Universitat Oberta de Catalunya

uoc.edu

Obtención de driver genes para subtipado en Colon adenocarcinoma mediante la integración de datos ómicos.

Sharon Martínez Quiroga
Máster en Bioinformática y Bioestadística

Área 2

Jaume Sastre Tomàs
Marc Maceira

06/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Obtención de driver genes para subtipado en Colon adenocarcinoma mediante la integración de datos ómicos.*
Nombre del autor:	*Sharon Martínez Quiroga*
Nombre del consultor/a:	*Jaume Sastre Tomàs*
Nombre del PRA:	*Marc Maceira Duch*
Fecha de entrega (mm/aaaa):	06/2021
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*2*
Idioma del trabajo:	*Español*
Número de créditos:	*15*
Palabras clave	*Integración de datos ómicos, herramientas para integración de multi-ómicas, subtipado.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

En los últimos años se ha abaratado los costes de la generación de datos ómicos, lo cual ha impulsado a estudiarlos en su conjunto para poder avanzar hasta una medicina personalizada. Hoy en día, la integración de datos nos está permitiendo estudiar para las patologías sus subtipos, biomarcadores, y posibles medicamentos o sus combinaciones.

El objetivo de este TFM es conocer y aplicar los nuevos métodos de integración de datos disponibles, y comparar sus resultados con los de los análisis de ómicas independientes. Además, se verá parte de sus múltiples aplicaciones.

Se integrará datos de genómica, metilómica y transcriptómica mediante iClusterBayes, un método de última generación que ha demostrado tener la misma calidad que sus predecesores pero que ahorra mucho tiempo. Además, se integrarán pathways y otros datos disponibles en repositorios pudiéndose ver el gran potencial de la integración de datos.

Se analizará datos de Adenocarcinoma de Colon (COAD) procedentes del proyecto TCGA. Los resultados de la integración, muy distintos a los de análisis independientes, mostraron que el COAD tiene 5 subtipos. Se obtuvieron 1028 genes con CNA, 331 genes con metilación diferencial y 3073 genes con expresión diferencial que actúan como driver genes de esta patología. Además, se observó que las rutas más afectadas en la patología están relacionadas con la homeostasis y el desarrollo.

Abstract (in English, 250 words or less):

The generation of omic data has reduced its costs in the last years, this has boost its studies and enables to move forward to personalized medicine. Nowadays, the data integration is allowing us to study the subtypes, biomarkers, and medicines for pathologies.

The objective of this TFM is to know and apply the new methods of data integration availables, and compare results with the independent omics data analysis. Also, it presents part of the multiple applications.

This is an integrative multi-omics análisis that integrates genomics, methylomics and transcriptomics though iClusterBayes, which is a last generation method that has demonstrated to achieve the same quality than its predecessors with tradition. Nevertheless, it has the distinctive feature that it saves a lot of time. Furthermore, there are pathways integrated and some other available data in repositories, which show the big potential of data integration.

Beyon that, the Colom Adenocarcinoma (COAD) from the TCGA proyect has been analyzed in this proyect. The results of the integrative analysis were hightly different from the results achieved in the individual analysis. The results of integration were that COAD has 5 subtypes, and got 1028 genes with CNA, 331 genes with differential methylation and 3073 genes with differential expresión that act as drivers of this pathology. Moreover, it was observed that the most affected pathways are related with the homeostasis and the development.

Índice

Lista de figuras

Lista de tablas

1 Resumen

En los últimos años se ha abaratado los costes de la generación de datos ómicos, lo cual ha impulsado a estudiarlos en su conjunto para poder avanzar hasta una medicina personalizada. Hoy en día, la integración de datos nos está permitiendo estudiar para las patologías sus subtipos, biomarcadores, y posibles medicamentos o sus combinaciones.

El objetivo de este TFM es conocer y aplicar los nuevos métodos de integración de datos disponibles, y comparar sus resultados con los de los análisis de ómicas independientes. Además, se verá parte de sus múltiples aplicaciones.

Se integrará datos de genómica, metilómica y transcriptómica mediante iClusterBayes, un método de última generación que ha demostrado tener la misma calidad que sus predecesores pero que ahorra mucho tiempo. Además, se integrarán pathways y otros datos disponibles en repositorios pudiéndose ver el gran potencial de la integración de datos.

Se analizará datos de Adenocarcinoma de Colon (COAD) procedentes del proyecto TCGA. Los resultados de la integración, muy distintos a los de análisis independientes, mostraron que el COAD tiene 5 subtipos. Se obtuvieron 1028 genes con CNA, 331 genes con metilación diferencial y 3073 genes con expresión diferencial que actúan como driver genes de ésta patología. Además, se observó que las rutas más afectadas en la patología están relacionadas con la homeostasis y el desarrollo.

2 Introducción

2.1 Contexto y justificación del Trabajo

2.1.1 Contexto:

El Trabajo de Fin de Máster (TFM) consiste en realizar el análisis de datos de Colon Adenocarcinoma mediante la integración de tres ómicas distintas. Primero se estudia el perfil de expresión y metilación diferencial de forma independiente entre muestras sanas y patológicas. Después, se integran los datos de expresión génica con datos de metilación, solo de muestras patológicas. Finalmente se integran los datos de transcriptómica, metilómica y genómica de muestras patológicas [1]. Los resultados de los diferentes niveles de análisis se comparan para observar el efecto de la integración de datos sobre la obtención de driver genes, subtipado y observación de rutas afectadas. Por último, se estudia las rutas afectadas en la patología. Se podrá apreciar la gran utilidad de la integración de datos multi-ómicos a la hora de realizar subtipados, obtener biomarcadores, y estudiar rutas afectadas por una enfermedad.

La integración se realizará mediante varios métodos elegidos de entre todos los disponibles; usando como criterios de elección que funcionen bien con el tamaño del set de datos a usar en el análisis, que sean capaces de integrar correctamente las ómicas que se ha elegido para este TFM. [2,3]

2.1.2-Justificación del trabajo

En los últimos años el avance tecnológico ha permitido que se generen grandes cantidades de datos biológicos que necesitan ser procesados por profesionales que tengan los conocimientos adecuados. Gracias a la obtención de estos datos, la sociedad se dirige cada vez más hacia la medicina personalizada. Para poder llegar a ella será necesario recabar el máximo conocimiento posible sobre los sistemas biológicos. Según la naturaleza de la molécula estudiada se generan distintas ómicas, las cuales interaccionan entre sí para poner en funcionamiento la célula y el organismo, por lo que la integración de datos ómicos es el método que puede llevar a obtener el máximo conocimiento sobre los sistemas biológicos y la medicina personalizada [5].

El objetivo personal para realizar este TFM es aprender a integrar datos ómicos para, por un lado, en un futuro contribuir a los avances científicos que lleven a desarrollar la medicina personalizada capaz de salvar el mayor número de vidas con menos efectos secundarios para los pacientes. Y, por otro lado, lograr que se reduzca el uso de animales en la investigación.

Considero que este TFM será un paso más en la adquisición de los conocimientos que permitirán mi desarrollo personal en la disciplina de Biología

Molecular de Sistemas y que, además, consolidará y ampliará los conocimientos adquiridos en el máster.

2.2 Objetivos del Trabajo

2.2.1 Objetivos Generales:

1-Revisión del estado del arte. Determinar la metodología y recursos disponibles para integración de datos ómicos, y aquellos métodos que se utilizarán en este TFM.

2-Realizar el análisis independiente e integrativo de las ómicas.

3-Comparar los resultados de los distintos análisis [2,3].

2.2.2-Objetivos específicos:

1-Determinar la metodología y recursos disponibles para integración de datos ómicos.

1.1-Explicar que es la Biología de Sistemas y otros conceptos clave.

1.2-Indicar los inconvenientes por la elevada generación de datos. Indicar algunos repositorios de datos públicos.

1.3-Enumerar y explicar las distintas clases de métodos de integración.

1.4-Indicar herramientas y paquetes disponibles para la integración.

1.5-Indicar soluciones para los elevados requerimientos computacionales de la integración.

1.6-Indicar las características del Adenocarcinoma de colon.

1.7-Seleccionar e indicar los métodos a utilizar en este TFM.

2- Realizar el análisis independiente e integrativo de las ómicas.

2.1-Analizar los datos de las ómicas de forma independiente:

2.1.1- Analizar transcriptómica.

2.1.2- Analizar metilómica.

2.2- Analizar los datos integrando las ómicas a pares:

2.2.1- Integrar transcriptómica y metilómica.

2.3- Analizar los datos por integración de multi-ómicas: genòmica, transcriptómica, y metilómica. Estudiar las rutas afectadas mediante integración de pathways y enriquecimiento.

3-Comparar los resultados obtenidos:

3.1- Visualización y comparación de resultados obtenidos.

3.2- Comparación de drive genes obtenidos en los distintos análisis.

2.3 Enfoque y método seguido

Este TFM se realizará de forma que se cumplan los objetivos detallados previamente. Para poder realizar este trabajo es fundamental realizar una revisión bibliográfica inicial sobre el estado del arte de la integración de datos ómicos. Como en los últimos años se ha incrementado la disponibilidad de datos, el número de ómicas a estudiar, y las herramientas para procesarlas e integrarlas; será necesario elegir tanto las ómicas a estudiar en este TFM, como los métodos más adecuados para la integración de las mismas. [2,3]

Se usará datos de Adenocarcinoma de Colon (COAD) porque en los repositorios hay una cantidad significativa de los mismos para diversas ómicas. Los datos a usar pertenecen al proyecto Cancer Genome Atlas Program, TCGA [4]. Para realizar la integración se han seleccionado genómica (CNA), transcriptómica y epigenómica (metilómica), cuya interacción se ha demostrado que es esencial para el desarrollo del cáncer [1].

Se analizará primero las ómicas por separado con muestras normales y patológicas, después se integrará la transcriptómica con metilómica, y, por último, se integrarán las tres ómicas de tal forma que se pueda observar el efecto de la integración sobre los resultados de los análisis de datos. Finalmente, se estudiará qué rutas se están viendo afectadas, los posibles subtipos de forma visual, y se visualizarán las anotaciones de genes prognósticos y dianas terapéuticas de COAD mediante integración de datos.

2.4 Planificación del Trabajo

2.4.1-Tareas:

1-Realizar una exhaustiva búsqueda bibliográfica para determinar la metodología y recursos disponibles para integración de datos ómicos. Seleccionar los métodos y paquetes a usar. (17/03/21-28/04/2021)

1.1-Buscar y explicar qué es la Biología de Sistemas, algunas de las ómicas más investigadas, y conceptos clave.

1.2-Buscar e indicar los problemas surgidos por la elevada generación de datos ómicos y las posibles dificultades de la búsqueda de datos públicos en repositorios. Buscar los repositorios de datos públicos más relevantes.

1.4-Buscar en la bibliografía las distintas clases de métodos de integración y entender sus fundamentos teóricos.

1.5-Localizar en la bibliografía las últimas herramientas y paquetes disponibles para la integración de datos multi-ómicos y clasificarlas según si son herramientas para programadores o son herramientas para usuarios con bajo conocimiento informático.

1.6-Indicar soluciones para los problemas de requerimientos computacionales.

1.7-Indicar las características del Adenocarcinoma de colon.

1.8-Seleccionar los métodos que se usarán en el TFM.

2- Realizar el análisis independiente e integrativo de las ómicas. (28/04/2021-06/06/2021)

2.1-Analizar los datos de las ómicas de forma independiente:

2.1.1- Analizar transcriptómica:

- Filtrado de datos de RNAseq.
- Estudio de la calidad y normalidad de los datos.
- Transformación de los datos.
- Análisis de expresión diferencial.
- Visualización de resultados: Heatmap y PCA [6,7].
- Enriquecimiento.

2.1.2- Analizar metilómica.

- Filtrado de datos de metiloma
- Estudio de la calidad y normalidad de los datos
- Análisis de metilación diferencial.
- Visualización de resultados: Heatmap y PCA [7].
- Enriquecimiento.

2.2- Analizar los datos integrando las ómicas a pares:

2.2.1- Integrar transcriptómica y metilómica: Integrar los datos de genes diferencialmente expresados y metilados.

2.2.2-Visualización de resultados

2.2.3- Realizar enriquecimiento de los resultados [8,9].

2.3- Analizar los datos por integración de multi-ómicas.

2.3.1- Se integrará los datos de:

- Genómica: genes en regiones con CNA que hayan mostrado expresión diferencial.
- Transcriptómica: genes con expresión diferencial que estén entre los 30% genes más significativos.
- Metilómica: genes con metilación diferencial que tengan efecto funcional sobre la expresión génica.

2.3.2-Se realizará integración de pathways y enriquecimiento de los gene drivers. Además de integrarse datos de supervivencia y de medicamentos.

3-Comparar los resultados obtenidos: (06/06/2021-08/06/2021)

3.1- Visualización y comparación de resultados obtenidos mediante heatmaps y plotPCA.

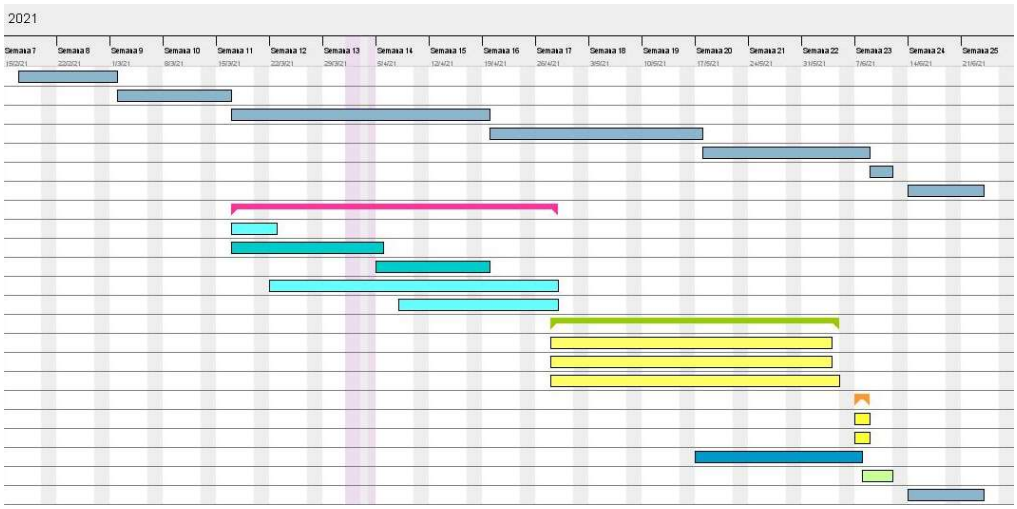
3.2- Comparación de drive genes obtenidos en los distintos análisis mediante diagrama de Venn.

3.3-Comparación de resultados de enriquecimiento.

2.4.2-Fechas clave:

- PEC0: Definición de los contenidos del trabajo (01/03/2021)
- PEC1: Plan de Trabajo (16/03/2021)
- Objetivo 1 (28/04/2021)
- PEC2: Desarrollo del trabajo fase 1 (19/04/2021)
- Objetivo 2 (06/06/2021)
- PEC3: Desarrollo del trabajo fase 2(17/05/2021)
- Objetivo 3 (08/06/2021)
- PEC4: Cierre de la memoria (08/06/2021)
- PEC5a: Elaboración de la presentación (13/06/2021)
- PEC5b: Defensa pública (23/06/2021)

2.4.3-Calendario:



Nombre	Fecha de inicio	Fecha de fin
PEC0-Definición de contenidos del trabajo	17/2/21	1/3/21
PEC1-Plan de Trabajo	2/3/21	16/3/21
PEC2-Desarrollo del trabajo fase 1	17/3/21	19/4/21
PEC3-Desarrollo del trabajo fase 2	20/4/21	17/5/21
PEC4-Cierre de la memoria	18/5/21	8/6/21
PEC5a-Elaboración de la presentación	9/6/21	13/6/21
PEC5b-Defensa pública	14/6/21	23/6/21
Objetivo 1: Búsqueda bibliográfica. Selección y comprensión de métodos.	17/3/21	28/4/21
Bibliografía Adenocarcinoma de Colon	17/3/21	22/3/21
Búsqueda y compra de un ordenador con el perfil adecuado para este TFM	17/3/21	5/4/21
Instalación de R, Rstudio y paquetes a utilizar	5/4/21	19/4/21
Bibliografía integración de datos ómicos	22/3/21	28/4/21
Selección de métodos para descubrir driver genes y planear los workflow	8/4/21	28/4/21
Objetivo 2: Integración de datos ómicos y anotación en la memoria.	28/4/21	4/6/21
Análisis diferencial de datos normales y patológicos	28/4/21	3/6/21
Integraciones a pares	28/4/21	3/6/21
Integración multi-ómica mediante iClusterBayes	28/4/21	6/6/21
Objetivo 3: Comparación de resultados	7/6/21	8/6/21
Comparación de gene drivers obtenidos	7/6/21	8/6/21
Visualización y comparación de resultados	7/6/21	8/6/21
Memoria: Revisión	17/5/21	7/6/21
Presentación	8/6/21	13/6/21
Defensa pública	14/6/21	23/6/21

Figura 1: Diagrama de grantt[12].

2.4.4-Análisis de riesgos:

- **Riesgo 1:** relacionado con el tiempo requerido por la dificultad de la búsqueda de datos.
- **Riesgo 2:** relacionado con el tiempo requerido por el aprendizaje de uso de las herramientas.
- **Riesgo 3:** La gran cantidad de datos con la que hay que trabajar puede dar lugar a que los recursos computacionales disponibles no sean suficientes.
- **Riesgo 4:** La gran cantidad de herramientas disponibles para la integración de datos puede suponer que se elija una herramienta que parece adecuada pero que no genere resultados concluyentes o no sea la mejor para realizar el análisis.
- **Riesgo 5:** Durante la realización del TFM se puede descubrir que sea necesario realizar más pasos de los considerados con anterioridad para obtener los resultados que se busca.

2.5 Breve sumario de contribuciones y productos obtenidos

- *Memoria*: Documento de máximo 90 páginas incluyendo anexos, que contiene los contenidos más importantes del trabajo de fin de máster.
- *Informes de Rmarkdown*: Anexos con los Workflows utilizados y sus resultados.
- *Presentación virtual*: Breve resumen de los contenidos de la memoria.
- *Autoevaluación del proyecto*: Respuesta a las preguntas realizadas por el jurado de evaluación del trabajo de fin de máster.

2.6 Breve descripción de los otros capítulos de la memoria.

- *Capítulo 3: Estado del arte*. Se realizará una revisión sobre COAD y sobre la integración de datos ómicos. Constará de los siguientes subapartados:
  3.1-Biología de Sistemas y Ómicas.
  3.2-Repositorios de datos públicos disponibles y dificultades de la búsqueda de datos.
  3.3-Clases de métodos de integración de datos ómicos.
  3.4-Herramientas y paquetes para la integración de datos ómicos.
  3.5-Soluciones para manejar el big data.
  3.6-Caracterización del Adenocarcinoma de Colon.

