM362959 GSM362960 GSM362961 GSM362962 GSM362963 GSM362964 GSM362965
1007_s_at  2.781570  2.935083  2.984589  2.735955  2.832688  2.765323  2.803434  2.729879
1053_at    2.217541  2.098622  2.087463  2.037382  1.973795  1.955685  2.048236  2.212258
          GSM362966 GSM362967 GSM362968 GSM362969 GSM362970 GSM362971 GSM362972 GSM362973
1007_s_at  2.885574  2.697774  2.713036  2.723559  2.714136  2.781989  2.535804  2.695326
1053_at    2.163821  2.090176  1.933950  1.944484  2.246104  2.043345  2.080658  1.978562
```

Figura 4.16: Datos de expresión del array GPL3921 normalizados.
Fuente: Código del proyecto

Una vez que tenemos los datos normalizados vamos a hacer una reestructuración de estos para dividirlos en los dos grupos que vamos a comparar "Pacientes sanosz "Pacientes enfermos". En el array de la plataforma GPL571, además de los pacientes sanos y enfermos encontramos datos de expresión referentes a un grupo de donantes. Sin embargo, estos datos no los vamos a analizar puesto que equivalen únicamente a 2 muestras y no obtendríamos ninguna significación clínica a partir de su análisis.

Para llevar a cabo este proceso creamos un nuevo DataFrame para cada plataforma en el que se almacene el identificador único de dicha muestra junto a la descripción de la misma. En función de la descripción de la misma, que puede ser "Liver Tumor Tissue.º "Liver Non-Tumor Tissue", cambiamos la descripción de estos por los grupos que vamos a crear para estudiar la diferencia entre la expresión diferencial de los genes, "Patogénicosz "Sanosrespectivamente.

### Análisis de la expresión diferencial

Para estudiar la expresión diferencial de los distintos genes que aparecen en las sondas de nuestro experimento se emplea el paquete limma de BioConductor.

Para llevar a cabo dicho análisis se realizan los siguientes pasos:

- Se aplica la función 'lmFit' mediante la que realizamos un ajuste del modelo lineal a cada una de las sondas de los genes que encontramos en la matriz de expresión. Esto nos permite realizar comparaciones entre las condiciones experimentales que se han empleado.

- Se aplica la función 'eBayes' a los resultados del ajuste lineal, que lo que hace es realizar un ajuste bayesiano a las estadísticas de estos modelos lineales que hemos ajustado anteriormente, mejorando así el potencial y precisión estadísticos del análisis que estamos realizando, pudiendo establecer los genes diferencialmente expresados con un mayor grado de confianza.

- A continuación se aplica la función 'topTable' que nos permite extraer los resultados del análisis de la expresión diferencial que hemos realizado, así como ordenarlos en base a su diferencia en la expresión.

- Por último, como tenemos una gran cantidad de genes, vamos a seleccionar aquellos que presentan una evidencia clara de estar diferencialmente expresados. Para ello, vamos a extraer los datos con un p-valor ajustado inferior a 0.05, condición que se utiliza normalmente en el ámbito científico para establecer la significación clínica de los datos.

Las sondas pertenecientes a los genes expresados diferencialmente y que presentan una significación clínica junto al resumen de su análisis estadístico se almacenan en los objetos 'pvalor571' y 'pvalor3921' en función del array utilizado para la expresión, tal y como podemos ver en las Figuras 4.17 y 4.18 respectivamente.

```
> pvalor_GPL571
               logFC  AveExpr        t      P.Value   adj.P.Val        B
222358_x_at 0.1646779 2.673542 6.117073 2.331156e-07 0.005191019 6.692003
```

Figura 4.17: Resultados del análisis de la expresión diferencial en la plataforma GPL571.
Fuente: Código del proyecto

### Representación de los datos

Entre los datos que obtenemos del análisis estadístico que hemos realizado, el valor que más nos importa es el valor 'logFC', que es una abreviatura de 'Log Fold Change' o 'Logaritmo de Cambio de Pliegue' y que se utiliza tanto en el campo de la Bioinformática como en el campo de la Biología Molecular para comparar los niveles de expresión de los genes entre dos condiciones diferentes. En nuestro caso concreto, por ejemplo, vamos a

```
> pvalor_GPL3921
                    logFC  AveExpr          t      P.Value    adj.P.Val        B
212952_at        0.14829347 3.075325  12.041052 4.046552e-29 6.040511e-25 55.32281
215259_s_at     -0.12271637 1.770177 -11.970273 7.728971e-29 6.040511e-25 54.68675
208461_at       -0.10707347 1.744662 -11.964624 8.137926e-29 6.040511e-25 54.63607
208522_s_at     -0.10336801 1.748434 -11.830054 2.769561e-28 1.541815e-24 53.43233
217422_s_at     -0.09811215 1.817244 -11.726259 7.088784e-28 3.157061e-24 52.50869
208495_at       -0.09553109 1.751604 -11.545482 3.605726e-27 1.338205e-23 50.91027
```

Figura 4.18: Resultados del análisis de la expresión diferencial en la plataforma GPL3921.