- *Capítulo 4: Metodología.* Se indicarán los métodos que se han utilizado para la integración de datos ómicos.
  4.1-Datos a integrar.
  4.2-Filtrado de datos.
  4.3-Transformación de datos.
  4.4-Análisis de una ómica: análisis diferencial de expresión o metilación por regresión lineal.
  4.5-Integración de datos:

4.5.1-Integración de datos ómicos mediante correlación de variables.
4.5.2-Integración de datos multi-ómicos mediante métodos bayesianos.
4.7-Métodos de comparación y visualización de resultados.
4.8-Integración de pathways, visualización de redes y enriquecimiento.

- *Capítulo 5: Resultados*. Resultados de integración de datos ómicos.
  5.1.Subtipado
  5.2.Driver genes
  5.3.Integración de Pathways
  5.4.Enriquecimiento
  5.5.Integración de datos de supervivencia y medicamentos
- *Capítulo 6: Discusión:* Se comparan los resultados de los distintos métodos entre sí y se comparan con la bibliografía existente para COAD.

- *Capítulo 7: Conclusiones.*
  7.1-Conclusiones: se indican las conclusiones de los análisis.
  7.2-Líneas de futuro: posibles ampliaciones o mejoras del trabajo.
  7.3-Seguimiento de la planificación.

- *Capítulo 8: Glosario.*

- *Capítulo 9: Abreviaturas*

- *Capítulo 10: Bibliografía.*

3 Estado del arte

3.1-Biología de Sistemas, Ómicas y otros conceptos importantes:

En este apartado se hará una breve explicación de varios conceptos importantes relacionados con la Integración de datos ómicos, y se profundizará en aquellos que estén más relacionados con los métodos a utilizar en este TFM.

3.1.1-Biología de Sistemas:

La biología de sistemas es la ciencia que estudia los fenómenos biológicos como si fueran parte de un sistema donde los componentes, moléculas, interactúan entre sí y con el ambiente a diferentes niveles, para con ello dar lugar a las características funcionales que caracterizan al sistema en estudio.

Un ejemplo de la visión de una célula como sistema se puede observar en que, en una célula tumoral, la acumulación de mutaciones o variación de copias de regiones cromosómicas genera cambios en la expresión génica, incluyendo genes estructurales y reguladores de la expresión de otros genes. Lo cual deriva en cambios en el perfil proteico y en el perfil de los elementos reguladores de la expresión génica como la metilación de DNA. Todo esto genera cambios en las rutas, pathways, entre ellas las de transmisión de señales y las metabólicas; cambios en el plegamiento y funcionalidad de proteínas; alteraciones en la replicación celular; entre otros muchos fenómenos anormales [2].

El desarrollo de la biología de sistemas será esencial para alcanzar la medicina personalizada. El estudio de los elementos que conforman los distintos niveles del sistema se separa en grupos denominados ómicas. Hay ómicas que estudian componentes internos del individuo, y ómicas que estudian componentes externos al individuo y cómo le afectan [5].

3.1.2-Ómicas que estudian componentes internos del individuo:

- **Genómica:** es la ciencia que estudia las secuencias de DNA de los organismos, es decir, sus genomas [13,14]. Hay dos ramas dentro de la genómica:
    - *Genómica estructural:* se ocupa de crear mapas genómicos estudiando y localizando las distintas estructuras genómicas. [13,15] Entre ellas están los polimorfismos de un único nucleótido (SNP), variación del número de copias (CNV) de genes o regiones cromosómicas en células germinales, o alteración en el número de

copias (CNA) de genes o regiones cromosómicas en células somáticas, y mutaciones [16]. *En este TFM se integrará solo CNA.*

- *Genómica funcional:* se ocupa de estudiar la función de los genes, sean genes estructurales (codifican proteínas) o reguladores (codifican RNA funcionales), para comprender el funcionamiento del genoma [13,14].

- **Transcriptómica:** es la ciencia que estudia el perfil de expresión génica, transcriptómica, de un determinado organismo, tejido o célula, en determinadas condiciones de estudio. Su estudio es importante para obtener biomarcadores para patologías [13,14,17].

  Dentro de la transcriptómica se pueden estudiar por separado distintos productos de expresión génica, los cuales se pueden clasificar en RNA mensajeros (mRNA), que son los RNA codificantes de proteínas; y RNA no codificantes (ncRNA). Destaca el estudio del perfil de expresión de mRNA, y dentro de los ncRNA, prevalece el estudio de microRNA (miRNA), cuya función es regular la expresión de genes [14,17*]. En este TFM se integrará solo mRNA.*

- **Epigenómica**: es la ciencia que estudia el conjunto de modificaciones reversibles, que no implican cambios en la secuencia del DNA, y que regulan la expresión génica. Estas modificaciones incluyen metilación del DNA [18]; remodelación de la cromatina para controlar el acceso a los genes, entre otros mediante metilación, acetilación o ubiquitinación de histonas o unión de factores de transcripción [19-21]; y conformación cromosómica, de la que depende la interacción entre regiones distantes incluso en cromosomas distintos. Se ha observado que su estudio es importante para obtener biomarcadores para patologías [22].

  Como hasta ahora la metilación del DNA es el área de la epigenómica más estudiada, se le ha dado el nombre de metilómica [23]. *En este TFM se integrará solo metilómica.*

- **Proteómica**: es la ciencia que estudia el perfil de expresión proteica en organismos, tejidos o células, en condiciones específicas. También estudia las isoformas proteicas, las modificaciones que sufren las proteínas, sus funciones y las rutas en las que participan. Destaca en el estudio de biomarcadores [24].

- **Metabolómica**: es la ciencia que estudia los metabolitos que hay en organismos, tejidos o células, en determinadas condiciones

experimentales. Se entiende por metabolitos los intermediarios metabólicos, que son moléculas intermediarias en las rutas metabólicas; hormonas; o metabolitos secundarios [25]. Actualmente la **lipidómica** se considera como una rama de la metabolómica y estudia el perfil lipídico en condiciones específicas. En los últimos años empieza a cobrar importancia en el estudio de biomarcadores para patologías [26,27].

- **Interactómica**: ciencia que estudia la interacción entre las moléculas en un sistema biológico [5,28].

- **Secretómica**: ciencia que estudia las proteínas que secreta un organismo, célula o tejido en determinadas condiciones experimentales. Está cobrando importancia en el estudio de biomarcadores de medibles en muestras de fácil obtención [29].

- **Citómica:** ciencia que estudia los tipos de células distintas que forman los organismos o tejidos. Ayuda a entender los procesos bioquímicos y los fenotipos en las células individuales [30].

3.1.3-Algunas ómicas que estudian componentes externos del individuo:

- **Microbiomica:** es la ciencia que estudia el material genético de los microorganismos que viven en un nicho específico dentro de otro organismo. Puede tener valor diagnóstico en patologías.

- **Farmacogenómica:** es la ciencia que estudia los genes que se ven afectados en un organismo en respuesta a un determinado fármaco [5].

3.1.4-Integración de datos Ómicos

Es un mecanismo de ampliación de conocimientos aplicado por la Biología de Sistemas, que se ha desarrollado en los últimos años gracias a las técnicas de última generación. Consiste en analizar cantidades masivas de datos, obtenidos de distintas ómicas, de forma conjunta, para poder observar la interacción de los elementos en los distintos niveles del sistema [2]. Al analizar las ómicas de forma integrada se obtiene información adicional a la obtenida analizando cada ómica de forma independiente, y se puede observar la complejidad del sistema [13].

3.1.5-Conceptos importantes

- **Métodos de representar la interacción entre moléculas:** el uso de estos métodos permite extraer información de la cantidad masiva de datos

inicial, incrementando el poder estadístico del análisis mediante la reducción de hipótesis. Hay dos formas de graficar esta interacción:

- **Pathways o rutas**: tienen estructura esquemática y lineal. Describen las reacciones bioquímicas que ocurren en un determinado proceso. Dan información de cómo se organizan los componentes celulares en un determinado sistema biológico o en una ruta de reacción. Normalmente encontramos rutas de señalización, metabolismo, apoptosis, entre otras. En los pathways se incluye información sobre la interacción entre ácidos nucléicos, proteínas y metabolitos [31].

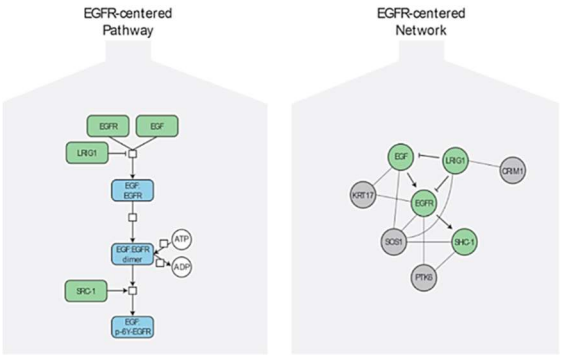


Figura 2: Diferencias estructurales entre Pathways y Networks [31].

- **Networks o redes:** describen la interacción entre moléculas, normalmente gen-gen o proteína-proteína, pero también pueden incluir metabolitos. Su estructura está formada por nodos que representan los genes o proteínas, y vectores que conectan los nodos entre sí cuando existe una relación entre ellos. Hay tres tipos de vectores, los que tienen forma de línea sin dirección (undirected edges), que representan interacción física entre los nodos; las que terminan en punta, y que representan una relación de activación (directed edges); y las que terminan con una línea vertical, que representan una relación de represión de la actividad (directed edges) [31,32].

  - **Network Clustering**: es el proceso de agrupar nodos de los networks en sets, clusters o módulos, donde los nodos son similares de alguna forma [31].
  - **Pathway o Network enrichment, integración de Pathways o Networks**: es el proceso de detectar qué rutas o redes están siendo afectadas por la patología. Este tipo de análisis permite estudiar las rutas funcionales afectadas, y detectar nuevos biomarcadores y dianas terapéuticas. Tras anotar las rutas o redes afectadas es posible visualizarlas con herramientas específicas [33].

3.2-Repositorios de datos públicos disponibles y dificultades de la búsqueda de datos.

Actualmente se ha puesto a disposición de la comunidad científica repositorios de datos de acceso público, con el objetivo de que sea posible reproducir y verificar los descubrimientos, además de acelerar el avance de la ciencia hacia una medicina de precisión. Sin embargo, esto ha generado ciertos problemas de ámbito ético, legal y tecnológico [34,35].

Para solucionar los problemas de ámbito ético y legal, cada repositorio indica su política de acceso, uso y difusión de los datos que pone a disposición de los usuarios [34,35]. Ver *Anexo 1: Tipos de repositorios según su política de uso de los datos.*

Los otros conflictos que han surgido son tecnológicos. El primero es la disponibilidad de múltiples tecnologías e instrumentos generadores de datos, además de distintos formatos de almacenamiento de los mismos. Esto ha supuesto que sea muy importante definir correctamente el tipo de datos que se está almacenando, el formato y la técnica con la que se han obtenido; ya que, a la hora de usar datos del mismo tipo en un estudio, obtenidos de diversas fuentes, es esencial que coincidan en la técnica de obtención, y si es posible en formato, o bien, que se pueda transformar todos los datos al mismo formato [34].

El segundo problema tecnológico que ha surgido ha sido la necesidad de disponer de una gran capacidad computacional de almacenamiento y de procesamiento de datos [36,37]. Para solucionarlo en los últimos años ha aparecido la posibilidad de almacenar datos y trabajar en la nube en una máquina virtual, solución que se adoptó para realizar la integración de datos multi-ómicos.

Hay múltiples repositorios disponibles en la red, Algunos se pueden clasificar según la ómica de la que contengan datos, por ejemplo:

- ***ProteomeXchange****:* consorcio que agrupa varios repositorios de proteómica como PRIDE o iProx, entre otros. Permite buscar los datos de proteómica dentro de los repositorios que agrupa, los cuales se encuentran clasificados principalmente por especie, proyecto, condiciones experimentales y técnicas de obtención [38].

- ***Expression Atlas****:* repositorio con datos de transcriptómica para distintas condiciones y especies, obtenidos con diversas técnicas [39].

- ***methDB**:* repositorio con datos de metilómica para distintas condiciones y especies, obtenidos con diversas técnicas [23].

Otros repositorios almacenan datos ómicos de una patología específica o un grupo de patologías relacionadas, para una única ómica o para varias ómicas. Algunos ejemplos son:

- ***CPTAC Data Portal**:* repositorio de proteomas para distintos tipos de cáncer, obtenidos a partir de varios estudios mediante espectrometría de masas. Los datos que contiene proceden del Proteomics Tumor Analysis Consortium (CPTAC) [40].

- ***cBioPortal**:* entre otros servicios es un repositorio de datos genómicos pertenecientes a varios proyectos y tipos de cáncer. Se encuentran, dependiendo del tipo de cáncer, datos sobre mutaciones, número de copias(CNV o CNA), transcriptómica, metilómica y proteómica [41].

- ***GDC Data Portal**:* repositorio de datos de múltiples ómicas para distintos tipos de cáncer, obtenidos a partir de varios programas, entre ellos The Cancer Genome Atlas (TCGA), y mediante distintas técnicas. Posee datos de acceso libre y otros de acceso controlado por medio de dbGap [42].

- ***Broad GDAC Firehose**:* entre otros servicios es un repositorio de datos pre-procesados de acceso libre pertenecientes a múltiples ómicas para distintos tipos de cáncer, obtenidos específicamente por el programa TCGA mediante distintas técnicas. En Broad GDAC Firehose se filtran los datos antes de publicarlos eliminando réplicas o muestras erróneas. De esta plataforma se extrajeron los datos que se han usado en el TFM [4].

- ***LinkedOmics**:* Entre otros servicios funciona como repositorio de datos para distintos tipos de cáncer, múltiples ómicas y obtenidos por varias técnicas. Los datos que contiene proceden del TCGA de forma mayoritaria, pero también se han añadido datos de CPTAC [43].

Hay repositorios que almacenan datos ómicos de múltiples ómicas, obtenidos mediante diversas técnicas y en distintas condiciones, como, por ejemplo:

- ***Database of Genotypes and Phenotypes (dbGap)**:* repositorio de datos de humanos para genómica, transcriptómica y epigenómica. Requiere permiso de acceso por parte de su comité. Pertenece al National Center of Biothecnology Information (NCBI) [44].

- ***Gene Expression Omnibus (GEO)****:* repositorio de datos de acceso libre para genómica, transcriptómica, epigenómica y proteómica; de diversas especies y obtenidos mediante distintas técnicas [45].

- ***BioStudies****:* es tanto un consorcio de repositorios como un repositorio en sí mismo de datos de acceso libre para genómica, transcriptómica, epigenómica, metabolómica y proteómica; de diversas especies y obtenidos por distintas técnicas [46].

En integración de datos ómicos se pueden integrar redes de interacción, relación o reacción molecular, también conocidas como pathways y networks, que nos permiten estudiar alteraciones funcionales [47]:

- ***Kyoto Encyclopedia of Genes and Genomes Pathways (KEGGpathways)****:* es un repositorio de pathways creados de forma manual que incluyen interacción, relación o reacción molecular, que permite visualizarlos en forma de mapas, y que da un código a cada molécula y pathway denominado código KEGG [47].

- ***Reactome Pathway Database (Reactome)****: e*s una plataforma que contiene herramientas de visualización, integración y análisis de pathways, y que además actúa como repositorio de pathways [48].

- ***The Network Data Exchange(NDEX)****:* es un consorcio de repositorios de pathways al que se accede desde la plataforma ndexbio, antes conocida como ***NCI-Nature Pathway Interaction database*** (PID). Esta herramienta permite descargar los pathways desde la plataforma o te da acceso a un dropbox de Networks donde puedes acceder a los pathways que deseas e incluso subir los pathways propios, teniéndolos de acceso privado o público según se desee. Además, permite descargar o subir datos fácilmente desde la herramienta de análisis y visualización de redes **Cytoscape**. Desde ella se puede acceder a pathways almacenados en los repositorios Reactome, KEGGpathways entre otros [49].

Finalmente, se debe comentar la existencia de Database Commons, un catálogo de repositorios de datos gestionado por Nacional Genomics Data Center de China [50]. Por otra parte, existen catálogos de recursos, que contienen tanto bases de datos como herramientas de análisis bioinformático, algunos de ellos son OPENEBENCH y Bio.Tools [51,52].

3.3-Clases de métodos de integración de datos ómicos.

Los algoritmos para la realización de integración de datos se pueden clasificar según las clases de métodos que utilizan. En muchos casos un mismo algoritmo puede estar basado en más de una clase de método, ya que su workflow puede constar de varios pasos, por lo que se puede decir que las clases no son excluyentes. Se verán paquetes que usan estos métodos en el siguiente apartado. Las clases son [3]:

- **Métodos basados en Redes (Network-based methods)**: Son métodos que estudian la relación entre los datos, crean subredes y estudian la interacción entre ellas. Usan métodos de inferencia bayesiana para crear redes bayesianas o bien métodos heurísticos, como modelos Markov [3] o Artificial Neural Networks(ANN) [53]. Permiten estudiar las redes de interacción, comprobar la correlación funcional entre las distintas ómicas, visualizar los procesos afectados en determinada enfermedad, y estudiar posibles biomarcadores y medicamentos para una determinada patología [3].

- **Métodos basados en Clusters (Clustering methods)**: son métodos que agrupan las muestras según la semejanza de las variables que se han medido, entre ellos encontramos k-mean clustering. Permiten descubrir subtipos dentro de una patología y observar las diferencias entre ellos [3,2].

- **Métodos basados en Deep learning (Deep Networks methods)**: son métodos que crean subredes y estudian la interacción entre ellas. Utilizan métodos de deep learning, para integración destaca Autoencoders. Se utilizan para identificar subtipos de la patología de estudio o bien para extraer características como driver genes, biomarcadores de prognosis y respuesta medicamentos [3,54,55].

- **Métodos basados en extracción o selección de características (Featured Extraction Methods)**: Son métodos que reducen las dimensiones del dataset seleccionando las variables más relevantes para la condición de estudio, entre ellos encontramos como método no lineal Autoencoders [54,55,56], Random Forest [53] o Support Vector Machine Recursive Feature Elimination(SVM-RFE) [57], y como métodos lineales los shrinked methods como regularización de lasso [58]. Permiten estudiar subtipos dentro de la patología, obtener driver genes, pathways y posibles biomarcadores [3].

- **Métodos basados en transformación de características (Featured Transformation methods):** Son métodos que obtienen nuevas variables a partir de las variables medidas, estas nuevas variables representan las características más importantes de las variables iniciales, e incluso pueden dar lugar a obtener información nueva. Entre ellos encontramos SVM, modelado de variables latentes, regresión de componentes principales (PCR), mínimos cuadrados parciales (PLS) o Análisis de Correlación Canonica regularizada [3, 10, 58,59,60]. Permiten reducir las

dimensiones de los datasets creando nuevas variables, permiten realizar subtipados, y encontrar biomarcadores [3].

- **Métodos basados en factorización (Factorization methods):** son métodos que descubren relaciones muy complejas ocultas, factores, en grandes cantidades de datos. Normalmente son relaciones que son muy difíciles de encontrar con los métodos tradicionales. Permiten realizar subtipados, y encontrar biomarcadores [3].

Otra forma de clasificación de las clases de métodos es:

- **Métodos supervisados**: son aquellos que usan los datos introducidos, características predictoras o inputs, para predecir un determinado valor diana, característica respuesta o output. Por tanto, son métodos que determinan la relación entre las variables predictoras y la respuesta, y crean modelos predictivos a partir de ella [61.62].

- **Métodos no supervisados**: Son aquellos que buscan resumir los datos iniciales extrayendo información para finalmente crear un modelo descriptivo. Se suelen utilizar para obtener patrones, y son los que se va a usar para obtener driver genes y subtipado [61.62].

Finalmente, las últimas formas de clasificar los algoritmos son según las ómicas que pueden integrar, o bien, según el tipo de información que se puede extraer de su utilización. En este último tipo de clasificación se encuentra las siguientes clases:

- **Subtipado:** integración de datos ómicos para descubrir subgrupos de pacientes con la misma patología pero que habitualmente presentan diferente prognosis, o distintos resultados terapéuticos frente a determinado medicamento.

- **Descubrimiento de biomarcadores:** integración de datos ómicos para detectar características ómicas que permiten diagnosticar, evaluar el estado de una enfermedad, pronosticar la evolución de dicha enfermedad, susceptibilidad de padecerla, o bien, detectar la exposición a determinado agente o patógeno.

- **Análisis de pathways o redes:** integración de datos ómicos que permite descubrir la relación entre variables medidas en las distintas ómicas, como genes, proteínas, metabolitos u otras biomoléculas, para una determinada condición.

- **Reutilización o descubrimiento de medicamentos:** nuevos medicamentos, o bien, medicamentos ya existentes utilizados para tratar otras patologías [3].

Habitualmente un mismo algoritmo de integración puede extraer información para varios objetivos, pero solo si lo aplicas en un workflow con otros

procesamientos. *La mayoría de algoritmos extrae características específicas de la patología, driver genes, que luego se pueden utilizar para subtipado, estudio de nuevos biomarcadores, dianas terapéuticas, obtención de redes, o estudios de respuesta a medicamentos, los resultados que se quiera obtener condicionarán los procesos extra que se deban utilizar* [61].

3.4-Herramientas y paquetes para la integración de datos ómicos.

Actualmente hay disponibles múltiples recursos para realizar un correcto análisis de datos ómicos. Dichos recursos pueden aparecer en formato de herramientas online, o bien, como paquetes para R o python. Para explicar cada recurso se indicará el método que utiliza, la información que se puede extraer con dicho método, y el grupo de ómicas que se pueden integrar usándolo. Se separarán las herramientas en función de si sus usuarios objetivos son bioinformáticos, o bien, científicos con bajo conocimiento informático.

Algunas de las herramientas desarrolladas para uso de los bioinformáticos son:

- **iClusterPlus**: es un paquete del proyecto bioconductor de R que permite integrar distintas ómicas para obtener gene drivers y realizar subtipado. Si se combina con un análisis de supervivencia permite obtener biomarcadores [3, 54, 61,63]. Maneja datos de genómica (Mut y CNV), proteómica, transcriptómica y metilómica, obtenidos mediante arrays o secuenciación de última generación (NGS o HTS), pero no maneja missing values. Este paquete tiene tres métodos [2,59,63] de integración pertenecientes a la clase de métodos de transformación de variables [3] y métodos no supervisados [61]. De los tres, el último que se ha desarrollado es iClusterBayes, el cual genera la misma calidad de resultados que los otros dos métodos, pero reduce significativamente el tiempo del análisis y las necesidades computacionales [59,63]. Por ello iClusterBayes será el método de integración multi-ómica que se usará en este TFM.

- **Similarity Network Fusion (SNF):** esta herramienta está disponible como paquete de R o de MATLAB, y permite integrar información clínica, transcriptómica, metilómica y datos en imagen, las ómicas pueden ser obtenidas por arrays o secuenciación [64]. Pertenece a los métodos basados en redes [3] y a los métodos no supervisados [54] y no maneja missing values[65]. Se utiliza para subtipado [54].

- **MixOmics**: es un paquete del proyecto bioconductor de R. Permite integrar datos de transcriptómica obtenidos por NGS, metabolómica, proteómica, microbioma, metagenómica e imágenes,[10] y maneja missing values[54]. Tiene ocho métodos distintos de análisis de multivariantes para realizar subtipado u obtener biomarcadores para una determinada patología [10]. Todos sus métodos pertenecen a los métodos de transformación de variables [3,54]. pero se pueden agrupar en dos grupos según si son métodos supervisados o no supervisados [10,54,60].

- **Pathway Recognition Algorithm using Data Integration on Genomics Models (PARADIGM):** es un paquete desarrollado en lenguaje Python[54,61] con un algoritmo que permite inferir, mediante integración de datos multi-omicos, los pathways afectados en un paciente con una determinada patología, lo cual puede tener información prognóstica [54,66]. Permite integrar genómica (CNV), transcriptómica, y epigenómica, obtenidas mediante array o secuenciación [66]. Su algoritmo pertenece a los métodos basados en redes y a los métodos no supervisados [54].

- **Multi-omics factor analysis (MOFA)**: es un paquete desarrollado tanto en lenguaje *R* como en *Python* [54,67] con un algoritmo que permite extraer relaciones complejas entre las ómicas cuando otras herramientas no son capaces. Realiza *imputación* de datos para corregir los *missing values* y *reconoce los outliers* [67]. Permite integrar datos de *genómica(Mut), transcriptómica y epigenómica* [67,68] obtenidos tanto por array como por NGS[67]. Pertenece a los *métodos basados en factorización*, a los *métodos basados en transformación de variables* [3] y a los métodos no supervisados [54,61]. Se usa para descubrimiento de *biomarcadores* y para *subtipado* [61]. También se puede utilizar para *análisis multi-ómicos single-cell* [67], aunque para este uso se ha desarrollado MOFA+ que corrige ciertas deficiencias de MOFA [68].

- **Multi-Omics Late Integration (MOLY)**: es un paquete desarrollado en lenguaje *Python* [61]. Esta herramienta permite crear, mediante integración de datos ómicos, un modelo para *predecir la respuesta a un medicamento* para una determinada patología [61,69]. Se clasifica como un *método basado en Deep Networks* [69] y como métodos supervisados [61]. Integra *genómica (Mut y CNV), transcriptómica(mRNA), y proteómica* obtenidas mediante *arrays o NGS* [61,69].

- **NEighborhood based Multi-Omics clustering (NEMO)**: es un paquete desarrollado en lenguaje *R* [54], cuyo algoritmo permite integrar datos de *transcriptómica y metilómica*, obtenidos mediante *array o NGS*, para realizar *subtipado*. Permite manejar *missing values* [54,70], y se clasifica como *método basado en clustering* [3,54] y métodos no supervisados [54].

- **MethylMix**: es un paquete desarrollado en R cuyo algoritmo permite extraer los griver genes con una metilación diferencial, altamente relacionada con la variación de la expresión génica en dichos genes en la condición patológica. Integra datos de *metilómica* y *transcriptómica* obtenidos por arrays o por NGS. Se clasifica entre los métodos basados en extracción de características y en los métodos supervisados. Permite hacer filtrado e imputación de los missing values con algunas de sus funciones. Sus resultados en Diferential metilation values(DM-values) permiten un buen subtipado mediante técnicas de clustering [9].

- **NetDx**: es un paquete desarrollado en lenguaje *R* [61] que permite integrar *genómica (Mut y CNV), transcriptómica, metilómica y proteómica* [61], obtenidas mediante *arrays o NGS*, para realizar *subtipado* [61,71] y

extraer las características específicas de los mismos. Además, a partir de las características obtenidas, si se aplican otros métodos, permite estudiar la *respuesta a medicamentos o biomarcadores de prognosis* [71]. No maneja *missing values* y se clasifica entre los *métodos basados en extracción de características* [3] y métodos supervisados [61,71].

- **Amaretto**: es un paquete del proyecto Bioconductor de *R* [72] que permite integrar *genómica (CNV), transcriptómica (mRNA) y metilómica, o*btenidos mediante *array y NGS* [61,72], para obtener características específicas que permiten obtener biomarcadores de diagnóstico y pathways específicos de la patología [61,72,73]. No maneja los missing values [73] y pertenece a los métodos basados en redes [3], y a los métodos no supervisados [61].

- **DrugComboExplorer**: es un paquete desarrollado en java y Python [61], cuyo algoritmo permite extraer, mediante integración de datos multi-ómicos, las redes de señalización afectadas por la patología en un paciente, y estudiar combinaciones de medicamentos que puedan actuar sobre dichas redes de señalización [74]. Este algoritmo integra datos de genómica (DNA y CNV), transcriptoma (mRNA), y metiloma, obtenidos por array o NGS [3, 74]. Se clasifica en los métodos basados en redes [3] y métodos supervisados [61].

Ver tabla resumen en *Anexo 2: Resumen de herramientas de integración*.

Además, cada vez son más las herramientas disponibles para usuarios con bajo conocimiento informático. Se pueden separar en dos grupos:

- Herramientas para cualquier condición:

  - **GalaxEast**: es una plataforma web que pertenece al proyecto Galaxy, y que da acceso a las herramientas del proyecto Galaxy desarrolladas para el análisis de las ómicas de forma independiente, y a herramientas de integración de datos multi-omicos(OpenOmics y primo-multiomics) [75,76,77]. GalaxEast permite que el usuario cree workflows con las herramientas disponibles, le ofrece almacenamiento para los datos a analizar, y ejecuta en sus servidores el workflow creado [75].

  - **T-Bioinfo Platform**: es una plataforma web premium, que contiene herramientas para el análisis de ómicas de forma independiente y para integración de datos ómicos. Esta plataforma utiliza las herramientas en workflows bajo una interfaz gráfica que es accesible para personas sin conocimiento bioinformático [78].

  - **OneOmics$_{TM}$Suite**: es una plataforma premium que permite analizar de forma independiente e integrar genómica, transcriptómica, metabolómica y proteómica, esta última obtenida por espectrometría

de masas. Permite localizar biomarcadores para la patología de estudio [79].

- **QIAGEN Ingenuity Pathway Analysis (IPA)**: es una plataforma premium que permite tanto analizar de forma independiente como integrar genómica (SNP), transcriptómica, metabolómica y proteómica obtenidos mediante arrays o NGS. Permite localizar biomarcadores y dianas terapéuticas. Este software permite comparar los resultados finales con los obtenidos por otros usuarios para estudios iguales o similares, para comprobar que son correctos [80].

- Herramientas para integración de datos ómicos de cáncer:

  - **UCSC TumorMap:** es una plataforma web que permite integrar datos ómicos de cáncer y que localiza patrones de similaridad entre las muestras para realizar subtipado y detectar posibles biomarcadores. Para ello la plataforma realiza una representación gráfica interactiva de las muestras, que permite visualizar juntas aquellas con las mismas características en la variable que se elija. Es un método basado en clustering y unsupervised method, e integra genómica (Mut, CNV, SNP), transcriptómica, metilómica y proteómica, obtenidas por array o por NGS [81].

  - **FireBrowse**: es una plataforma web interactiva que analiza 38 tipos de cáncer cuyos datos se han obtenido en el proyecto TCGA. Tiene disponible la herramienta iCoMut que permite visualizar, al mismo tiempo, los resultados del análisis de cada una de las ómicas. El usuario puede interactuar con las pistas de resultados seleccionando cada vez una única ómica, y las muestras analizadas se agruparán en clusters por semejanza de perfil en la ómica, lo cual permite integrar dicha ómica con las demás de forma visual. Así el usuario puede realizar su propio análisis [54,82].

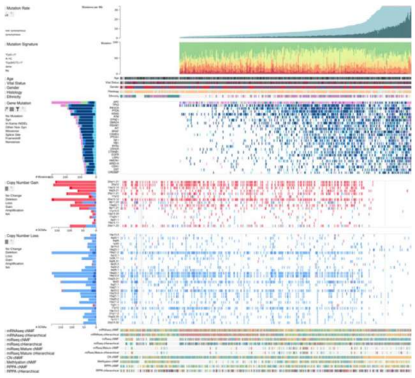


Figura 3: Visualización de ómicas mediante iCoMut. Muestras ordenadas por tasa de mutación [82].

  - **LinkedOmics**: Es un portal web que actúa, por un lado, como base de datos, y, por otro, pone a disposición del usuario tres herramientas para realizar un análisis de integración de datos completo. El análisis lo puede realizar con los datos que almacena la plataforma, o bien, y con datos que

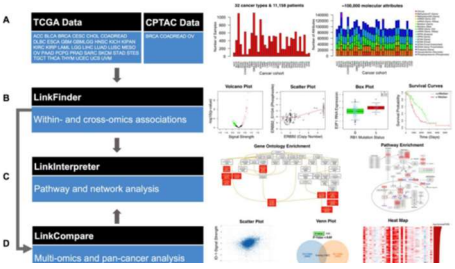


Figura 4: Flujo de trabajo de las herramientas de LinkedOmics y resultados que devuelve cada herramienta [43]

suba el propio usuario. Las herramientas son:

- LinkFinder: Permite que el usuario busque características específicas entre los resultados de la integración. Es la herramienta que se debe usar para encontrar biomarcadores, visualizar el efecto de determinada mutación sobre las demás ómicas, o estudiar la supervivencia de los subtipos.
- LinkInterpreter: Permite realizar, sobre los resultados de la integración, enriquecimiento con Gene Ontology, integración de pathways, y obtención de redes afectadas por la patología.
- LinkCompare: Esta herramienta gráfica los resultados de la integración obtenidos por LinFinder, deseados por el usuario [43,54,83].

- **cBioPortal**: Este portal contiene datos de distintos tipos de cáncer procedentes de múltiples proyectos, y además, permite al usuario subir sus propios datos para analizarlos. Esta plataforma permite, por un lado, visualizar los resultados de la integración de datos almacenados en la plataforma, son datos pertenecientes a un estudio, o incluso a más de uno. Por otro lado, permite realizar el análisis de integración de datos multi-ómicos para datos propios, incluyendo análisis de supervivencia, enriquecimiento, obtención de pathways y redes alteradas [41,84].

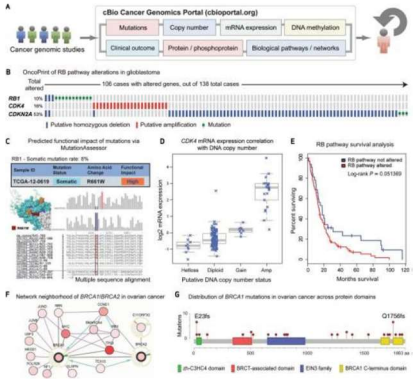


Figura 5: Imagen con algunas de las representaciones de datos que realiza cBioPortal. A) Omicas integradas; B) Pathways Alterados; C) Enriquecimiento; D) Estudios de correlación de datos de distintas ómicas; E) Análisis de supervivencia; F) Obtención

Estas son solo algunas de las múltiples herramientas de integración de datos ómicos disponibles. Se han seleccionado de forma que hubiera por lo menos un ejemplo de cada clase de método explicado en el apartado 3.3, además de tener en cuenta el número de citas con respecto a la fecha de publicación. Para encontrar otras herramientas se puede visitar catálogos como OPENEBENCH y Bio.tools [51,52], buscar en la bibliografía, o bien en revisiones sobre el estado del arte de la integración de datos ómicos.

Hay algunos métodos de integración de datos ómicos que se han definido en papers, cuyos autores no han desarrollado herramientas para aplicarlos, o incluso hay métodos que para aplicarlos se usan paquetes no específicos para ciencias [3].

Estas herramientas demandan muchos recursos computacionales, por lo que se necesita un equipo de alto rendimiento para utilizarlas o bien usarlas en un servidor externo mediante una máquina virtual [85]. Otra opción es usar alguna

de las herramientas para no bioinformáticos que te dan almacenamiento y realizan el análisis en sus servidores [43,83].

3.5-Soluciones para manejar el Big Data.

El desarrollo de las tecnologías de última generación ha generado una acumulación masiva de datos que habitualmente se ponen a disposición de los investigadores. Esto hace que los estudios se puedan realizar integrando datos de varias ómicas y usando mayor cantidad de datos por ómica, lo cual supone que sean estudios de Big Data, y, por tanto, se necesitan requerimientos computacionales específicos para realizarlos [86].

Actualmente hay dos formas de cumplir con estos requerimientos, y son las siguientes:

- **Disponer de un ordenador capacitado para realizar ciencia de datos**:
    - Que disponga de un procesador i7 o superior. Cuantos más núcleos y velocidad de procesamiento mejor.
    - Memoria RAM como mínimo de 30 Gb. Cuanta más RAM mejor.
    - Cuanta mayor capacidad de almacenamiento de datos mejor, dependerá del almacenamiento que se vaya a necesitar en el estudio y de si se va a instalar una máquina virtual en el dispositivo. Lo ideal es un disco duro SSD de 1 Tb de almacenamiento [87,88].
    - Para bioinformáticos en específico, se necesitará disponer de software Linux, las opciones más fáciles disponer de él serán, realizando una partición en el dispositivo e instalar linux en una de las particiones; o bien, instalar una máquina virtual con Linux dentro del ordenador con el sistema operativo de uso habitual [88].

- **Disponer de una cuenta en la nube en alguno de los espacios creados para investigación científica.** Estos espacios ponen a disposición de sus usuarios:
    - Capacidad de almacenamiento de datos propios.
    - Datos de acceso público.
    - Herramientas de bioinformática para procesar los datos.
    - Herramientas para crear workflows con las herramientas bioinformáticas deseadas.
    - Una máquina virtual en la nube con los requerimientos que necesites para ejecutar el workflow que hayas creado.
    - Espacios de comunicación con otros investigadores.
    - Algunas incluso te permiten trabajar con R y python desde Rstudio o Jupyter.

Actualmente hay múltiples opciones premium, las más conocidas son **Cloud Life Science** de Google; Amazon Web Services(**AWS**); **Terra** de Broad Institute, Varily y Microsoft; y **Cancer Genome Collaboratory** de la Universidad de Toronto entre otros[85,89,90,91].

Finalmente, otra opción es **utilizar alguna de las herramientas de integración de datos multi-ómicos desarrolladas para usuarios no bioinformáticos** ya explicadas en el apartado 3.4. Algunas permiten crear workflows con las herramientas que ofrecen para luego ejecutarlos en sus servidores en la nube, como **GalaxEast** [75]; otros ponen a disposición del usuario uno o varios pipelines que el usuario puede ejecutar en los servidores en la nube de la plataforma, sobre sus propios datos, o sobre datos disponibles en la misma plataforma, un ejemplo es **cBioPortal** [41].

3.6-Adenocarcinoma de Colon:

El colon junto con el recto dan lugar al intestino grueso, el cual a su vez forma parte del Sistema Digestivo o Gastrointestinal. La mayor parte del intestino grueso pertenece al colon [92]. Su función es extraer el agua y los electrolitos de los alimentos procesados provenientes del intestino delgado, y enviar los residuos al recto donde esperarán hasta ser defecados [92,93].

Los cánceres en colon y recto comparten muchas características, por lo que habitualmente se pueden encontrar referencias a ellos como cáncer colorrectal, y se especifica el tipo, colon o recto, según el origen de las células tumorales. [92] Sin embargo, en este TFM se va a trabajar solo con cáncer de colon.