Fuente: Código del proyecto

diferenciar entre los tejidos 'Sanos' y los tejidos 'Patogénicos' de hígado, es decir, los tejidos que han desarrollado una neoplasia.

Para el cálculo de este valor se utiliza la siguiente fórmula:

$$\text{logFC} = \log_2 \left( \frac{Expresiónenlacondiciónexperimental}{Expresiónenlacondiciónnativa} \right)$$

Como en nuestro proyecto, la condición experimental es el estado en el que los tejidos son diagnosticados con cáncer hepático y la condición nativa cuando estos tejidos no sufren de esta patología, podemos interpretar los valores de la siguiente manera:

- **Valor de logFC positivo:** Indica que el gen se encuentra sobreexpresado en la condición experimental, es decir, en la neoplasia hepática en nuestro caso. Esto ocurre sobre todo con los oncogenes, genes que cuanto mayor se expresen, mayor probabilidad de que se desarrolle un tumor existe.

- **Valor de logFC negativo:** Indica que los niveles de expresión de dicho gen son menores en la condición experimental con respecto a la condición nativa, en nuestro caso diríamos que el nivel de expresión del gen en el tejido canceroso es menor que en el tejido sano. Este sería el caso de los genes supresores de tumores, ya que cuanto menos se expresen mayor probabilidad habrá de que los individuos afectados sufran un proceso neoplásico.

Por lo tanto, lo que hacemos ahora es una remodelación de la forma en la que se presentan los datos de la expresión diferencial.

Para realizar dicha transformación seguimos los siguientes pasos:

- Utilizamos las bibliotecas que utiliza cada una de las plataformas 'HGU133A2' y 'HT HGU133A' respectivamente para crear un nuevo Da-

taFrame que relacione cada una de las sondas utilizadas con su gen correspondiente y eliminamos las muestras que no correspondan a ningun gen.

- Creamos una función que combine dos DataFrames en base a una columna que ambos presenten en común, es decir, creamos una función que realice un merge join.

- Filtramos el DataFrame resultante para que solo contenga la columna con los identificadores de los genes y el valor de logFC, para de esta forma visualizar fácilmente los genes que se han expresado diferencialmente y si estos se encuentran sobre o subexpresados.

Llegados a este punto, no nos queda ninguna de las muestras que han sido analizadas en el array GPL571, debido a dos motivos principales. Solo había una muestra que presentaba significación clínica, y la sonda de esta no se encontraba relacionada con ningun gen.

Los genes de la plataforma GPL3921 junto a su valor de logFC se pueden observar en la Figura 4.19. Además, hemos añadido a este DataFrame una nueva columna que indica si se encuentran sobreexpresados o subexpresados en función del valor de la columna logFC.

```
> datos_genes_GPL3921 [1:10, ]
     SYMBOL       logFC         status
1    A4GALT -0.05608114  Subexpresados
2      AAK1 -0.07577308  Subexpresados
3      AAK1 -0.04076325  Subexpresados
4      AAK1  0.03424780 Sobrexpresados
5     AAMDC -0.01893911  Subexpresados
6     AAMDC  0.02572030 Sobrexpresados
7     AANAT -0.03321977  Subexpresados
8      AAR2  0.03852956 Sobrexpresados
9  AASDHPPT  0.15426609 Sobrexpresados
10 AASDHPPT  0.06125324 Sobrexpresados
```

Figura 4.19: Datos expresados diferencialmente en el cáncer hepático.
Fuente: Código del proyecto

Sin embargo, en las muestras analizadas en la plataforma GPL3921 encontramos un gran número de genes expresados diferencialmente. Sin embargo, aunque con un valor de logFC >0.1 o <-0.1 ya se considera que se expresan diferencialmente con seguridad, filtramos aquellos genes con

un valor de logFC < -0.15 o un valor de logFC <0.15 para obtener una mayor significación y los almacenamos en dos objetos distintos, 'genessubreexpGPL3921' y 'genessobreexpGPL3921' respectivamente, los cuales se pueden observar en las Figuras 4.20 y 4.21 respectivamente.

```
> genes_subexp_GPL3921
      SYMBOL      logFC        status
9539 UGT2B15 -0.1680315 Subexpresados
9540 UGT2B17 -0.1680315 Subexpresados
```

Figura 4.20: DataFrame que contiene los genes subexpresados en la plataforma GPL3921 junto a su valor de logFC.