La incidencia y mortalidad del cáncer de colon en el mundo en 2020, según GLOBOCAN, indica que es el cuarto cáncer con mayor incidencia, con 1.148.515 casos nuevos, lo que supone un 6% del total de nuevos casos; y el quinto con mayor número de muertes en el mundo, con 576.858 muertes, lo que supone un 5,8% del total de muerte por cáncer [94].

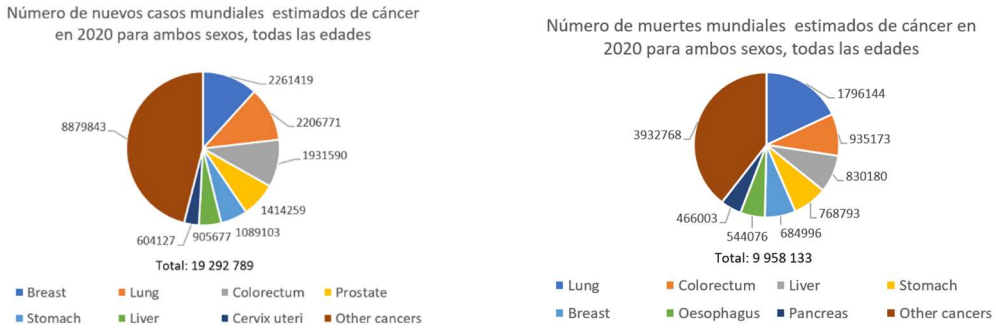


Figura 6: A) Número y porcentaje de nuevos casos por tipo de cáncer en 2020 en el mundo[94]. B) Número de muertes por tipo de cáncer en 2020 en el mundo [94].

Hay varios tipos de cáncer de colon, pero el más habitual es el Adenocarcinoma de Colon, presentándose en el 96% de los casos. Las células en las que se origina son las células glandulares secretoras de la mucosa que lubrica el colon [92]. Hay dos subtipos, adenocarcinoma mucinoso y adenocarcinoma de células en anillo de sello, teniendo este último un peor pronóstico [92,95]

4 Metodología

Se podrá ver la aplicación y los resultados de cada paso en los informes de RMarkdown anexos.

4.1-Datos a integrar:

Los datos que se van a integrar se obtuvieron del proyecto TCGA y pertenecen a 292 pacientes diagnosticados con Adenocarcinoma de colon con datos para las tres ómicas [96]. Cada muestra en cada ómica recibe un código de referencia que permite reconocer a qué participante pertenece, centro del que proviene, tipo de muestra, entre otros. Ver *Anexo 3: Código TCGA.*
Antes de ser compartidos los datos del proyecto TCGA en Broad GDAC firehose, los centros que participaron en el proyecto realizaron un proceso de filtrado de datos con varias etapas, en el que solo se compartieron aquellos que pudieran ser científicamente relevantes [4,97]. Ver *Anexo 4: Filtrado de los datos de TCGA.*

Los datos de TCGA pueden estar en distintos niveles de procesamiento, pero solo los datos de mayor nivel de preprocesamiento (3 y 4) están disponibles sin acceso controlado, por ello son los datos que se van a usar [1]. Ver los workflows de pre-procesamiento de los datos descargados en Firehose en *Anexo 5: Pre-procesamiento de los datos descargados.*

Se van a utilizar los datos de las ómicas: genómica (solo CNA), transcriptómica (solo mRNA) y epigenómica (solo metilación de DNA). Se han elegido estas ómicas porque en múltiples estudios se ha observado que su interacción es esencial para el desarrollo del cáncer, estas ómicas permiten establecer perfiles de tumores detectando los driver genes [1], y dando lugar a que se pueda establecer perfiles distintos, subtipos, dentro de un mismo tipo de tumor, que en muchos casos tienen distinta supervivencia. Esto hace que estas ómicas sean especialmente relevantes a la hora de buscar biomarcadores de COAD [5,6].

Los datos que se descargaron de cada ómica fueron:

- 449 muestras con datos de CNA ya anotados por gen mediante la herramienta GISTIC en G-scores.
- 333 muestras con datos de metiloma anotadas en beta values por probe.
- 500 muestras con datos de mRNAseq en FPKM por gen.

En las tres ómicas los datos estaban normalizados. Ver el pre-procesamiento que tienen los datos descargados en *Anexo 5: Pre-procesamiento de los datos descargados* [96]. Los datos de genómica y metilómica se obtuvieron mediante el paquete ***curatedTCGA*data** [98], y los de mRNAseq se descargan directamente de Firehose[4], ya que el paquete no tenía todos los datos disponibles de este tipo. De RNAseq se seleccionó los datos RSEM en FPKM porque tenían la mejor normalización y había muestras normales para el análisis de expresión diferencial. Por otra parte, se eligió los datos de GISTIC en G-

Scores porque son valores continuos, lo que facilita su uso en los análisis de integración.

4.2-Filtrado de datos:

El filtrado de datos constó de cinco partes:

- Filtrado de muestras con determinado porcentaje de missing values en todas las ómicas: los datos de CNA y mRNAseq no tenían missing values, pero los de metilación de DNA sí. Los dos primeros tipos de datos descargados tienen un nivel de procesamiento mayor que los datos de metilación.

  En este filtrado no se eliminaron muestras, porque ningún dato atípico superó el 70% indicado en Song et al [11] y Yang et al [99], que son dos de los protocolos que se han usado como base para este TFM, ni tampoco el 20% de Chaudhary et al [55], mucho más restrictivo.

- Filtrado de variables con determinado porcentaje de missing values: de nuevo los datos de CNA y mRNAseq no tenían missing values, pero los de metilación de DNA sí. Se observó en un boxplot que en la mayoría de las variables el porcentaje de missing values no superó el 70%, y las que lo hicieron tenían un 100% de missing values.

  Eliminar la gran cantidad de missing values presentes en un data frame aumenta el error en los resultados de los estudios, por ello una buena alternativa es filtrar las variables con más de un porcentaje de missing values e imputar los missig values que quedan. Imputar datos es sustituir los missing values por valores estimados a partir de la misma variable en las demás muestras del data frame. Es importante tener en cuenta que imputar un elevado número de missing values dentro de una misma variable también puede aumentar el error en los resultados, por lo que se debe llegar a un punto intermedio [100,101]. El método de imputación recomendado por Song et al [11], Chaudhary et al [55] y Yang et al [99] es K-Nearest Neighbors.

  En este TFM no se realizó la imputación de datos, porque no se dispone de recursos computacionales suficientes para hacerlo, por ello se eliminarán todas las variables con Missing Values.

- Filtrado de variables con baja expresión: Con 500 muestras se eliminó todas aquellas variables que tuvieran valor 0 en 495 muestras. Se eligió este número de muestras porque se observó un incremento anormal de variables con valores 0 a partir de las 495 muestras.

- Selección de muestras patológicas: para realizar la integración de datos, tanto de las ómicas a pares como de multi-ómicas se seleccionó las muestras patológicas, ya que eliminar las muestras normales permite extraer los driver genes que permiten realizar un subtipado o estudiar los biomarcadores de COAD. Además, se seleccionó en específico de tumor

primario, ya que de los demás tipos patológicos solo había 2 muestras en total para mRNA, por lo que mantenerlas sólo introduciría bias en el análisis. Tras este filtrado se tuvieron 449 muestras de CNA, 294 de metilación y 498 de mRNA. Para realizar este filtrado se usó la función ***TCGAsplitAssays*** del paquete ***TCGAutils*** [102] y la función **str_split** del paquete ***Stringr***[103].

- Selección de muestras que tienen datos para las tres ómicas a integrar: Tras este filtrado se obtuvieron 292 muestras de los mismos participantes en cada ómica. Para realizar este filtrado se usó la función ***MatchedAssayExperiment*** del paquete ***MultiAssayExperiment***[104] y la función ***match*** del paquete ***base***[105].

Tras el filtrado, de 20553 genes se redujo a 19804 genes en transcriptoma, y de 485577 probes a 374908 en metiloma.

4.3-Transformación de datos:

Solo se transformaron a escala logarítmica los datos de RNAseq para mejorar aún más la distribución de los datos. Para ello se sumó una unidad a todos los valores antes de hacer la transformación, de forma similar a como hace la transformación logarítmica de *CPM* el paquete *Edge* R, solo que como se tiene valores 0 no es útil sumar solo 0.5, así que se sumó 1 FPKM. Esto evitó los valores negativos que ocurren cuando hay FPKM<1 y los NA que aparecen cuando FPKM=0. Estos valores negativos no son soportados por la función *DGELis*t de *EdgeR* que crea el objeto que usa *limma* en el análisis diferencial [106].

4.4-Análisis de una ómica:

Para hacer el análisis diferencial de expresión y metilación se usaron dos grupos de muestra, normales y patológicas. En mRNAseq 41 normales y 456 patológicas, y en metiloma 38 normales y 294 patológicas. El análisis se realizó mediante el paquete *limma*.

El proceso de obtención de las variables diferencialmente expresadas o metiladas se hará creando primero un modelo de regresión lineal con la función *lmFit* de *limma* para cada gen o probe donde la variable Y será la expresión medida y las variables X serán binomiales con valor 1 para uno de los grupos y 0 para el otro cuyo valor dependerá del grupo al que pertenezca cada muestra. Los parámetros para cada gen del modelo son los parámetros alfa.

A partir de los parámetros alfa se calculan los parámetros beta, que permiten realizar contrastes, que son comparaciones entre los grupos para saber qué genes tienen distinta expresión en cada grupo. En este caso se testa hipótesis nula H0: normales-tumorales=0, es decir, la expresión génica es igual en las muestras normales y en las tumorales. Para ello se calculará el contraste beta=alfa1-alfa2 para cada gen. Para calcular beta se usa la función *contrasts.fit* de *limma,* los valores alfa guardados y la matriz de contraste que indica el contraste a realizar.

Posteriormente, se calcula mediante bayes la probabilidad de que cada uno de esos genes no esté diferencialmente expresado, es decir, que H0 sea cierta. El resultado serán varios parámetros como fold-change, t-value y p-value ajustado FDR al 5%. Se usará la función eBayes. Para ello se usa la función *eBayes*. [6,7]

Finalmente se clasifica con *decideTest* los genes entre no diferencialmente expresados, sobre expresados o hipo expresados; o bien los probes en isla CpG no diferencialmente metilada, hipo metilada o hiper metilada. Para ello se utilizó un nivel de significación de 0.05 como indica Song et al [11].

En el análisis se introdujo 374908 probes para el análisis de metilación diferencial, y 13705 genes en el análisis de expresión diferencia.

4.5-Integración de datos ómicos:

4.5.1-Integración de datos ómicos RNAseq y Metiloma mediante correlación inversa de variables.

La función *MethylMix* del paquete con el mismo nombre obtiene los driver genes del perfil metilómico haciendo primero un modelo de regresión lineal, $Y=\beta X+\varepsilon$, entre la expresión génica y los niveles de metilación de cada gen en las muestras patológicas. En el modelo, y_i es el valor de $\log_2$FPKM del gen y en la muestra i, y x_i el Beta valor del gen en la muestra i. Si la relación lineal entre X e Y es significativa e inversa, es decir, tienen correlación inversa significativa, el gen al que pertenezcan se definirá como driver gen metilómico, ya que el nivel de metilación tiene un elevado efecto sobre la expresión del mismo gen [107, 108].

Si se aporta a la función los Beta values de muestras normales, a continuación de obtener los gene drivers, calcula la diferencia de metilación de cada gen, Differential Metilation Values(DM-Values), que permiten seleccionar los gene drivers metilómicos relacionados con la patología [109,110].

Los DM-values se calculan haciendo, para cada muestra y gen, la diferencia entre la media de Beta values de las muestras sanas para dicho gen, y la media de Beta values del componente al que pertenece la muestra patológica de estudio en dicho. Un componente es un grupo de muestras con un perfil de metilación similar en un determinado gen. Para obtener los componentes *MethylMix* aplica un proceso semejante a K-mean clustering en cada gen, donde va comparando el Beta value de cada muestra con la media, y si hay varios Beta values muy alejados de la media los separa y hace un nuevo grupo de expresión. Repite este proceso hasta que obtiene por cada gen los grupos de muestras que se están expresando de forma similar, estos grupos son los componentes. [107,108]

Además, se ha observado que los DM-values obtenidos de MethylMix son muy útiles para hacer el subtipado de una patología mediante K-mean Clustering [109,110].

Para aplicar MethylMix se anotó a que gen pertenece cada probe diferencialmente metilado obtenido del análisis con limma, lo cual redujo la demanda computacional del análisis. Se filtró los probes que no tuvieron anotación, y después, se seleccionó de cada gen, el probe con mayor metilación diferencial, menor p-valor. Finalmente se obtuvo un data set con Beta values por gen. A continuación, se seleccionó los genes que tuvieran datos de RNAseq y metiloma, y se ordenaron sus data frames para que coincidieran las muestras y genes en las mismas posiciones. En el análisis se introdujo 292 muestras y 18576 genes, y se obtuvieron 1325 driver genes [107].

4.5.2-Integración de datos multi-ómicos mediante métodos bayesianos.

Lo primero que hace iClusterBayes es obtener los vectores latentes con una técnica análoga al **Análisis de Componentes Principales (PCA). Solo que en este caso lo importante son las variables latentes que están conformadas por los elementos z_i de cada componente principal para una muestra. PCA consiste en encontrar estructuras lineales entre variables dentro de bases de datos con gran densidad, es decir, muchas variables.** Los componentes Z_k (Z_1, Z_2, …, Z_k) representan las combinaciones lineales de las *p* variables. El número, k, de combinaciones lineales que se puede obtener será igual a p, número de variables predictoras. Como no todos los PC aportan información importante al modelo, el eliminar PC innecesarios hace que se reduzcan dimensiones. En iClusterBayes cada variable latente representa los datos de las tres ómicas en una misma muestra, por lo que p será igual al total de variables en las tres ómicas para esa muestra **[58,59].**

Posteriormente, a cada variable se le estimará un modelo de regresión lineal:

$$Y_{jt}=X\Gamma_{jt}\,\beta_{jt}+\varepsilon_{jt},\ j=1,\cdots,p_t;\ t\in(1,\cdots,m)$$

Donde:

- Y_{jt} es un vector con los valores observados para cada muestra i en la variable j del dataset t.
- X es una matriz con una variable latente cada muestra por cada fila, más una columna inicial de valores 1 que permitirá que el modelo tenga parámetro β_0. Para cada muestra X(1, z_i,…, z_k).
- Γ_{jt} es una matriz diagonal para la característica j en la ómica t, con la siguiente estructura: diag(1, γ_{jt},⋯, γ_{jt}) donde γ_{jt} puede tener valor 0 o 1 según si el valor β_{jt} es pequeño o elevado, y por tanto afecta menos o más al modelo de unión obtenido al final.
- ε_{jt} es un vector con el error obtenido para cada muestra de la diferencia entre Y_{ijt} estimado con el modelo y el real.

Los parámetros del modelo se estimarán mediante el modelado de bayes [59]. Para entender la técnica primero se debe recordar el teorema de bayes es:

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{P(B|A)P(A)}{P(B)}$$

Donde para iClusterBayes se redefine A como un vector de parámetros θ, y B como los datos observados Y y otros parámetros del modelo. Entonces:

- θ: variable aleatoria continua con los posibles valores del parámetro a determinar, por ejemplo β_{jt}. En bayes los parámetros se estudian como variables por ello tienen una determinada distribución.
- P (Y_{ijt}, X, γ_{jt}, σ_{jt}^2/ β_{jt}) : likelihood o función de verosimilitud. Se obtiene a partir de los datos observados. La función de verosimilitud es la densidad conjunta de los datos en función del parámetro θ. L(θ)
- P(β_{jt})=distribución anterior del parámetro β_{jt}, a partir de conocimiento previo.
- P(Y, X, γ, σ)=verosimilitud marginal o evidencia. Se obtiene a partir de simulaciones de Cadenas de Markov vía Monte Carlo. No depende del parámetro ni de su distribución inicial.
- P(β_{jt} /Y, X, gamma, sigma)=distribución posterior del parámetro β_{jt} habiéndose tomado en cuenta los datos observados.

Para determinar los valores de los parámetros del modelo para cada característica se va a tomar como que θ es el parámetro que se quiere determinar a partir de los datos observados, usando bayes [111,112].

Con Bayes se combina la información anterior sobre el parámetro P(θ) y la información aportada por Y, para obtener la distribución posterior, que se utilizará como función de verosimilitud para determinar mediante inferencia estadística el estimador de máxima verosimilitud, es decir, el valor del parámetro de interés con mayor probabilidad de ser real según los datos observados, y que es el que maximiza L(θ). De esta forma se estimarán los parámetros del modelo en cada variable. [111,112,113].

Las distribuciones previas propuestas por el modelo de iClusterBayes para los parámetros de las variables continuas, que es el caso de los datos a usar en este TFM, son:[59]

- $\beta_{jt} \sim MVN(\beta_{0t}, \sum_{0t}), \sigma_{jt}^2 \sim IG(\frac{\nu_0}{2}, \frac{\nu_0\, \sigma_0^2}{2}) \sim Bernoulli(q_t)$

Que tras combinarlas con la función de verosimilitud de los datos con respecto a los parámetros dio que σ^2_{jt} tiene distribución posterior gamma inversa con parámetro de escalado ($v_0\ \sigma_0^2 + (y_{jt} - X\Gamma_{jt}\beta_{jt})^T(y_{jt} - X\Gamma_{jt}\beta_{jt})$)/2 y parámetro de forma (v_0+n)/2(v_0+n)/2. La distribución de β_{jt} es normal multivariante. Las distribuciones de γ_{jt} y z_i no están claras por eso es necesario usar el algoritmo Metropolis-Hasting para estimarlas. Para hacer el modelo se debe indicar la probabilidad anterior γ, se usó 0.5 para las tres ómicas, como indica el manual [59,111]. Distribuciones posteriores estimadas:

- $P(\sigma_{jt}^2 | y_{jt}, Z, \gamma_{jt}, \beta_{jt}) \sim IG\left(\frac{\nu_0+n}{2}, \frac{\nu_0\, \sigma_0^2 + (y_{jt} - X\Gamma_{jt}\beta_{jt})^T(y_{jt} - X\Gamma_{jt}\beta_{jt})}{2}\right)$
- $P(\beta_{jt} | y_{jt}, Z, \sigma_{jt}^2, \gamma_{jt}) \sim MVN(m, V)$

$$m = (\frac{\Gamma_{jt}^T\, X^T X \Gamma_{jt}}{\sigma_{jt}^2} + \sum_{0t}^{-1})^{-1} (\frac{\Gamma_{jt}^T\, X^T X \Gamma_{jt}}{\sigma_{jt}^2} + \sum_{0t}^{-1} \beta_{0t})$$

$$V = (\frac{\Gamma_{jt}^T X^T X \Gamma_{jt}}{\sigma_{jt}^2} + \sum_{0t}^{-1})^{-1}$$

- $$P(\gamma_{jt}|y_{jt}, Z, \sigma_{jt}^2, \beta_{jt}) \propto e^{\left(-\frac{(y_{jt}-X\Gamma_{jt}\beta_{jt})^T(y_{jt}-X\Gamma_{jt}\beta_{jt})}{2}\right)} P(,\gamma_{jt})$$

Finalmente, para poder calcular los estimadores de máxima verosimilitud es necesario obtener distribuciones posteriores generalizadas es decir que se aproximen a la distribución real, ya que las distribuciones obtenidas se basan solo en un set de datos. Para hacerlo se usa el método de Cadenas de Markov vía Monte Carlo (MCMC), cuando no se conoce bien la forma de la distribución se calcula mediante el algoritmo Metropolis-Hasting; y si se conoce claramente la distribución, como en β_{jt} y σ_{jt}, se usará el algoritmo de Gibbs, que hace que la distribución posterior generalizada sea semejante a la obtenida experimentalmente [59,111,114].

Las MCMC son cadenas de variables θ en las que cada variable θ depende únicamente de la variable θ anterior. La variable θ de inicio de la MCMC es la variable θ obtenida a partir de los datos de estudio, de la que se tiene una distribución posterior. El tipo de algoritmo usado para obtenerlas depende de si se conoce dicha distribución claramente o no. A partir de la variable θ_0 va generando nuevas variables θ_t y con estudios de probabilidad determina si se aceptan o no, cuando no se aceptan se sustituye el valor de dicha variable nueva por el de la anterior.

Así ocurre tantas veces, iteraciones, como variables queramos que tenga la MCMC. Para obtener la distribución a posteriori generalizada con MCMC se busca llegar a la distribución estacionaria ρ_i, que es aquella en la que no se observa la influencia de la variable θ_0 sobre la nueva variable θ_t, sino que dicha variable solo está afectada por la anterior θ_{t-1}. Esta distribución se usa como distribución a posteriori aproximada a la real para realizar inferencia. Para llegar a ella se necesitan realizar múltiples iteraciones en las que las primeras, burn-in, deben descartarse. Se le indicó a iClusterBayes que hiciera 16000 iteracines burn-in y 10000 iteraciones draw. Se usó menos que Mo et al[2] para reducir el tiempo de ejecución.

Una vez obtenida la distribución posterior generalizada de los parámetros se estima sus valores mediante la estimación de máxima verosimilitud. Después, se obtiene la distribución de unión, joint likelihood, de *y* y *z* para cada característica j en cada ómica. Con la distribución de unión se observa que variables y_j afectan más a las variables latentes, y por tanto, son características ómicas diferenciadoras importantes para realizar el clustering. Se debe recordar que la clave de la integración de las tres ómicas son las variables pues agrupan las tres ómicas. El modelo general de distribución de unión es:

$$P(y_{jt}, z_i|\beta_{jt}, \gamma_{jt}) = \prod^{m}\prod^{n}\prod^{p_i} P(y_{jt}|\, z_i, \beta_{jt}, \gamma_{jt})\, P(z_i),$$

$$i = 1, \dots, n; j = 1, \dots, p_t, t = 1, \dots, m$$

Donde, como las variables son todas continuas, se usará la función de densidad de la distribución normal:

$$P(y_{jt}|z_i, \beta_{jt}, \gamma_{jt}) \propto \sigma_{jt}^{-1} e^{\left(\frac{-\left(y_{ijt} - X_i \Gamma_{jt} \beta_{jt}\right)^2}{2\sigma_{jt}^2}\right)}$$

Y $P(z_i)$ es la distribución inicial del vector latente.

Se conservará como características relevantes j para el clustering aquellas con elevados valores de beta y con gamma=1.

Para separar las muestras se utilizará el método de K-mean clustering.[2] El número de clusters en los que se separarán las muestras será g, siendo g=k+1, donde k es el número de elementos que conforman la variable latente. K-mean clustering tiene dos etapas diferenciadas. La etapa de iniciación consiste en elegir aleatoriamente g muestras que funcionarán de muestras de referencia, posteriormente se calcula la distancia euclidiana, a partir de los vectores latentes, de cada una de las muestras de no referencia con respecto a cada una de las muestras de referencia.

A continuación, se agrupan las muestras con respecto a la muestra de referencia con menor distancia euclidiana. En la siguiente etapa se llevan a cabo iteraciones que solo paran cuando ninguna muestra cambia de cluster. Esta etapa consiste en, primero, obtener un vector por cluster con las medias de los elementos, en cada posición, de los vectores latentes presentes en dicho cluster. Después, se calcula la distancia euclidiana entre el vector latente de cada una de las muestras y cada uno de los vectores latentes medios de los clusters. Posteriormente se reagrupan las muestras en los clusters con los que tengan menor distancia euclidiana. Ahora vuelve a repetirse esta etapa tantas veces como sea necesario hasta que as muestras no cambien de cluster [62].

Al final de la integración iClusterBayes te devuelve las variables, genes en este caso, más relevantes para diferenciar los subtipos y que muestras pertenecen a cada subtipo(cluster).

Como no se sabe cuál es el número de subtipos, g, adecuados para la patología de estudio, hay que realizar la integración de 1 a k veces, siendo k el número de elementos en el vector latente, el número de subtipos a generar menos 1 y el número de veces que se realiza la integración. En este caso se ha integrado 6 veces para generar 6 modelos, el primero con 2 subtipos y el último con 7 subtipos. Se ha elegido 6 porque es el recomendado por el manual [62]

4.6-Métodos de comparación y visualización de resultados.

El mejor modelo de integración de los 6 entrenados, k=1:6, como indica Mo et al[59] se elegirá mediante BIC y el ratio de desviación. Además, como los datos de COAD tienen ruido, y esto hace que BIC tienda a disminuir y el ratio de desviación a aumentar según aumenta k, lo que distorsiona los resultados, se comprobará con heatmaps si el modelo elegido es el mejor, como indica el manual.

- Ratio de desviación o deviance ratio: interpretado en iClusterBayes como porcentaje de variación explicado. Se obtiene dividiendo el log-likelihood ratio, deviance, del modelo obtenido con iClusterBayes a partir de las variables más significativas y el modelo nulo, sin variables predictoras;

entre el deviance del modelo obtenido con todas las variables introducidas y el modelo nulo. El Deviance o log likelihood ratio testa la hipótesis nula de que el modelo sin predictores, nulo, es tan bueno como el modelo ajustado o completo, en cada caso. El logLikelihood será la función de máxima verosimilitud del modelo [115].

- Bayesian Information Criterion(BIC): mide la estimación relativa de la información perdida cuando se usa el modelo que se está estudiando, Por tanto, siempre se elegirá como mejor modelo el que tenga un BIC menor, porque se habrá perdido menos información. Lo usa iClusterBayes para elegir el mejor modelo para cada valor de k, y se usará para elegir la k con el mejor modelo [58].

$$BIC = -2 * logLikelihood + p * \log(n)$$

Donde:
p: número de parámetros del modelo
n: número de muestras

- Heatmap: Se grafica muestras frente a variables(genes) el nivel de expresión, o metilación. Con sus valores transformados a z-scores(datos con media 0 y varianza 1; obtenidos restándole a cada dato la media de su variable y dividiendo el resultado por la desviación estándar de la misma) y ordenados calculando la correlación entre las muestras y características de forma que se agrupen, por un lado, las muestras y variables de mayor nivel de FPKM o beta-values, en cada caso, y por otro, las de menor nivel. Esto permite observar los posibles clusters o subtipos de muestras y elegir qué modelo genera los mejores clusters [116].

Para comparar los resultados de los distintos métodos de análisis se usaron Heatmaps y diagramas de Venn. Estos últimos muestran los genes comunes y no comunes que han aparecido como resultado en los análisis comparados [58].

4.7-Enriquecimiento, integración de pathways y otros datos, y obtención de redes.

Para integrar los pathways se utilizó las herramientas NDEx y Cytoscape las cuales se complementan integrando pathways y permitiendo su visualización. Ya se habló de estas herramientas en el estado del arte como repositorios de distintos repositorios primarios de pathways [49]. Con NDEx también se obtendrán redes por integración de datos de prognosis y medicamentos.

Adicionalmente, se realizará el enriquecimiento, en específico se estudiará la función molecular y posición celular de los driver genes obtenidos en ambas integraciones. Se anotará los términos de Gene Ontology con el paquete *ClusterProfiler* [117].

5 Resultados

5.1. Subtipado:

- ***Análisis de expresión diferencial:***

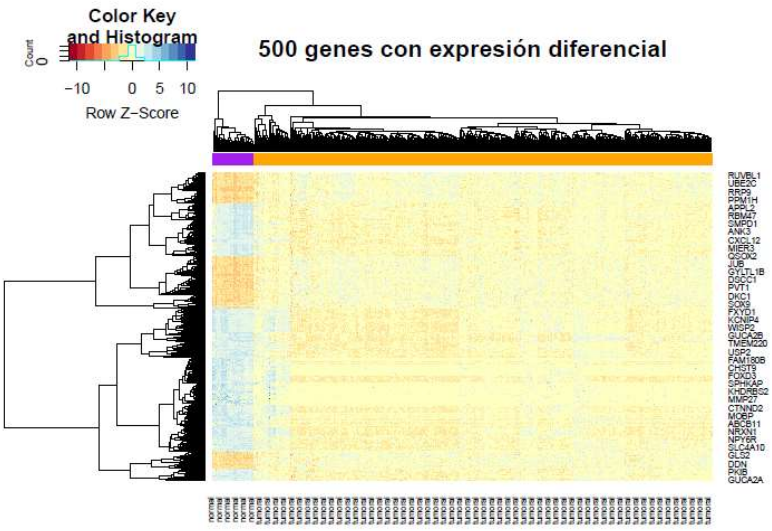


Figura 7: Heatmap de expresión diferencial.

En morado se observan las muestras normales y en naranja las patológicas. A simple vista con el heatmap es difícil diferenciar los subtipos dentro del grupo de muestras patológicas, aún así parecen haber por lo menos tres subtipos de COAD. Para hacer una determinación más exacta habría que hacer un clustering a partir de los genes diferencialmente expresados.

- ***Análisis de metilación diferencial:***

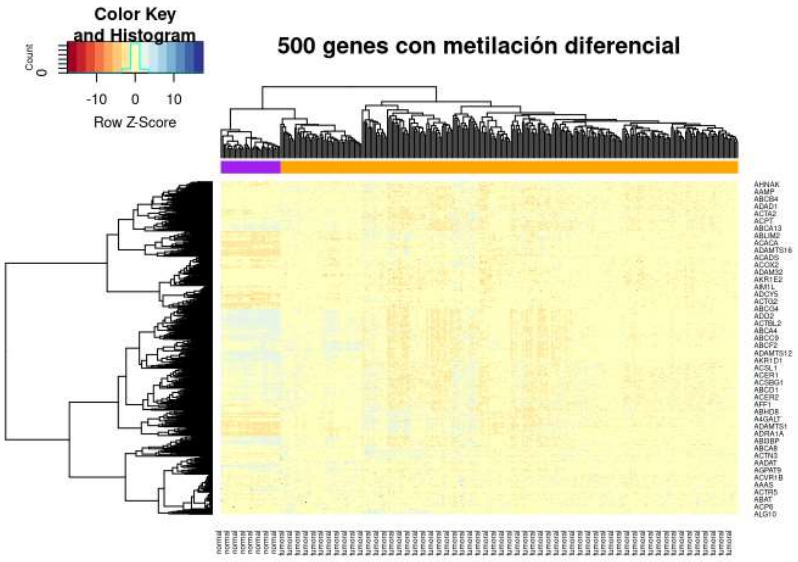


Figura 8. Heatmap de metilación diferencial

Solo con el análisis por metilación diferencial no parece haber subtipos de COAD claros, por lo que parece ser muy adecuado realizar la integración con datos de transcriptómica.

- ***Análisis de integración de transcriptómica y metilómica:***

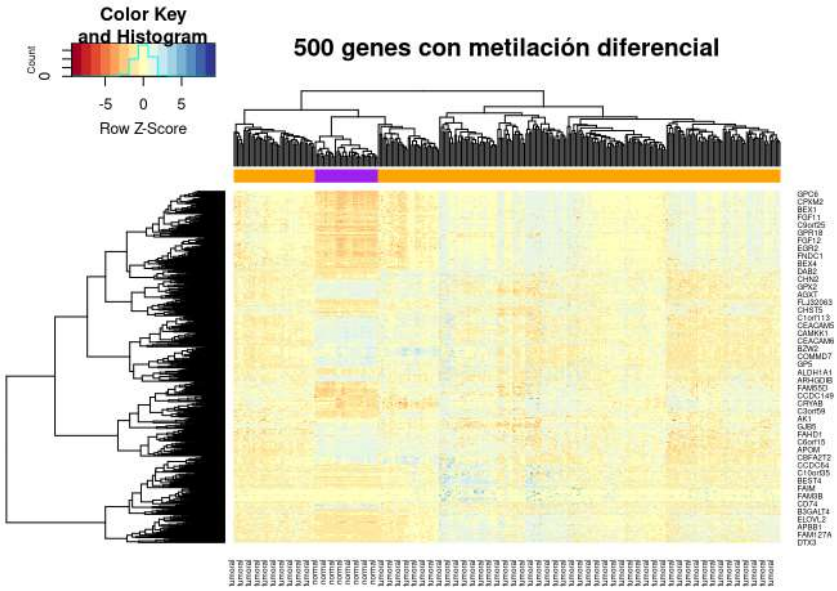


Figura 9. Heatmap de metilación diferencial y efecto funcional

Con los datos integrados de metilación con expresión diferencial parece haber 5 subtipos de COAD, dos más que en expresión diferencial. Se observa como la integración de ambas ómicas mejoró claramente los resultados. A partir de estos driver genes generados se podría hacer un buen subtipado mediante k-mean clustering utilizando los DM-values.

- ***Análisis de integración de multi-ómicas:***

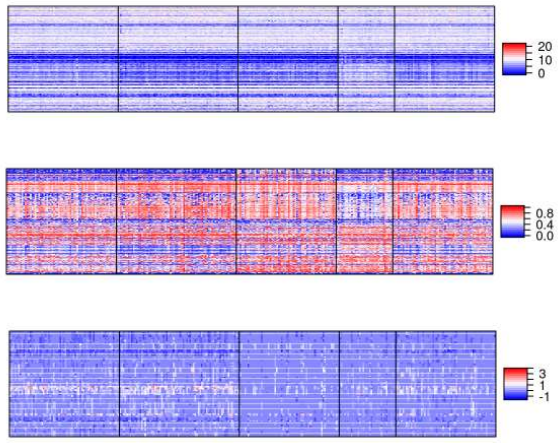


Figura 10: Heatmaps de Transcriptómica, metilómica y genómica de arriba abajo.

Se eligió el modelo K=4 con los 5 subtipos, ya que se diferencian muy claramente en todas las ómicas. Aunque los modelos con 2 y 4 subtipos, k=1 y k=3, también son bastante buenos. Se pudo comprobar como el ruido afectó al valor de BIC y deviance ratio, ya que mostraban que el mejor modelo es k=1. Se puede ver los distintos modelos en el *Anexo 4: Códigos e informes.*

5.2-Driver genes

En el análisis de expresión diferencial se obtuvieron 13.705 genes con expresión diferencial y en el análisis de metilación diferencial se obtuvieron 18.576 genes con metilación diferencial. En el análisis de integración de datos de transcriptómica y metilómica se obtuvieron 1.325 genes con metilación diferencial con efecto funcional. La integración con MethylMix da los driver genes de metiloma, hay que tomar en cuenta que pueden no ser los driver genes con la expresión diferencial más significativa. Por último, en la integración de multi-ómicas se obtuvieron 1.028 driver genes con alteraciones en el número de copias; 1.325 driver genes de metilación diferencial, y 3.073 driver genes de expresión diferencial.

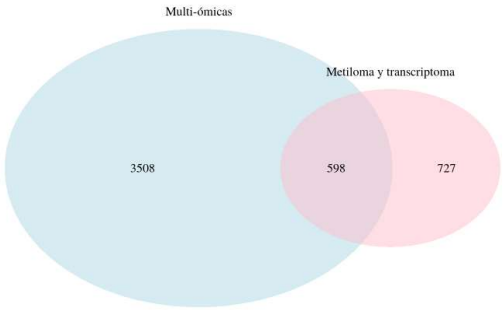


Figura 11: Diagrama de venn de gene drivers obtenidos en ambas integraciones.

En este diagrama de ven se puede observar como la integración de dos ómicas y de multi-ómicas solo comparten 598 driver genes. Para hacer el diagrama se unificó los drivers genes de las tres ómicas generados por iClusterBayes.

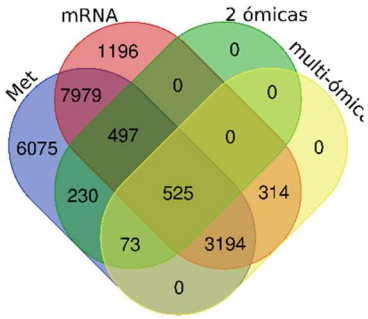


Figura 12: Diagrama de Venn de los genes obtenidos en los cuatro análisis [118].

Se observa cómo el análisis de integración con dos ómicas comparte mayor número de genes con los genes obtenidos en el análisis de metilación diferencial. En cambio, el análisis de multi-ómicas comparte mayor número de genes con los genes de expresión diferencial que con los de metilación diferencial. Por otra parte, también se ve que la mayor parte de los genes obtenidos en el análisis de una sola ómica son descartados por ambos análisis de integración.

5.3-Integración de Pathways

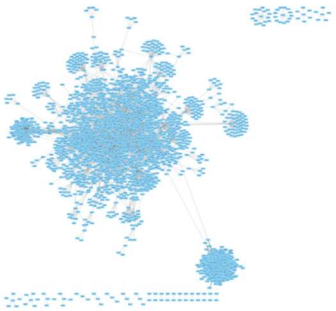


Figura 13: Pathways Reactome [119,120,121]

En la integración de pathways con NDEx a partir de los pathways almacenados en Reactome se obtuvo un network con 2676 nodos y 7584 aristas. Para obtener información de este network se necesitaría hacer un análisis muy exhaustivo para el que no ha habido tiempo en este TFM. Además, se podrían mezclar redes con Cytoscape para incluir anotaciones de repositorios distintos de pathways o de otros tipos de datos.

5.4-Enriquecimiento:

Se estudió el enriquecimiento solo para las dos integraciones:

- Integración de dos ómicas:

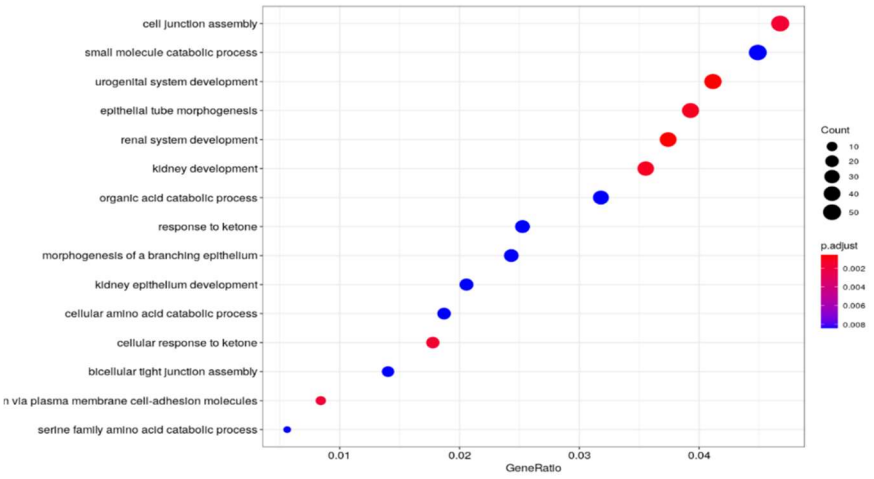


Figura 14: GO terms más significativos en integración de dos ómicas.