Fuente: Código del proyecto

```
> genes_sobreexp_GPL3921[1:10,]
       SYMBOL     logFC         status
9    AASDHPPT 0.1542661 Sobrexpresados
100     ACTR2 0.1763699 Sobrexpresados
313   ALDH5A1 0.1636496 Sobrexpresados
677  ATP6V0E1 0.2029485 Sobrexpresados
717  ATXN7L3B 0.2317468 Sobrexpresados
830     BLTP2 0.1551552 Sobrexpresados
1687    CPNE3 0.1662161 Sobrexpresados
1691 CPS1-IT1 0.1667551 Sobrexpresados
2010    DDX17 0.1914140 Sobrexpresados
2092     DHX9 0.1635728 Sobrexpresados
```

Figura 4.21: DataFrame que contiene los genes sobrexpresados en la plataforma GPL3921 junto a su valor de logFC.
Fuente: Código del proyecto

## Datos de las variantes de los genes

En cuanto a las variantes el procedimiento seguido es el siguiente:

Se han definido una serie de funciones que interactúan con la API de la base de datos Ensembl. A través de estas vamos obteniendo datos acerca de un gen, como sus variantes, tanto benignas como patogénicas, sus exones, su coding sequence (CDS), etc.

Para realizar los alineamientos de las secuencias y poder observar aquellas zonas que se encuentran más conservadas, zonas más importantes en el desarrollo de la función de su producto génico y menos conservadas, necesitamos obtener la secuencia de estas variantes, la cual obtenemos a través de una función que modifica la secuencia consenso del gen introduciendo dichas variantes.

Tenemos otra función que se encarga de guardar estas secuencias en formato SeqRecord en un archivo de tipo fasta.

Para realizar el alineamiento se utiliza una función que lanza una solicitud a la API de Clustal Omega para realizar el alineamiento. Y por último, se hace un mapa de calor que represente el resultado del alineamiento, pudiendo observar de una forma mucho más visual tanto las zonas más como menos conservadas en la secuencia.

# Resultados

En este capítulo se incluyen los resultados obtenidos durante la realización del presente proyecto, así como una explicación y discusión de los mismos.

## 5.1. Resumen de resultados.

### Genes expresados diferencialmente en el cáncer hepático

En un primer análisis de los resultados, observamos que cuando aparece un proceso neoplásico en el hígado se ve alterada la expresión de 10.163 genes de los aproximadamente 25.000 genes que tenemos en el cuerpo humano. Sin embargo, este valor no es realmente significativo, puesto que en los cálculos de expresión diferencial hay genes con un valor de logFC tan bajo que esta expresión diferencial puede deberse a la propia naturaleza de las muestras utilizadas.

Este es el motivo principal por el que se han establecido unos umbrales de significación clínica de la expresión diferencial entre [-0.15, 0.15], concluyendo que los valores de logFC que se encuentren por debajo del límite inferior o por encima del límite superior son los que presentan una expresión diferencial significativa. Por lo tanto, si tomamos como referencia estos umbrales que hemos predefinido en base a la propia naturaleza de las muestras empleadas durante el desarrollo del proyecto, podemos concluir que tenemos un total de **42 genes diferencialmente expresados** en el cáncer hepático.

Dentro de esta variación en los niveles de expresión de los genes cuando las células pasan de su estado normal a su estado canceroso, observamos

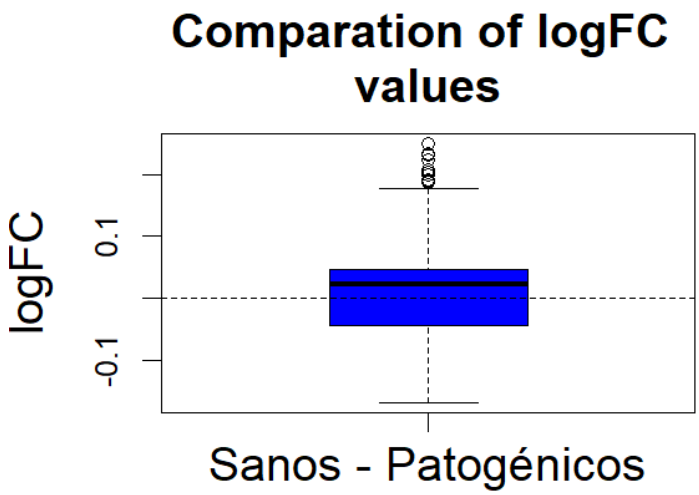



Figura 5.1: Boxplot que compara las muestras de pacientes sanos con las muestras de pacientes con cáncer hepático
Fuente: Código del proyecto

que existe una mayor cantidad de genes que se ven sobreexpresados que subexpresados. Concretamente, encontramos 40 genes que se encuentran sobreexpresados, mientras que solo dos se encuentran subexpresados, teniendo en cuenta el umbral que hemos establecido nosotros.

Podemos observar la variación en los niveles de expresión de los distintos genes en la Figura 5.1 que aparece a continuación.