Parece que en la integración de dos ómicas los genes con mayor alteración funcional relacionada con el nivel de metilación son los relacionados con el desarrollo ya que aparecen más GO terms relacionados con dicha función.

- Integración de multi-ómicas:

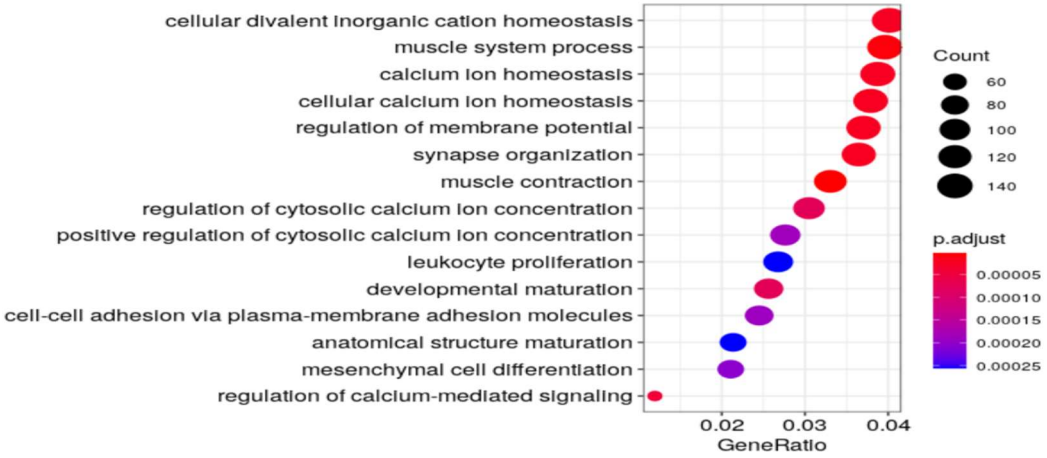


Figura 15: GO terms más significativos en integración de multi-ómicas

Se observa que al integrar además los datos de CNA se obtiene que los genes más alterados suelen ser los relacionados con homeostasis de sales. También se ve que podemos fiarnos más de los resultados de la integración de tres ómicas pues los resultados de enriquecimiento tienen menore p-valores,10^{-5}, en comparación con los p-valores de la integración de dos ómicas, 10^{-3}.

5.5-Integración de datos de supervivencia y medicamentos

- Supervivencia:

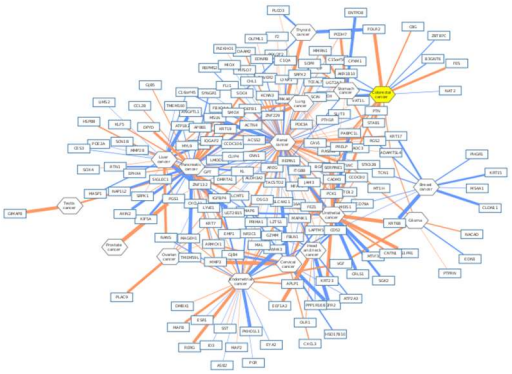


Figura 16: Driver genes prognósticos anotados [119,120,121]. Se ve en amarillo un nodo que conecta a todos los driver genes de COAD que tienen anotaciones de supervivencia. Azul: buena prognosis, naranja mala prognosis.

Se integró, en NDEx, los driver genes que estuvieran en todas las ómicas en la integración de multi-ómicas, con las anotaciones de supervivencia por gen y tipo de cáncer del Human Protein Atlas, y se obtuvo 15 genes con anotaciones de

valor prognóstico. Esta red tiene 195 nodos genes y 301 aristas, la mayoría de los nodos son genes con valor prognóstico para otros cánceres, pero que interactúan con los 15 genes anotados de COAD con valor prognóstico. Ver los genes prognósticos anotados en la siguiente tabla obtenida de NDEx:

Source Node	Interaction	Target Node	score	name	prognosis
		colorectal			
PRKAR2B	correlates-with	Colorectal cancer	6.38E-4	PRKAR2B (corre…	favourable
STAB1	correlates-with	Colorectal cancer	7.59E-4	STAB1 (correlate…	unfavourable
PABPC1L	correlates-with	Colorectal cancer	3.8E-4	PABPC1L (correl…	unfavourable
SERPINE1	correlates-with	Colorectal cancer	2.82E-4	SERPINE1 (corr…	unfavourable
AOC3	correlates-with	Colorectal cancer	3.68E-4	AOC3 (correlates…	unfavourable
UGT2A3	correlates-with	Colorectal cancer	3.31E-4	UGT2A3 (correla…	favourable
ADAMTSL4	correlates-with	Colorectal cancer	1.86E-4	ADAMTSL4 (corr…	unfavourable
NAT2	correlates-with	Colorectal cancer	4.61E-5	NAT2 (correlates…	favourable
FOLR2	correlates-with	Colorectal cancer	8.42E-4	FOLR2 (correlate…	unfavourable
C8G	correlates-with	Colorectal cancer	2.99E-4	C8G (correlates-…	unfavourable
FES	correlates-with	Colorectal cancer	5.78E-4	FES (correlates-…	unfavourable
ZBTB7C	correlates-with	Colorectal cancer	1.59E-4	ZBTB7C (correla…	favourable
PRELP	correlates-with	Colorectal cancer	9.7E-5	PRELP (correlat…	unfavourable
AKR1B10	correlates-with	Colorectal cancer	1.91E-4	AKR1B10 (correl…	favourable
B3GNT6	correlates-with	Colorectal cancer	2.31E-4	B3GNT6 (correla…	favourable

Total Items: 301 (Showing Items: 15)

Figura 17: Tabla de NDEx de genes con anotaciones de supervivencia [119,120,121]

Por lo tanto, esta es una forma de estudiar la bondad como biomarcadores de prognosis de los driver genes.

- Posibles medicamentos:

También se puede integrar los driver genes con los medicamentos existentes para los que son diana, anotados en DrugBank. El resultado fue el siguiente:

Figura 18: Drive genes cómo dianas terapéuticas y posibles medicamentos existentes [119,120,121].

Se encontraron 688 nodos, medicamentos y dianas terapéuticas, y 735 aristas, interacciones entre ellos. En específico hay 58 driver genes que pueden actuar como diana terapéutica para medicamentos que ya están en el mercado con anotaciones en DrugBank.

En cytoscape, al igual que en el visor de NDEx, se pueden agrandar las redes y evaluar las redes con detalle, e incluso se pueden añadir nuevos nodos y edges que se descubran experimentalmente.

6 Discusión

En los dos tipos de integración realizados se observó que hay 5 subtipos de COAD, si se compara con los datos publicados, Guinney et al[121] indica que hay 4 subtipos de cáncer colorectal, esta diferencia se puede deber a añadir las muestras de cáncer rectal al análisis, o bien, a una apreciación incorrecta de la autora de este TFM, que es lo más probable, pues en el estudio de Guinney et al[121] se integraron datos mutacionales, número de copias, transcriptómica, metilación, miRNA, y proteómica, lo que permite obtener resultados más precisos. El siguiente paso para comprobarlo sería hacer un análisis de prognosis a partir de los modelos con 4 y 5 subtipos para determinar cuál es el número de subtipos más adecuado en la integración de tres ómicas. Sin embargo, el tiempo disponible no lo ha permitido. Cabe recordar que se indicó en los resultados que eran buenos los modelos de 2, 4 y 5 subtipos, por lo que hay margen de error por mi parte. Sin embargo, se puede decir que el modelo elegido es relativamente bueno porque extrajo 15 genes con valor prognóstico, pero lo ideal sería observar cuantos biomarcadores prognósticos ya anotados extrae el modelo de 4 subtipos.

Según Xie et al[222], que realizó un análisis de subtipado según CNAs, hay tres subtipos de COAD. En el análisis de integración de multi-ómicas también se observó que el modelo de 3 subtipos era muy bueno si solo se tomaba en cuenta los CNA. En ambos casos se ha demostrado la importancia de integrar un mayor número de datos, al igual que espero haberlo podido demostrar con este TFM.

7 Conclusiones

7.1 Conclusiones

Se ha constatado que es esencial la integración de multi-ómicas para llegar a los mejores resultados posibles desde para realizar subtipado hasta para determinar posibles biomarcadores o medicamentos que generen la menor cantidad de efectos secundarios posibles realizando la mejor labor terapéutica. Este TFM solo dio algunas pinceladas sobre la amplia y compleja disciplina de la integración de datos, pero por mi parte han sido suficientes como para querer continuar profundizando en el ámbito de la biología de sistemas. El objetivo inicial de este TFM era muy distinto del TFM al que se ha llegado, sin embargo me encuentro satisfecha con el trabajo y esfuerzo realizado.

7.2 Líneas de futuro

El objetivo inicial de este trabajo era determinar posibles biomarcadores de COAD, por lo que en un futuro se realizará, como indican Mo et al[8] y Song et al[11], un análisis de supervivencia estudiando las curvas Kaplan Meier para evaluar la supervivencia en cada subtipo obtenido; el test log Rank para testar la H0: no hay diferencia entre las distintas poblaciones de estudio, en ningún momento en el tiempo, en cuanto a la probabilidad de que ocurra un determinado evento final, en este caso la muerte; y por último, una regresión de COX para evaluar cómo afectan las diferentes variables clínicas junto con el subtipo a la supervivencia.

Finalmente, si se comprueba el valor prognóstico se entrenaría un modelo de predicción mediante Supported Vector Machine(SVM) con los distintos driver genes obtenidos.

7.3 Seguimiento de la planificación

Ha sido imposible seguir las planificaciones, y se han tenido que ir readaptando continuamente porque se han cumplido, múltiples veces, todos los riesgos que comenté al inicio. Se dedicó tiempo a la compra y preparación de un equipo y servidor para realizar el TFM, se realizaron protocolos inviables porque alguna de las funciones finales necesitaba un paquete descatalogado en la última versión de R, o bien, ya no funcionaban por algún error interno. Entre otras muchas incidencias que se superaron con éxito.

8 Glosario

- **Análisis Single cell**: consiste en analizar una determinada ómica de forma independiente en cada célula de una muestra, y no del material obtenido de la digestión de toda la muestra cómo se realiza tradicionalmente [2,224] Este análisis, que está incrementando cada vez más su relevancia, tiene el objetivo de poder diferenciar las líneas celulares presentes en dicho tejido [2,225]. Para realizarlo se han combinado las técnicas de análisis habituales con técnicas de aislamiento celular mediante fluorescencia [2,226].

- Variación del número de copias (CNV): es la variación (por duplicación, deleción, traslocación o inserción) en el número de copias de un gen o región cromosómica en células germinales, por tanto, son variaciones hereditarias. Este es el tipo de variaciones estudiadas en las enfermedades genéticas hereditarias [14,227,228].

- Alteración en el número de copias (CNA): es la variación en el número de copias de un gen o región cromosómica en células somáticas, por tanto, no son variaciones hereditarias. Este es el tipo de variaciones estudiadas en los tumores. A veces se usa CNV y CNA indistintamente, por lo que hay que estar atentos al tema de estudio para tener claro a qué concepto se refiere [227,228].

- MATLAB es una plataforma desarrollada para programar, realizar cálculo numérico, analizar y visualizar datos, y crear modelos. Tiene su propio lenguaje MATLAB y permite trabajar con los lenguajes C/C++, java, python o fortran. [229]

- Incidencia: La incidencia de una patología es el número de nuevos casos de la patología que aparecen en una población específica en un determinado periodo de tiempo, normalmente un año). Se puede expresar en valores absolutos sobre el total de la población por año, o bien, en número de casos por cada 100.000 personas por año [230].

- Mortalidad: La mortalidad en una patología es el número de muertes debidas a dicha patología en una determinada población dentro de un periodo de tiempo específico, normalmente un año. Se puede expresar en valores absolutos sobre el total de la población por año, o bien, en número de casos por cada 100.000 personas por año [230].

- Tasa Bruta: Se calcula dividiendo el número de casos nuevos o muertes observadas dentro de un periodo de tiempo específico entre el número de individuos en la población de riesgo. Se suele expresar en una tasa anual por cada 100.000 personas al año [230]

9 Bibliografía

1.Shen H, Laird PW. Interplay between the cancer genome and epigenome. Cell. 2013 Mar 28;153(1):38-55. doi: 10.1016/j.cell.2013.03.008. PMID: 23540689; PMCID: PMC3648790.

2.de Anda-Jáuregui G, Hernández-Lemus E. Computational Oncology in the Multi-Omics Era: State of the Art. Front Oncol. 2020 Apr 7;10:423. doi: 10.3389/fonc.2020.00423. PMID: 32318338; PMCID: PMC7154096.

3.Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Front Oncol. 2020 Jun 30;10:1030. doi: 10.3389/fonc.2020.01030. PMID: 32695678; PMCID: PMC7338582.

4. Broad GDAC Firehose. [internet] Gdac.broadinstitute.org. 2021 [cited 4 April 2021] Available from: https://gdac.broadinstitute.org/

5. Fundación Instituto Roche. Informe Anticipando Ciencias Ómicas [Internet]. 2019. Available from: https://www.institutoroche.es/static/archivos/Informes_anticipando_CIENCIAS_OMICAS.pdf

6. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, Ritchie ME. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res. 2016 Jun 17;5:ISCB Comm J-1408. doi: 10.12688/f1000research.9005.3. PMID: 27441086; PMCID: PMC4937821.

7. limma [internet] Bioconductor. 2021[cited 5 June 2021] Available from: https://bioconductor.org/packages/release/bioc/html/limma.html

8. Mo Q, Li R, Adeegbe DO, Peng G, Chan KS. Integrative multi-omics analysis of muscle-invasive bladder cancer identifies prognostic biomarkers for frontline chemotherapy and immunotherapy. Commun Biol. 2020 Dic 17; 3 (784) . https://doi.org/10.1038/s42003-020-01491-2

9. MethylMix [internet] Bioconductor. 2021[cited 5 June 2021] Available from: https://www.bioconductor.org/packages/release/bioc/vignettes/MethylMix/inst/doc/vignettes.html

10. mixOmics[internet] mixomics.org. 2021[cited 5 June 2021] Available from: http://mixomics.org/methods/rcca/

11. Song Y, Yang K, Sun T, Tang R. Development and validation of prognostic markers in sarcomas base on a multi-omics analysis. BMC Med Genomics. 2021 Jan 28;14(1):31. doi: 10.1186/s12920-021-00876-4. PMID: 33509178; PMCID: PMC7841904.

12. GanttProject: free project management tool for Windows, macOS and Linux [Internet]. GanttProject. 2021 [cited 16 March 2021]. Available from: https://www.ganttproject.biz/

13. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018;19(5):299-310. doi:10.1038/nrg.2018.4

14. Glosario. [internet] Fundación Instituto Roche: Glosario de genética.2021 [cited 23 April 2021]. Available from: https://www.institutoroche.es/recursos/glosario/

15. Bernal L. La era de las ciencias ómicas. Academia de Farmacia “Reino de Aragón.” 2015; 64. doi:10.1016/j.ehb.2011.08.004

16. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719-24. doi: 10.1038/nature07943.

17. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017 May 18;13(5):e1005457. doi: 10.1371/journal.pcbi.1005457. PMID: 28545146; PMCID: PMC5436640

18. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(83):1-15. doi:10.1186/s13059-017-1215-1

29. Bell O, Tiwari VK, Thomä NH, Schübeler D. Determinants and dynamics of genome accessibility. Nat Rev Genet. (2011) 12:554. doi: 10.1038/nrg3017

20.Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. (2019) 20:207–20. doi: 10.1038/s41576-018-0089-8

21 .Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. (2015) 109:21–9. doi: 10.1002/0471142727.mb2129s109

22. Jia R, Chai P, Zhang H, Fan X. Novel insights into chromosomal conformations in cancer. Mol Cancer. (2017) 16:173. doi: 10.1186/s12943-017-0741-5

23. Grunau C, Renault E, Rosenthal A, Roizes G. MethDB--a public database for DNA methylation data. Nucleic Acids Res. 2001 Jan 1;29(1):270-4. doi: 10.1093/nar/29.1.270. PMID: 11125109; PMCID: PMC29842.

24.Van Eyk JE, Snyder MP. Precision Medicine: Role of Proteomics in Changing Clinical Management and Care. J Proteome Res. 2019;18(1):1-6. doi:10.1021/acs. jproteome.8b00504

25.Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014 Jun;46(6):543-550. doi: 10.1038/ng.2982. Epub 2014 May 11. PMID: 24816252; PMCID: PMC4064254.

26.Raghu P. Functional diversity in a lipidome. Proc Natl Acad Sci USA. 2020 May 26;117(21):11191-11193. doi: 10.1073/pnas.2004764117. Epub 2020 May 12. PMID: 32398365; PMCID: PMC7260965.

27.Perrotti F, Rosa C, Cicalini I, Sacchetta P, Del Boccio P, Genovesi D, Pieragostino D. Advances in Lipidomics for Cancer Biomarkers Discovery. Int J Mol Sci. 2016 Nov 28;17(12):1992. doi: 10.3390/ijms17121992. PMID: 27916803; PMCID: PMC5187792.

28.Feng S, Zhou L, Huang C, Xie K, Nice EC. Interactomics: toward protein function and regulation. Expert Rev Proteomics. 2015 Feb;12(1):37-60. doi: 10.1586/14789450.2015.1000870. Epub 2015 Jan 12. PMID: 25578092.

29.Song P, Kwon Y, Joo JY, Kim DG, Yoon JH. Secretomics to Discover Regulators in Diseases. Int J Mol Sci. 2019 Aug 9;20(16):3893. doi: 10.3390/ijms20163893. PMID: 31405033; PMCID: PMC6720857.

30. Gomase VS, Tagore S. Cytomics. Curr Drug Metab. 2008 Mar;9(3):263-6. doi: 10.2174/138920008783884731. PMID: 18336233.

31.Reactome. Pathways and Networks Overview [Internet]. 2020. Available from:https://reactome.org/docs/training/Pathways_&_Networks_Overview.pdf

32.Glosario. [internet] .2021 [cited 23 April 2021]. Available from: https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/introduction-to-graph-theory/graph-theory-graph-types-and-edge-properties/

33.Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, Merico D, Bader GD. Pathway enrichment

analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019 Feb;14(2):482-517. doi: 10.1038/s41596-018-0103-9. PMID: 30664679; PMCID: PMC6607905.

34. Cook-Deegan R, McGuire AL. Moving beyond Bermuda: sharing data to build a medical information commons. Genome Res. 2017 Jun;27(6):897-901. doi: 10.1101/gr.216911.116. Epub 2017 Apr 3. PMID: 28373484; PMCID: PMC5453323.

35. Jansen P, van den Berg L, van Overveld P, Boiten JW. Research Data Stewardship for Healthcare Professionals. 2018 Dec 22. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science [Internet]. Cham (CH): Springer; 2019. Chapter 4. PMID: 31314246.

36. *NIH STRATEGIC PLAN FOR DATA SCIENCE*. [internet] NIH, 2018[cited 4 April 2021]. Available from: https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

37. *Cloud Credits*. [internet] Commonfund.nih.gov. 2021.[cited 4 April 2021] Available from: https://commonfund.nih.gov/bd2k/cloudcredits .

38.Deutsch EW, Bandeira N, Sharma V, Perez-Riverol Y, Carver JJ, Kundu DJ, et al. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. Nucleic Acids Res. 2020 Jan 8;48(D1):D1145-D1152. doi: 10.1093/nar/gkz984. PMID: 31686107; PMCID: PMC7145525.

39.Papatheodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells.. Nucleic acids Res. 2019 Oct 30; 48 (D1): D77-D83. https://doi.org/10.1093/nar/gkz947

40. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, Madhavan S, Ketchum KA. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. J Proteome Res. 2015 Jun 14;14(6) 2707-2713. doi:10.1021/pr501254j. PMID: 25873244.

41.Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012 May;2(5):401-4. doi: 10.1158/2159-8290.CD-12-0095. Erratum in: Cancer Discov. 2012 Oct;2(10):960. PMID: 22588877; PMCID: PMC3956037.

42.*GDC* [internet] Portal.gdc.cancer.gov. 2021 [cited 4 April 2021] Available from: https://portal.gdc.cancer.gov/.

43. Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. 2018 Jan 4;46(D1):D956-D963. doi: 10.1093/nar/gkx1090. PMID: 29136207; PMCID: PMC5753188.

44.Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014 Jan;42(Database issue):D975-9. doi: 10.1093/nar/gkt1211. Epub 2013 Dec 1. PMID: 24297256; PMCID: PMC3965052

45.Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5. doi: 10.1093/nar/gks1193. Epub 2012 Nov 27. PMID: 23193258; PMCID: PMC3531084

46.Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A et al. The BioStudies database-one stop shop for all data supporting a life sciences study. Nucleic Acids Res. 2018 Jan 4;46(D1):D1266-D1270. doi: 10.1093/nar/gkx965. PMID: 29069414; PMCID: PMC5753238.

47.Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27-30. doi: 10.1093/nar/28.1.27. PMID: 10592173; PMCID: PMC102409.

48.Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020 Jan 8;48(D1): D498-D503. doi: 10.1093/nar/gkz1031. PMID: 31691815; PMCID: PMC7145712.

49.Pillich RT, Chen J, Rynkov V, Welker D, Pratt D. NDEx: A Community Resource for Sharing and Publishing of Biological Networks. Methods Mol Biol. 2017;1558:271-301. doi: 10.1007/978-1-4939-6783-4_13. PMID: 28150243.

50.Database Commons[internet] CNCB-NGDC. 2021 [cited 4 april 2021]. Available from: https://bigd.big.ac.cn/databasecommons/

51. Capella-Gutierrez S, De la Iglesia D, Haas J, Lourenco A, Fernández, Repchevsky D, Dessimoz C et al. Lessons Learned: Recommendations for Establishing Critical Periodic Scientific Benchmarking. BioRxiv. 2017 Aug 31 ; doi: https://doi.org/10.1101/181677

52.Ison J, Rapacki K, Ménager H, Kalaš M, Rydza E, Chmura P et al. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res. 2016 Jan 4;44(D1):D38-47. doi: 10.1093/nar/gkv1116. Epub 2015 Nov 3. PMID: 26538599; PMCID: PMC4702812.

53.Procedia Manufacturing: Machine Learning Approaches to Learning Heuristics for Combinatorial Optimization Problems. Elsevier [Internet]. 2021 [cited 11 April 2021];(17):102-109. Available from: https://www.sciencedirect.com/science/article/pii/S2351978918311351

54.Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020 Jan 31;14:1177932219899051. doi: 10.1177/1177932219899051. PMID: 32076369; PMCID: PMC7003173.

55. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018 Mar 15;24(6):1248-1259. doi: 10.1158/1078-0432.CCR-17-0853. Epub 2017 Oct 5. PMID: 28982688; PMCID: PMC6050171.

56.Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016 Jul 29;12(7):878. doi: 10.15252/msb.20156651. PMID: 27474269; PMCID: PMC4965871.

57.Kim S, Jhong JH, Lee J, Koo JY. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017 Jan 26;10:2. doi: 10.1186/s13040-017-0126-8. Erratum in: BioData Min. 2017 Feb 14;10 :8. PMID: 28149325; PMCID: PMC5270233.

58.James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R. 1st ed. London: Springer; 2013.

59.Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics. 2018 Jan 1;19(1):71-86. doi: 10.1093/biostatistics/kxx017. PMID: 28541380; PMCID: PMC6455926.

60.Rohart F, Gautier B, Singh A, Lê Cao KA. MixOmics: An R package for 'omics feature selection and multiple data integration'. PLOS Computational Biology. 2017 Nov 3; 13(11): e1005752. https://doi.org/10.1371/journal.pcbi.1005752

61.Vlachavas EI, Bohn J, Ückert F, Nürnberg S. A Detailed Catalogue of Multi-Omics Methodologies for Identification of Putative Biomarkers and Causal Molecular Networks in Translational Cancer Research. Int J Mol Sci. 2021 Mar 10;22(6):2822. doi: 10.3390/ijms22062822. PMID: 33802234; PMCID: PMC8000236.

62.Lantz B. Machine Learning with R: Expert Techniques for Predictive Modeling to Solve All Your Data Analysis Problems. Birmingham: Packt Publishing. 2nd. de; 2015.
63 iClusterPlus: Integrative clustering of multi-type genomic data. [Internet]. Bioconductor. 2018[cited 5 june 2021]. Available from: https://www.bioconductor.org/packages/devel/bioc/vignettes/iClusterPlus/inst/doc/ iManual.pdf

64.Similarity NEtwork FUSION (SNF)[internet]. Compbio.cs.toronto.edu 2021 [cited 16 April 2021]. Available from: http://compbio.cs.toronto.edu/SNF/SNF/Software.html

65.Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014 Mar;11(3):333-7. doi: 10.1038/nmeth.2810. Epub 2014 Jan 26. PMID: 24464287.

66.Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J et al. Inference of patient-specific pathway activities from multi-dimensional càncer genomics data using PARADIGM. Bioinformatics. 2010 Jun 15;26(12):i237-45. doi: 10.1093/bioinformatics/btq182. PMID: 20529912; PMCID: PMC2881367.

67.Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC et al. Multi-Omics Factor Analysis-a framework for unsupervised Integration of multi-omics data sets. Mol Syst Biol. 2018 Jun 20;14(6):e8124. doi: 10.15252/msb.20178124. PMID: 29925568; PMCID: PMC6010767.

68.Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. MOFA+: a statistical framework for comprehensive integration of multi-modal singlecell data. Genome Biol. 2020 May 11;21(1):111. doi: 10.1186/s13059-020-02015-1. PMID: 32393329; PMCID: PMC7212577.

69.Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019 Jul 15;35(14): i501-i509. doi: 10.1093/bioinformatics/btz318. PMID: 31510700; PMCID: PMC6612815.

70.Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multiomic data. Bioinformatics. 2019 Sep 15;35(18):3348-3356. doi: 10.1093/bioinformatics/btz058. PMID: 30698637; PMCID: PMC6748715.

71. Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: interpretable patient classification using integrated patient similarity networks. Mol Syst Biol. 2019 Mar 14;15(3):e8497. doi: 10.15252/msb.20188497. PMID: 30872331; PMCID: PMC6423721.

72. Shinde J, Everaert C, Bakr S, Nabian M, Xu J, Carey V, Pochet N, Gevaert O. AMARETTO: Regulatory Network Inference and Driver Gene Evaluation using Integrative Multi-Omics Analysis and Penalized Regression. Bioconductor R. 2020.

73.Champion M, Brennan K, Croonenborghs T, Gentles AJ, Pochet N, Gevaert O. Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Response. EBioMedicine. 2018 Jan;27:156-166. doi: 10.1016/j.ebiom.2017.11.028. Epub 2017 Dec 1. PMID: 29331675; PMCID: PMC5828545.

74.Huang L, Brunell D, Stephan C, Mancuso J, Yu X, He B et al. Driver network as a biomarker: systematic Integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. Bioinformatics. 2019 Oct 1;35(19):3709-3717. doi: 10.1093/bioinformatics/btz109. PMID: 30768150; PMCID: PMC6761967.

75.Galaxeast [Internet]. Galaxeast.fr 2021 [cited 20 April 2021]. Available from: http://www.galaxeast.fr

76.Repositories in Category Systems Biology [Internet]. Galaxy Tool Sheet 2021

[cited 20 April 2021]. Available from: https://toolshed.g2.bx.psu.edu/

77.OpenOmics[Internet]. BioMeCIS Lab. 2021 [cited 20 April 2021]. Available from: https://github.com/BioMeCIS-Lab/OpenOmics

78.T-Bioinfo Platform [Internet]. Server.t-bio.info. 2021 [cited 20 April 2021]. Available from: https://server.t-bio.info/

79.OneOmics [Internet]. Sciex. 2021 [cited 21 April 2021]. Available from: https://sciex.com/applications/life-science-research/oneomics

80.QIAGEN Ingenuity Pathway Analysis (IPA) [Internet]. QIAGEN . 2021 [cited 21 April 2021]. Available from: https://www.qiagen.com/us/products/discovery-andtranslational-research/next-generation-sequencing/informatics-and-data/interpretationcontent-databases/ingenuity-pathway-analysis/

81.Newton Y, Novak AM, Swatloski T, McColl DC, Chopra S, Graim K et al. TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. Cancer Res. 2017 Nov 1;77(21):e111-e114. doi: 10.1158/0008- 5472.CAN-17-0580. PMID: 29092953; PMCID: PMC5751940.

82.FireBrowse [Internet]. Broad Institute. 2021 [cited 21 April 2021]. Available from: http://firebrowse.org/

83.LinkedOmics [Internet]. LinkedOmics.org . 2021 [cited 21 April 2021]. Available from: http://linkedomics.org/login.php

84.cBioPortal [Internet]. cBioPortal.org . 2021 [cited 21 April 2021]. Available from: http://www.cbioportal.org/

85.Terra [Internet]. Terra.bio. 2021 [cited 22 April 2021]. Available from: https://terra.bio/

86.de.NBICloud [Internet]. de.NBI. 2021 [cited 22 April 2021]. Available from: https://www.denbi.de/cloud

87.Requisitos Técnicos [Internet]. Universidad de Navarra. 2021 [cited 22 April 2021]. https://www.unav.edu/web/master-en-big-data-science/plan-de-estudios/requisitos-tecnicos

88.Hardware Requirements [Internet]. Bioinformatics.stackexchange . 2021 [cited 22 April 2021]. Available from: https://bioinformatics.stackexchange.com/questions/6936/hardware-requirements-specs-for-bioinformatics-dedicated-desktop/6939

89.Cloud Life Science [Internet]. Cloud Google. 2021 [cited 22 April 2021]. Available from:https://cloud.google.com/life-sciences

90.Amazon Web Services(AWS) [Internet]. AWS Amazon. 2021 [cited 22 April 2021]. Available from: https://aws.amazon.com/es/health/genomics/

91.Cancer Genome Collaboratory [Internet]. Cancercolaboratory.org. 2021 [cited 22 April 2021]. Available from:https://cancercollaboratory.org/

92.¿Qué es el cáncer colorrectal? [Internet]. Cancer.org. 2021 [cited 20 March 2021]. Available from: https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/acerca/que-es-cancer-de-colon-o-recto.html

93. Moran BJ, Jackson AA. Function of the human colon. Br J Surg. 1992 Nov;79(11):1132-7. doi: 10.1002/bjs.1800791106. PMID: 1467882.

94-Cancer Today [Internet]. Global Cancer Observatory. 2021 [cited 20 March 2021]. Available from: https://gco.iarc.fr/

95-Tratamiento del cáncer de colon, versión para profesionales [Internet]. Cancer.org. 2021 [cited 20 March 2021]. Available from: https://www.cancer.gov/espanol/tipos/colorrectal/pro/tratamiento-colorrectal-pdq

96. COAD Sample Report [Internet]. GDAC Broad Institute. 2021 [cited 20 March 2021].Available from: http://gdac.broadinstitute.org/runs/stddata__latest/samples_report/COAD.html

97.COAD Sample Report [Internet]. GDAC Broad Institute. 2021 [cited 20 March 2021].Available from: http://gdac.broadinstitute.org/runs/stddata__latest/samples_report/COAD.html

98. Ramos M. curatedTCGAData: Curated Data From The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects. R package version 1.12.1. Bioconductor R; 2021

99.Yang C, Zhang Y, Xu X, Li W. Molecular subtypes based on DNA methylation predict prognosis in colon adenocarcinoma patients. Aging (Albany NY). 2019 Dec 18;11(24):11880-11892. doi: 10.18632/aging.102492. Epub 2019 Dec 18. PMID: 31852837; PMCID: PMC6949097.

100.Liao SG, Lin Y, Kang DD, Chandra D, Bon J, Kaminski N, Sciurba FC, Tseng GC. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC Bioinformatics. 2014 Nov 5;15(1):346. doi: 10.1186/s12859-014-0346-6. PMID: 25371041; PMCID: PMC4228077.

101.Baneshi MR, Talei AR. Does the missing data imputation method affect the composition and performance of prognostic models? Iran Red Crescent Med J. 2012 Jan;14(1):31-6. Epub 2012 Jan 1. PMID: 22737551; PMCID: PMC3372019.

102.Ramos M, Schiffer L, Waldron L. TCGAutils: TCGA utility functions for data management. R package version 1.10.1. Bioconductor R; 2021.

103. Wickham H. Simple, Consistent Wrappers for Common String Operations. R package versión 1.4.0. R-cran. 2021

104.Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Cabrera CR et al. Software For The Integration Of Multi-Omics Experiments In Bioconductor. *Cancer Research*, **77(21);** e39-42. 2017.

105. match[internet] R-documentation. 2021[cited 5 june 2021]. Available from: https://www.google.com/search?q=match+R&oq=match+r&aqs=chrome.0.69i59j0l6j69i60.6014j0j7&sourceid=chrome&ie=UTF-8

106. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140. 2021. doi: 10.1093/bioinformatics/btp616.

107.Reference Manual MethylMix[internet] Bioconductor. 2021[cited 5 June 2021] Available from: https://www.bioconductor.org/packages/release/bioc/manuals/MethylMix/man/MethylMix.pdf

108.Gavaertlab/MethylMix[internet] Github. 2021[cited 5 June 2021] Available from: https://github.com/gevaertlab/MethylMix/blob/master/R/MethylMix.R

109.Gevaert 0. MethylMix: an R package for identifying DNA Methylation driven genes. Bioinformatics (Oxford, England). 2015;31(11):1839- 41. doi:10.1093/bioinformatics/btv020.

110.Gevaert O, Tibshirani R, Plevritis SK. Pancancer analysis of DNA Methylation driven genes using MethylMix. Genome Biology. 2015;16(1):17. doi:10.1186/s13059-014-0579- 8

111. Ortiz MT. Estadística Computacional: Análisis Bayesiano[internet] Github. 2021[cited 5 June 2021. Available from: https://tereom.github.io/est-computacional-2018/analisis-bayesiano.html

112.Kruschke J. Doing Bayesian Data Analysis. Boston: Academic Press. 2nd edition. 2015.

113. Trevor H, Tibshirani R, Friedman J. The Elements of Statistical Learning. New YorK: Springer New York Inc. Springer Series in Statistics. 2001.

114. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. London: Chapman & Hall/CRC. 1st edition. 1996.

115. Logistic Regression[internet] Penn State Eberly College of Science. 2021. [cited 5 June 2021. Available from: https://online.stat.psu.edu/stat501/lesson/15/15.1

116. Heatmap [internet] Penn State Eberly College of Science. 2021. [cited 5 June 2021. Available from: https://online.stat.psu.edu/stat555/node/87/

117. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012 May;16(5):284-7. doi: 10.1089/omi.2011.0118. Epub 2012 Mar 28. PMID: 22455463; PMCID: PMC3339379.

118. Bioinformatics & Evolutionary Genomics[internet] Ghent University. 2021[cited 8 June 2021] Available from: http://bioinformatics.psb.ugent.be/webtools/Venn/

119. Pratt et al. NDEx, the Network Data Exchange. Cell Systems, Vol. 1, Issue 4: 302-305 (2015).

120.Pillich et al. NDEx: A Community Resource for Sharing and Publishing of Biological Networks. Methods Mol Biol, 1558: 271-301 (2017).

221.Pratt et al. NDEx 2.0: A Clearinghouse for Research on Cancer Pathways. Cancer Res. Nov 1;77(21):e58-e61 (2017).

222. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa E Melo F, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam GC, Broom B, Boige V, Perez-Villamil B, Laderas T, Salazar R, Gray JW, Hanahan D, Tabernero J, Bernards R, Friend SH, Laurent-Puig P, Medema JP, Sadanandam A, Wessels L, Delorenzi M, Kopetz S, Vermeulen L, Tejpar S. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015 Nov;21(11):1350-6. doi: 10.1038/nm.3967. Epub 2015 Oct 12. PMID: 26457759; PMCID: PMC4636487.

223. Xie T, D' Ario G, Lamb JR, Martin E, Wang K, Tejpar S, Delorenzi M, Bosman FT, Roth AD, Yan P, Bougel S, Di Narzo AF, Popovici V, Budinská E, Mao M, Weinrich SL, Rejto PA, Hodgson JG. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. PLoS One. 2012;7(7):e42001. doi: 10.1371/journal.pone.0042001. Epub 2012 Jul 31. PMID: 22860045; PMCID: PMC3409212.

224.Yuan Y. Spatial heterogeneity in the tumor microenvironment. Cold Spring Harb Perspect Med. 2016; 6:a026583. doi: 10.1101/cshperspect.a026583

225. Ren X, Kang B, Zhang Z. Understanding tumor ecosystems by single- cell sequencing: promises and limitations. Genome Biol. 2018; 19:1-14.doi: 10.1186/s13059-018-1593-z

226.Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. (2015) 16:133-45. doi: 10.1038/nrg3833

227.Griffiths AJF, Miller JH, Suzuki DT, et al. An Introduction to Genetic Analysis. 7th ed. New York: W. H. Freeman; 2000. Somatic versus germinal mutation. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21894/

228.Li W, Lee A, Gregersen PK. Copy-number-variation and copy-number-alteration region detection by cumulative plots. BMC Bioinformatics. 2009 Jan 30;10 Suppl 1(Suppl 1):S67. doi: 10.1186/1471-2105-10-S1-S67. PMID: 19208171; PMCID: PMC2648736.