Como ya sabemos, en la caja central que se representa en este tipo de gráficos hace referencial al percentil 50, es decir, en su interior se encuentran el 50 por ciento de las muestras analizadas. Asimismo, a partir de este gráfico podemos inferir que la distribución de los niveles de expresión de los genes de las muestras es muy variada ya que las líneas comunmente conocidas como bigotes son de gran longitud.

El gráfico evidencia también la mayor cantidad de genes sobreexpresados en referencia a los subexpresados que hemos mencionado anteriormente. Dicha mediana se representa con una línea negra en la caja central, y como podemos observar se encuentra en la parte superior de la misma.

Además, podemos observar una gran cantidad de outliers en la representación, los cuales equivalen a puntos con una diferencia de expresión entre las dos condiciones (sano y patogénico) muy elevada, por lo tanto, serán los genes con una mayor relevancia clínica a la hora de estudiar los mecanismos de aparición del cáncer hepático. Estos se encuentran en la parte superior del gráfico, por lo tanto, refuerzan la teoría de que existe una mayor cantidad de genes sobreexpresados que subexpresados.

Asimismo, en la Figura 5.2 podemos observar los niveles de expresión individuales para cada una de las muestras que han sido analizadas en la plataforma GPL3921.

Esta gráfica refuerza lo que hemos mencionado antes, puesto que si observamos los niveles de expresión de las muestras relativas a genes subexpresados se encuentran más homogéneamente repartidos, existiendo una menor cantidad de muestras con niveles de expresión más alejados de la media, mientras que si observamos las muestras relativas a genes subexpresados vemos una mayor cantidad de muestras que contienen genes con valores de expresión más alejados de la media, es decir, muestras más significativas a nivel clínico, ya que estos son los genes que varían en mayor medida su expresión según se desarrolla el proceso tumoral.

Por lo tanto, a partir de los resultados de los niveles de expresión podemos establecer que los genes sobreexpresados en mayor cantidad son los siguientes:

- **Gen ATXN7L3B:** Este gen codifica para la proteína del mismo nombre (Ataxin 7-like protein 3B) cuya actividad se relaciona con la función 'stemness' de las células madre. Por lo tanto, en lo referente al cáncer hepático, la sobreexpresión de este gen podría llevar asociado una mayor progresión de esta neoplasia, así como un mayor desarrollo de resistencia ante antibióticos.

Por lo tanto, el estudio de este gen podría ser crucial para la comprensión del desarrollo del cáncer hepático, así como su potencial utilización como diana terapéutica.

- **Gen PRKAB2:** Este gen también se encuentra sobreexpesado en las muestras de cáncer hepático que hemos analizado, jugando una función fundamental en la regulación del metabolismo energético, una acción fundamental para la supervivencia de las células cancerosas. Por lo tanto, al verse sobreexpresado, las células cancerosas escapan de los procesos de apoptosis y se producen sin ningún tipo de control.

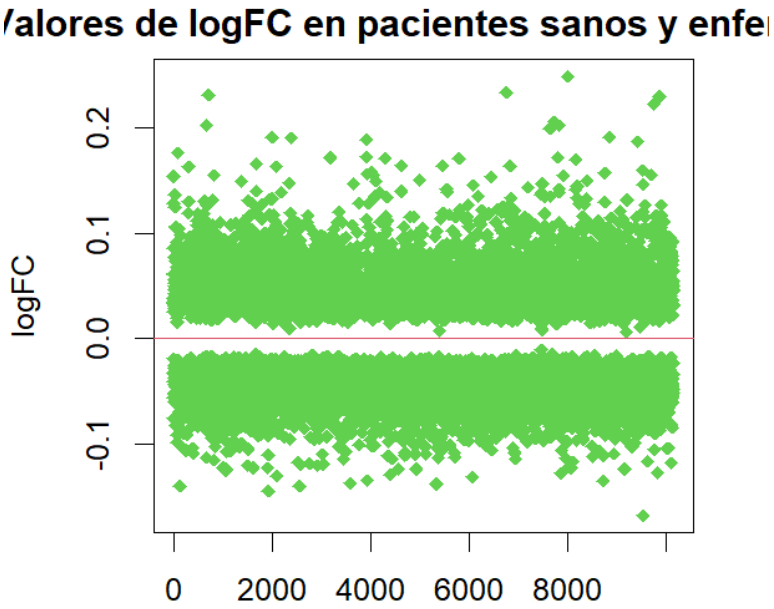



Figura 5.2: Figura que representa los niveles de expresión de cada una de las muestras.
Fuente: Código del proyecto

SLC23A2: Identificado como un transportador de vitamina C, SLC23A2 mostró sobreexpresión en CHC, sugiriendo una posible función en el manejo del estrés oxidativo en el microambiente tumoral.

- **Gen SLC23A2:** Este gen, el cual también se encuentra sobreexpresado codifica para un transportador de vitamina C, por lo tanto, la sobreexpresión en el tejido tumoral puede acarrear un cambio en el manejo del ambiente oxidativo en el microambiente neoplásico.

## Programación de funciones

Por su parte, en el desarrollo del código que permite la identificación de las regiones de los diferentes genes que sufren mutaciones, encontramos lo siguiente:

### Preprocesamiento de los datos

Esta etapa del proyecto ha sido crucial para poder realizar el análisis de las zonas de los genes que sufren variaciones. Entre las funciones principales del código encontramos las siguientes:

**1. Carga de Datos:** Lectura de archivos CSV con datos de expresión génica.

**2. Normalización:** Aplicación de transformaciones logarítmicas para normalizar los datos de expresión, asegurando la comparabilidad entre muestras.

**3. Filtrado:** Eliminación de datos faltantes y selección de genes de interés para el análisis detallado.

Estas etapas fueron fundamentales para garantizar la calidad y precisión de los análisis posteriores.

### Análisis de las Variantes Genéticas

Hemos realizado un análisis de las variantes, obteniendo sus exones, sus variantes tanto benignas como patogénicas, etc.

Hemos creado dos scripts que contienen las variantes y secuencias de los genes TP53 y CTNNB1 respectivamente, ya que son los que mayor implicación en el cáncer hepático presentan. Entre los resultados específicos encontramos los siguientes:

**- CTNNB1:** Se identificaron variantes específicas que podrían estar asociadas con la activación de vías oncogénicas en CHC. Estas variantes ofrecen una visión detallada de los mecanismos moleculares implicados en la progresión del cáncer.

**- TP53:** Se detectaron mutaciones frecuentes en TP53, que contribuyen a la disfunción del control del ciclo celular y la apoptosis en CHC. Estas mutaciones son críticas para entender la biología del tumor y desarrollar posibles estrategias terapéuticas.

### Alineamiento de secuencias

Esta se ha realizado utilizando la herramienta Clustal Omega y ha sido un alineamiento de tipo múltiple en el que se alinearon las variantes de genes ampliamente implicados en CH, como son el TP53 o el CTNNB1.

### Identificación de patrones

Al representar el resultado del alineamiento se han podido observar regiones más conservadas entre las variantes de los genes y regiones más divergentes o menos conservadas, lo cual al realizar su representación nos

proporciona información valiosa sobre la evolución y funcionalidad de las variantes que hemos identificado previamente. Identificación de Patrones: Los alineamientos revelaron patrones conservados y divergentes entre las secuencias, proporcionando información valiosa sobre la evolución y funcionalidad de las variantes identificadas.

# Conclusiones

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas.

## 6.1. Resumen del Proyecto

Este proyecto se centró en el análisis de la expresión diferencial de genes en el carcinoma hepatocelular (CHC), utilizando un enfoque bioinformático integral. Los objetivos principales incluyeron la identificación de genes diferencialmente expresados, la evaluación de su potencial como biomarcadores y objetivos terapéuticos, y la mejora de la comprensión de los mecanismos moleculares subyacentes al CHC.

## 6.2. Identificación de Genes Diferencialmente Expresados

A través del análisis de datos de expresión génica del conjunto de datos GSE14520, se identificaron varios genes con expresión diferencial significativa en muestras de CHC en comparación con tejidos hepáticos normales. Entre estos genes se encuentran ATXN7L3B, PRKAB2 y SLC23A2, los cuales mostraron una sobreexpresión notable en las muestras de cáncer.

**ATXN7L3B:** Este gen se relaciona con la promoción de la capacidad de las células madre en el CHC, sugiriendo un papel importante en la progresión del cáncer y en la posible resistencia a los tratamientos. **PRKAB2:**

Como subunidad de la proteína quinasa activada por AMP, PRKAB2 está implicada en la regulación del metabolismo energético, un factor crucial en la supervivencia y proliferación de las células cancerosas hepáticas. **SLC23A2:** Este transportador de vitamina C está sobreexpresado en el CHC, indicando una posible función en el manejo del estrés oxidativo en el microambiente tumoral.

## 6.3. Metodologías Implementadas

El proyecto empleó herramientas bioinformáticas avanzadas para la descarga, normalización y análisis de datos de expresión génica:

**R y Paquetes Asociados:** Se utilizó R junto con paquetes como GEOquery y limma para obtener y procesar datos de expresión génica. Estos paquetes permitieron una normalización efectiva y la identificación precisa de genes diferencialmente expresados. **Python y Jupyter Notebook:** Para el procesamiento adicional y la visualización de datos, se emplearon Python y Jupyter Notebook, utilizando bibliotecas como pandas, numpy y matplotlib. **Relevancia de los Resultados** Los genes identificados como diferencialmente expresados tienen el potencial de servir como biomarcadores para el diagnóstico temprano del CHC. Además, la sobreexpresión de genes como ATXN7L3B y PRKAB2 sugiere posibles dianas terapéuticas, lo cual es crucial para el desarrollo de tratamientos más efectivos y personalizados.