229. MATLAB [Internet]. Mathworks. 2021 [cited 16 April March 2021]. Available from:https://es.mathworks.com/products/matlab.html

230. 1. About [Internet]. Global Cancer Observatory. 2021 [cited 20 March 2021]. Available from:https://gco.iarc.fr/today/about

Abreviaturas

- Alteración en el número de copias(CNA)
- Artificial Neural Networks(ANN)
- Componentes Principales CP.
- Database of Genotypes and Phenotypes (dbGap)
- Differential metilation values(DM-Values)
- Memorial Sloan Kettering Cancer Center (MSK)
- MicroRNA (miRNA)
- Mínimos cuadrados parciales (PLS)
- Multi-omics factor analysis (MOFA)
- Multi-Omics Late Integration (MOLY)
- National Cancer Institute(NCI)
- National Center of Biothecnology Information (NCBI)
- National Institute of Health(NIH)
- NEighborhood based Multi-Omics clustering (NEMO)
- Polimorfismos de un único núcleótido (SNP)
- Protein-Protein Interaction Networks (PPINs)
- Proteomics Tumor Analysis Consortium (CPTAC)
- RNA no codificante (ncRNA)
- RNA mensajeros (mRNA)
- Regresión de componentes principales (PCR)
- Support Vector Machine Recursive Feature Elimination(SVM-RFE)
- The Cancer Genome Atlas (TCGA)
- Variación del número de copias (CNV)
- Pathway Recognition Algorithm using Data Integration on Genomics Models(PARADIGM)
- QIAGEN Ingenuity Pathway Analysis (IPA)
- Secuenciación de última generación (NGS o HTS)
- Similarity Network Fusion (SNF)

Anexos

Anexo 1: Tipos de repositorios según su política de uso de los datos.

En algunos repositorios, como ENCODE, los usuarios pueden acceder y usar cualquiera de los sets de datos libremente, solo es necesario citar la fuente de los mismos [1,2].

En otros, como CPTAC Data Portal o dbGap, además de citar, es necesario fijarse que algunos sets de datos tienen un Periodo de Embargo, es decir, se pueden utilizar dichos datos pero no se puede realizar ninguna publicación con los mismos hasta que transcurra la fecha indicada en el Periodo Embargo. Esto permite a los investigadores que generaron los datos tener la exclusiva de la primera publicación [3,4].

En otros, como dbGap, los usuarios deben solicitar un permiso de acceso a los datos que desean utilizar [5,6] La solicitud de acceso solo la puede realizar el miembro responsable del proyecto, el cual debe ser empleado fijo de la institución solicitante y científico senior. En la solicitud se debe indicar el proyecto de estudio a realizar; objetivos; diseño del estudio y plan de análisis; explicación de cómo se ajustarán a las limitaciones en los permisos de uso de los datos; si se va a usar almacenamiento o una máquina virtual en la nube para manejar los datos, y en su caso el proveedor del servicio; si va a haber o se ha planteado colaboración con otra u otras instituciones; y en el caso de que la institución solicitante tenga poco tiempo de existencia, o sea pequeña, se solicitará también currículum vitae de los miembros que manejan los datos, información sobre dicha institución y un plan de seguridad[3] para proteger los datos a los que se va acceder [5]. Los estudiantes no pueden solicitar acceso a estos datos, por lo que descartamos este tipo de repositorios para la búsqueda de los datos que vamos a usar en el TFM [6].

Bibliografía:

1.ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247. PMID: 22955616; PMCID: PMC3439153.

2.Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018 Jan 4;46(D1):D794-D801. doi: 10.1093/nar/gkx1081. PMID: 29126249; PMCID: PMC5753278.

3.Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, Madhavan S, Ketchum KA. The CPTAC Data Portal: A Resource for Cancer

Proteomics Research. J Proteome Res. 2015 Jun 14;14(6) 2707-2713. doi:10.1021/pr501254j. PMID: 25873244.

4.Osr.ucsf.edu. 2021. Embargo dates for dbGaP datasets | UCSF Office of Sponsored Research. [online] Available at: <https://osr.ucsf.edu/content/embargo-dates-dbgap-datasets> [Accessed 4 April 2021].

5.Jansen P, van den Berg L, van Overveld P, Boiten JW. Research Data Stewardship for Healthcare Professionals. 2018 Dec 22. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science [Internet]. Cham (CH): Springer; 2019. Chapter 4. PMID: 31314246.

6.NCBI dbGap, 2015. Tips for PreparingaSuccessful Data Access Request. [online] Available at: <https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi? document_name=GeneralAAInstructions.pdf> [Accessed 4 April 2021].

Anexo 2: Resumen de las herramientas de integración

Herramienta	iClusterPlus	SNF	MixOmics	PARADIGM
Lenguaje	R	R o MATLAB	R	Python
Clase de método en que se basa	Transformación de variables	Redes	Transformación de variables	Redes
Clase de Método según tipo de aprendizaje	Unsupervised	Unsupervised	Unsupervised y Supervised	Unsupervised
Maneja Missing Values	No	No	Si	NA
Datos que integra	Genómica, transcriptómica y metilómica	Datos clínicos, imagen, transcriptómica y metilómica	Transcriptómica, metabolómica, proteómica, microbiómica, metagenómica e imagen.	Genómica, Transcriptómica y epigenómica.
Información que obtiene	Subtipado y Biomarcadores	Subtipado	Subtipado y Biomarcadores	Pathways y prognosis con pathways alterados

Herramienta	MOFA	MOLY	NEMO	NetDX
Lenguaje	R o Python	Python	R	R
Clase de método en que se basa	Factorización y Transformación de variables	Deep Network	Clustering	Extracción de variables
Clase de Método según tipo de aprendizaje	Unsupervised	Supervised	Unsupervised	Supervised
Acepta Missing Values	Si	NA	Si	No

Datos que integra	Genética, transcriptómica y epigenómica	Genómica, transcriptómica y proteómica.	Transcriptómi cay metilómica.	Genómica, transcriptómica, metilómica y proteómica.
Información que obtiene	Subtipado y Biomarcadores	Predecir respuesta a un medicamento.	Subtipado	Subtipado

Herramienta	**Amaretto**	**DrugCombo-Explorer**
Lenguaje	R	Java y Python
Clase de método en que se basa	Redes	Redes
Clase de Método según tipo de aprendizaje	Unsupervised	Supervised
Acepta Missing Values	No	NA
Datos que integra	Genómica, transcriptómica, y metilómica	Genómica, transcriptómica, y metilómica
Información que obtiene	Subtipado, Biomarcadores y Pathways	Combinaciones de medicamentos

Anexo 3: Código TCGA

El código que define los datos de cada paciente se denomina código de referencia o de barras de TCGA, y tiene la siguiente estructura:[1]

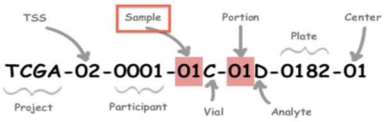


Figura 1: Estructura y ejemplo del código de referencia.[1]

- Proyecto: Código del proyecto al que pertenece el caso. El proyecto del que se extrajeron nuestros datos es TCGA.
- TSS: Código correspondiente al centro de obtención de la muestra, a la enfermedad y al tejido del que se ha obtenido. En el adenocarcinoma de colon hay 29 posibles TSS.
- Participante: Código del paciente.
- Sample: código del tipo de muestra. En el adenocarcinoma de colon se podrá encontrar: [1,2]
    - Tumor primario sólido, TP: 01
    - Tumor sólido recurrente, TR: 02
    - Metastásico: 06
    - Células sanguíneas normales, NB: 10
    - Tejido normal sólido, NT: 112
- Vial de la muestra.
- Portion: número correspondiente a la alícuota de la que se obtuvo el analito
- Analyte: código que representa el tipo de analito según el protocolo de extracción.
- Plate: código que se le otorga a cada caso y analito.
- Center: centro que obtuvo la muestra y los datos [1].

Bibliografía:

1. Frequently Asked Questions [Internet]. Broad Institute. 2021 [cited 20 March 2021].Available from: https://broadinstitute.atlassian.net/wiki/spaces/GDAC/pages/844334036/FAQ#FAQ-ClinicalQ%3AYourdocumentationdoesnotdescribewhattheclinicaltermX.Y.Zmeans%2CwherecanIfindthis%3F

2.Sample type codes [Internet]. Gdc.cancer.gov . 2021 [cited 20 March 2021].Available from: https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes

Anexo 4: Filtrado previo de los datos TCGA compartidos en Firehose.

Los métodos de filtrado fueron [1.2]:

- Redacción: Consistió en eliminar casos antes de ser publicados en la base de datos. Dichos datos se eliminaron por no tener el código de identidad del paciente; discordancia genotípica con los datos clínicos (ej: se anotó un género en los datos clínicos y genotípicamente se observó otro); tipo de cáncer, tejido u órgano erróneo [3].

- Selección de una réplica por muestra: habitualmente, para evitar bias, se realiza varias veces el mismo análisis mediante la misma técnica o distinta, con la misma alícuota y/o con varias alícuotas de un mismo paciente y del mismo tejido en la misma condición. Cada repetición se conoce como réplica o replicante. Sin embargo, Firehose solo acepta un replicante por técnica y paciente en determinada condición, lo que hizo necesario que se seleccionara la mejor. Para elegirla se aplicaron los criterios siguientes:
  - Filtro de replicantes de analito: Si las moléculas de estudio, denominadas analito, se obtuvieron y analizaron mediante distintas técnicas en cada replicante, se dice que se han obtenido mediante protocolos de extracción o amplificación distintos. En este caso se seleccionaron los datos de la o las réplicas obtenidas mediante el mejor protocolo de extracción o amplificación de los establecidos por TCGA [4]. El orden de prioridad de los protocolos de extracción usado en el filtrado de replicantes de analito es:
    - Cuando el analito a analizar sea RNA el orden de prioridad de los protocolos será H > R > T.
    - Cuando el analito sea DNA el orden de prioridad de los protocolos será D > G, W o X a menos que G, W o X tengan un mayor número “plate” en el código de referencia [4.5].

  - Filtro de ordenación de replicantes: Si tras realizar el filtrado anterior continuaron habiendo varias réplicas, se eligió la que tuviera el mayor número “plate”. [4,6] Específicamente, se elegirán siempre los datos de la alícuota con mayor “portion” y/o número “plate” cuando el resto del código de referencia sea el mismo. Ejemplo:
    - TCGA-37-4130-01A-01D-1097-01
    - TCGA-37-4130-01A-01D-1969-01

    Se elegiría la segunda alícuota pues tiene un mayor “plate”[5].

- Lista negra de muestras: Se añadieron a la lista negra aquellos casos en los que se había elegido una alícuota de peor calidad en el filtrado realizado en la selección de una réplica por muestra. Para realizar este paso se debe hacer un análisis manual de las réplicas. En el caso del

adenocarcinoma de colon no hubo ningún caso que se añadiera a la lista negra [6].

Por lo tanto, en el análisis no se dispondrá de replicantes para analizar los posibles biases, sino que se compensará con el estudio de un elevado número de muestras.

Bibliografía:

1.COAD Sample Report [Internet]. GDAC Broad Institute. 2021 [cited 20 March 2021].Available from: http://gdac.broadinstitute.org/runs/stddata__latest/samples_report/COAD.html

2.FireHose Broad GDAC[Internet]. GDAC Broad Institute. 2021 [cited 20 March 2021]. Available from: http://gdac.broadinstitute.org/#

3.COAD Redactions [Internet]. GDAC Broad Institute. 2021 [cited 20 March 2021].Available from: http://gdac.broadinstitute.org/runs/stddata__latest/samples_report/COAD_Redactions.html

4.COAD Replicate Samples [Internet]. GDAC Broad Institute. 2021 [cited 20 March 2021]Available from: http://gdac.broadinstitute.org/runs/stddata__latest/samples_report/COAD_Replicate_Samples.html

5. Frequently Asked Questions [Internet]. Broad Institute. 2021 [cited 20 March 2021].Available from: https://broadinstitute.atlassian.net/wiki/spaces/GDAC/pages/844334036/FAQ#FAQ-ClinicalQ%3AYourdocumentationdoesnotdescribewhattheclinicaltermX.Y.Zmeans%2CwherecanIfindthis%3F

6.COAD Blacklisted Samples [Internet]. GDAC Broad Institute. 2021 [cited 20 March 2021]Available from: http://gdac.broadinstitute.org/runs/stddata__latest/samples_report/COAD_Blacklisted_Samples.html

Anexo 5: Pre-procesamiento de los datos descargados.

Los datos de TCGA pueden estar en distintos niveles de procesamiento:

- Nivel 1: datos crudos, es decir, los directamente provenientes del array o secuenciador. Estos datos tienen control de acceso. Se obtienen de GDC.
- Nivel 2: datos procesados de acceso controlado, como por ejemplo los datos de Single Nucleotide Polimorphisms (SNPs).[1,2]
- Nivel 3: datos procesados, en ocasiones segmentados según la información que aportan, y de acceso libre. Un ejemplo es que los datos de RNAseq en la web de Firehose se encuentran fragmentados según si tienen los FPKM de mRNAseq por gen, isoforma, o exón.[1,3] En este TFM se usarán los datos de mRNA por gen.
- Nivel 4: contiene los datos de determinada región de interés ya anotados, y son de libre acceso. Un ejemplo de ello son los datos de CNA procesados por GISTIC disponibles en Firehose, que dan los CNA presentes por gen omitiendo los CNA de otras regiones cromosómicas.[1,3,4]

El procesamiento que tienen los datos a usar es:

- Copy Number Aberrations(CNA): Se han obtenido mediante Affymetrix Genome-Wide Human SNP Array 6.0.5 Se anotaron los CNA (duplicaciones y deleciones) y sus intensidades a partir de los datos del array con la herramienta Birdsuit, antes de ello se eliminó el ruido mediante normalización tangencial.

  Con el paquete DNAcopy del proyecto Bioconductor se agruparon los CNA por regiones cromosómicas, y se obtuvo los segment mean values o marker levels que se calculan con log2(copy-number/2) para cada región. Se divide entre 2 para que cuando haya solo dos copias(diploidía) el valor sea de 0, así las regiones amplificadas tienen valores positivos y las delecionadas negativos [6,7].

  Finalmente se anotó los CNA por gen mediante la herramienta GISTIC [6], esta herramienta detecta aquellos genes que tienen más alteraciones de las que ocurrirían por azar de forma que mide la patogenicidad de las mismas.[8]

  El archivo descargado contiene los CNA por gen en G-scores. A partir de los marker levels GISTIC obtiene los scores que miden el nivel de CNAs por gen (gene level). El método que usa para obtener los scores se denomina método extremo. Para ello GISTIC toma como que un mismo gen puede tener más de un CNA cada uno con su correspondiente marker level. El score de cada gen en cada muestra es el marker level de mayor valor del gen en dicha muestra. Con este método se asegura de que el score anotado pertenezca al CNA, entre los presentes en dicho gen, que

tenga mayor amplificación, o bien, el menor valor de deleción. Los valores pueden ir de -1 a 3 [1,8].

- RNA mensajeros obtenidos por RNAseq: La secuenciación se realizó mediante Illumina HiSeq 2000 RNA Sequencing.[5] Los resultados de la secuenciación se alinearon y normalizaron mediante la herramienta RSEM. En esta herramienta el alineamiento se realizó mediante la herramienta STAT, y se obtuvieron los row counts, es decir, fragmentos secuenciados y alineados por gen y muestra. A continuación se transformaron los datos a Counts per Million(CPM) y se normalizaron dividiéndolos por el número de bases de su gen correspondiente y tomando en cuenta que paired-end RNA-seq dos counts pueden pertenecer a un mismo fragmento, la nueva medida del nivel de expresión normalizada se denomina FPKM (Fragments Per Kilobase Million)[6,9,10]

- Metilación de DNA en CpG: los que se va a usar se han obtenido con HumanMethylation450 array[5]. A partir de las intensidades del array se realiza la normalización de las mismas con respecto a los probes control, y posteriormente se calculan los valores Beta=M/(M+U)[6,11]. Donde M es la intensidad medida en el probe metilado y U es la intensidad medida en otro probe con la misma secuencia no metilado, por lo tanto, lo que se está calculando el ratio entre la intensidad del probe metilado y la intensidad total [6,12]. Los B-values tienen un rango de valores que van entre 0 y 1, donde los valores 0 indican que la isla CpG de ese probe no está metilada y los valores 1 que está completamente metilada [13,14].

1. Silva TC, Colaprico A, Olsen C, D'Angelo F, Bontempi G, Ceccarelli M, Noushmehr H. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Res. 2016 Jun 29;5:1542. doi: 10.12688/f1000research.8923.2. PMID: 28232861; PMCID: PMC5302158.

2. GDC [internet] Portal.gdc.cancer.gov. 2021 [cited 4 April 2021] Available from: https://portal.gdc.cancer.gov/.

3. Broad GDAC Firehose. [internet] Gdac.broadinstitute.org. 2021 [cited 4 April 2021] Available from: https://gdac.broadinstitute.org/

4.Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41. Epub 2011 Apr 28. PMID: 21527027; PMCID: PMC3218867.

5.COAD Sample Report [Internet]. GDAC Broad Institute. 2021 [cited 20 March 2021].Available from: http://gdac.broadinstitute.org/runs/stddata__latest/samples_report/COAD.html

6.Data processing GDC [internet] Portal.gdc.cancer.gov. 2021 [cited 4 April 2021] Available from: https://gdc.cancer.gov/about-data/data-processing/genomic-data-processing#Overview

7.Tabak B, Saksena G, Oh C, Gao GF, Hill-Meyers B, Reich M et al. The Tangent copy-number inference pipeline for cancer genome analyses. BioRxiv. 2019. doi: https://doi.org/10.1101/566505

8.Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41. Epub 2011 Apr 28. PMID: 21527027; PMCID: PMC3218867.

9.RNA Seq Data [internet] Penn State Eberly College of Science2021 [cited 4 May 2021]Available from: https://online.stat.psu.edu/stat555/node/13/

10.FAQ [internet] National Cancer Institute [cited 4 May 2021]. Available from: https://btep.ccr.cancer.gov/question/faq/what-is-the-difference-between-rpkm-fpkm-and-tpm/

11.RSEM (RNA-Seq by Expectation-Maximization)[internet]biostat.wisc.edu. 2018[cited 20 May 2021] Available from:
http://deweylab.biostat.wisc.edu/rsem/rsem-calculate-expression.html#output

12. A cross-package Bioconductor workflow for analysing methylation array data [internet]. Bioconductor. 2021[cited 17 May 2021] Available from: http://bioconductor.org/packages/release/workflows/vignettes/methylationArrayAnalysis/inst/doc/methylationArrayAnalysis.html

13. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010 Nov 30;11:587. doi: 10.1186/1471-2105-11-587. PMID: 21118553; PMCID: PMC3012676.

14.Silva TC, Colaprico A, Olsen C, D'Angelo F, Bontempi G, Ceccarelli M, Noushmehr H. *TCGA Workflow*: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Res. 2016 Jun 29;5:1542. doi: 10.12688/f1000research.8923.2. PMID: 28232861; PMCID: PMC5302158.

Anexo 6: Código e informes:

Se anexan los informes de Rmarkdown con los resultados del workflow realizado en html y pdf. Además, se puede acceder al repositorio de github para ver el código utilizado.

https://github.com/SharonMQ/TFM_Sharon_Martinez_Quiroga.git

Por último, se anexa un archivo con los modelos obtenidos de iClusterBayes y los datos de mRNAseq.