Además, los resultados obtenidos pueden ser de gran utilidad en los siguientes campos: 1. Diagnóstico Temprano: La detección de biomarcadores específicos del CHC puede mejorar significativamente las tasas de supervivencia al permitir un diagnóstico más temprano y preciso. 2. Terapias Dirigidas: Los genes identificados ofrecen nuevas oportunidades para el desarrollo de terapias dirigidas, que pueden inhibir rutas específicas implicadas en la progresión del cáncer hepático.

## 6.4. Validación y Reproducibilidad

Para asegurar la validez y reproducibilidad de los resultados, se implementaron varios enfoques:

1. Validación Cruzada: Se realizaron validaciones cruzadas internas para verificar la consistencia de los resultados obtenidos.

2. Publicación de Métodos: Los scripts y métodos utilizados fueron documentados y puestos a disposición de la comunidad científica, permitiendo la replicación de los análisis y la validación independiente de los hallazgos.

## 6.5. Impacto en la Investigación del CHC

Este proyecto contribuye significativamente al campo de la investigación del carcinoma hepatocelular al proporcionar nuevos conocimientos sobre los mecanismos moleculares de la enfermedad y al identificar potenciales biomarcadores y dianas terapéuticas. Los hallazgos pueden servir como base para futuras investigaciones y desarrollo de tratamientos clínicos.

En conclusión, este proyecto ha logrado identificar genes diferencialmente expresados en el carcinoma hepatocelular, destacando su potencial como biomarcadores y objetivos terapéuticos. A través del uso de herramientas bioinformáticas avanzadas y la validación de resultados.

Los análisis realizados en este proyecto han proporcionado una comprensión detallada de los genes diferencialmente expresados y las variantes genéticas en el carcinoma hepatocelular. Los genes ATXN7L3B, PRKAB2 y SLC23A2 emergieron como potenciales biomarcadores y dianas terapéuticas, mientras que las variantes en CTNNB1 y TP53 ofrecen información crucial para la caracterización molecular del CHC.

El preprocesamiento adecuado de datos, el análisis exhaustivo de variantes y los alineamientos de secuencias han sido fundamentales para obtener resultados robustos y reproducibles. La representación visual de los alineamientos ha facilitado la identificación de patrones importantes y regiones conservadas en los genes clave.

Por lo tanto, se ha contribuido significativamente a la comprensión de los mecanismos moleculares del CHC y se ha sentado una base sólida para futuras investigaciones y desarrollos clínicos, mejorando las estrategias de diagnóstico y tratamiento del carcinoma hepatocelular.

# Lineas de trabajo futuras

En este capítulo vamos a exponer diferentes líneas de continuación del proyecto en un futuro, así como la explicación relativa a cada una de ellas.

Un hecho muy interesante que se podría añadir en el desarrollo del proyecto es el estudio acerca de cómo cambian los niveles de expresión de los distintos genes cuando se trata de un tumor primario o metástasis o como van cambiando los niveles de expresión a medida que evoluciona el proceso oncológico.

Además, sería muy interesante validar experimentalmente los resultados, sobre todo los genes expresados diferencialmente, para que el proyecto presente una mayor significancia clínica. Para llevar a cabo este proceso se pueden emplear técnicas tales como la PCR en tiempo real (qPCR), Western blot o ensayos funcionales en lineas celulares de cáncer hepático y modelos animales.

También se podrían añadir nuevos análisi integrando datos procedentes de otras ómicas, como son los datos de metilación del ADN, datos de proteómica o incluso de metabolómica. Con esta mejora seremos capaces de obtener una visión más completa de los mecanismos celulares de patogénesis que afectan al CH.

Otra opción es el desarrollo de modelos predictivos. Se pueden crear modelos predictivos para el diagnóstico y pronóstico del CH basados en los genes y vías identificada a través de algoritmos de aprendizaje automático y validación cruzada, con el obejtivo de desarrollar y probar modelos predictivos en cohortes de pacientes.

Por lo tanto, aunque con los resultados que hemos obtenido con el presente proyecto obtenemos un conocimiento bastante amplio del cáncer

hepático, de los genes con mayor influencia en el mismo y su desarrollo, existen todavía una gran variedad de pasos para que este proyecto se pueda emplear y tenga una significancia clínica real.

# Bibliografía


[Abalozz, 2024] Abalozz (2024). Bibliotecas en programación: ¿Qué son y para qué sirven? https://abalozz.es/bibliotecas-en-programacion-que-son-y-para-que-sirven/. Acceso realizado el 10 de mayo de 2024.

[Biopython Contributors, 2024a] Biopython Contributors (2024a). Biopython. https://biopython.org/. Acceso realizado el 11 de mayo de 2024.

[Biopython Contributors, 2024b] Biopython Contributors (2024b). Biopython Wiki: Seq. https://biopython.org/wiki/Seq. Acceso realizado el 12 de mayo de 2024.

[Biopython Contributors, 2024c] Biopython Contributors (2024c). Biopython Wiki: SeqIO. https://biopython.org/wiki/SeqIO. Acceso realizado el 12 de mayo de 2024.

[Biopython Contributors, 2024d] Biopython Contributors (2024d). Biopython Wiki: SeqRecord. https://biopython.org/wiki/SeqRecord. Acceso realizado el 12 de mayo de 2024.

[Bosch et al, 2004] Bosch et al (2004). Primary liver cancer: Worldwide incidence and trends. *Gastroenterology*, 127(5):S5–S16.

[Burgos San Juan, 2008] Burgos San Juan, L. (2008). Colangiocarcinoma. actualización, diagnóstico y terapia. *Revista Médica de Chile*, 136(2):240–248.

[Chapman, Brad and Chang, Jeffrey, 2000] Chapman, Brad and Chang, Jeffrey (2000). Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter*, 20(2):15–19.

[Códigos Python, 2024] Códigos Python (2024). Python 2 vs Python 3: diferencias y consideraciones. https://codigospython.com/python-2-vs-python-3-diferencias-y-consideraciones/. Acceso realizado el 30 de marzo de 2024.

[Davis, Sean and S. Meltzer, Paul , 2007] Davis, Sean and S. Meltzer, Paul (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23(14):1846–1847.

[Ensembl Genome Browser, 2024] Ensembl Genome Browser (2024). Ensembl. https://www.ensembl.org. Acceso realizado el 23 de junio de 2024.

[Fatty Liver Disease, 2024] Fatty Liver Disease (2024). Stages of Liver Cancer. https://fattyliverdisease.com/stages-of-liver-cancer/#. Acceso realizado el 15 de marzo de 2024.

[Gordon K. Smyth and others, 2024] Gordon K. Smyth and others (2024). limma: Linear Models for Microarray and RNA-Seq Data. https://www.bioconductor.org/packages/release/bioc/html/limma.html. Acceso realizado el 17 de mayo de 2024.

[Hertie Coding Club, 2024] Hertie Coding Club (2024). How to Install R and RStudio. https://www.hertiecodingclub.com/learn/rstudio/install-r-studio/. Acceso realizado el 28 de abril de 2024.

[Hu, Hui and Li, Hong and Jiao, Feng and Han, Tao and Zhuo, Mei and Cui, Jian and Li, Yang and Hu, Hui and Li, Hong and Jiao, Feng and Han, Tao and Zhuo, Mei and Cui, Jian and Li, Yang and Wang, Li (2018). Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. *Genome Medicine*, 10(1):42.

[International Agency for Research on Cancer, 2024] International Agency for Research on Cancer (2024). IARC Global Cancer Observatory. https://gco.iarc.fr/en. Acceso realizado el 3 de mayo de 2024.

[Jiawei Sun and Zizhen Zhang and Jiaru Cai and Xiaoping Li and Xiaoling Xu, 2024] Jiawei Sun and Zizhen Zhang and Jiaru Cai and Xiaoping Li and Xiaoling Xu (2024). Identification of Hub Genes in Liver Hepatocellular Carcinoma Based on Weighted Gene Co-expression Network Analysis. *Biochemical Genetics*, 62(1):1–14.

[Jiazhou Ye and Yan Lin and Xing Gao and Lu Lu and Xi Huang and Shilin Huang and Tao Bai and Jiazhou Ye and Yan Lin and Xing Gao and Lu Lu and Xi Huang and Shilin

Huang and Tao Bai and Guobin Wu and Xiaoling Luo and Yongqiang Li and Rong Liang (2022). Prognosis-Related Molecular Subtypes and Immune Features Associated with Hepatocellular Carcinoma. *Cancers*, 14(22):5721.

[Learners' Galaxy, 2024] Learners' Galaxy (2024). Python Programming for Data Science. https://learnersgalaxy.ai/courses/python-programming-for-data-science/. Acceso realizado el 30 de marzo de 2024.

[Mayo Clinic, 2024] Mayo Clinic (2024). Liver cancer - Symptoms and causes. https://www.mayoclinic.org/es/diseases-conditions/liver-cancer/symptoms-causes/syc-20353659. Acceso realizado el 18 de marzo de 2024.

[National Cancer Institute, 2024] National Cancer Institute (2024). ¿Qué es el cáncer de hígado? https://www.cancer.gov/espanol/tipos/higado/que-es-cancer-dehigado#:~:text=%C2%BFQu%C3%A9%20es%20el%20c%C3%A1ncer%20de%20h%C3%ADgado%3F%20El%20c%C3%A1ncer,uno%20de%20los%20%C3%B3rganos%20m%C3%A1s%20grandes%20del%20cuerpo/. Acceso realizado el 12 de marzo de 2024.

[National Center for Biotechnology Information, 2024a] National Center for Biotechnology Information (2024a). Gene Expression Omnibus (GEO). https://www.ncbi.nlm.nih.gov/geo/. Acceso realizado el día 8 de mayo de 2024.

[National Center for Biotechnology Information, 2024b] National Center for Biotechnology Information (2024b). NCBI GEO - GSE14520. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14520. Acceso realizado el día 9 de mayo de 2024.

[Python Software Foundation, 2024a] Python Software Foundation (2024a). JSON - Python 3 Documentation. https://docs.python.org/3/library/json.html. Acceso realizado el 14 de mayo de 2024.

[Python Software Foundation, 2024b] Python Software Foundation (2024b). Python Documentation. https://docs.python.org/. Acceso realizado el 28 de marzo de 2024.

[Python Software Foundation, 2024c] Python Software Foundation (2024c). Requests: HTTP for Humans. https://pypi.org/project/requests/. Acceso realizado el 15 de mayo de 2024.

[Python Software Foundation, 2024d] Python Software Foundation (2024d). sys - System-specific parameters and functions. https://docs.python.org/3/library/sys.html. Acceso realizado el 14 de mayo de 2024.

[R Project, 2024] R Project (2024). R: The R Project for Statistical Computing. https://www.r-project.org/. Acceso realizado el 20 de abril de 2024.

[RDocumentation, 2024] RDocumentation (2024). RDocumentation: Easy-to-search documentation for R. https://www.rdocumentation.org/. Acceso realizado el 20 de abril de 2024.

[Requests: HTTP for Humans, 2024] Requests: HTTP for Humans (2024). Requests: HTTP for Humans — Requests 2.28.2 documentation. https://requests.readthedocs.io/en/latest/. Acceso realizado el 15 de mayo de 2024.

[Roche Pacientes, 2024] Roche Pacientes (2024). Cáncer de Hígado. https://rochepacientes.es/cancer/higado.html. Acceso realizado el 20 de marzo de 2024.

[RStudio, PBC, 2024a] RStudio, PBC (2024a). RStudio. https://www.rstudio.com/. Acceso realizado el 25 de abril de 2024.

[RStudio, PBC, 2024b] RStudio, PBC (2024b). RStudio Community. https://community.rstudio.com/. Acceso realizado el 25 de abril de 2024.

[RStudio, PBC, 2024c] RStudio, PBC (2024c). RStudio Support. https://support.rstudio.com/. Acceso realizado el 25 de abril de 2024.

[RStudio, PBC, 2024d] RStudio, PBC (2024d). RStudio Webinars. https://www.rstudio.com/resources/webinars/. Acceso realizado el 26 de abril de 2024.

[Sharma, D., Subbarao, G., and Saxena, R, 2017] Sharma, D., Subbarao, G., and Saxena, R (2017). Hepatoblastoma. *Seminars in Diagnostic Pathology*, 34(2):192–200.

[Sievers, Fabian and Wilm, Andreas and Dineen, David and Gibson, Toby J and Karplus, Kevin an Sievers, Fabian and Wilm, Andreas and Dineen, David and Gibson, Toby J and Karplus, Kevin and Li, Weizhong and Lopez, Rodrigo and McWilliam, Hamish and Remmert, Michael and Soding, Johannes and Thompson, Julie D and Higgins, Desmond G (2011). Fast, scalable

generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539.

[Smith G. K., Data] Smith G. K. (2005], note = Chapter: limma: Linear Models for Microarray Data). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer.

[Srivatanakul P, Sriplung H, Deerasamee S,, 2004] Srivatanakul P, Sriplung H, Deerasamee S, (2004). Epidemiology of Liver Cancer: An Overview. *Asian Pacific J Cancer Prev*, 5(2):118–125.

[Statista, 2024] Statista (2024). Lenguajes de programación más usados del mundo. https://es.statista.com/grafico/16580/lenguajes-de-programacion-mas-usados-del-mundo/. Acceso realizado el 1 de abril de 2024.

[The Cancer Genome Atlas, 2024] The Cancer Genome Atlas (2024). The Cancer Genome Atlas (TCGA). https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. Acceso realizado el 23 de junio de 2024.

[UNIR, 2024] UNIR (2024). Lenguaje R: Big Data y Analítica Predictiva. https://www.unir.net/ingenieria/revista/lenguaje-r-big-data/. Acceso realizado el 22 de abril de 2024.

[Wikipedia, 2024] Wikipedia (2024). Biopython. https://es.wikipedia.org/wiki/Biopython. Acceso realizado el 11 de mayo de 2024.