Integración de datos ómicos y clínicos: nuevos avances en la frontera entre Biología y Medicina

Candelaria Lucía Hernández de la Fuente
Máster Universitario en Bioinformática y Bioestadística
Área 3: Subárea 1: *Análisis e integración de datos ómicos*

Ricardo Gonzalo Sanz
Ferrán Prados Carrasco

24/06/2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Integración de datos ómicos y clínicos: nuevos avances en la frontera entre Biología y Medicina*
Nombre del autor:	*Candelaria Lucía Hernández de la Fuente*
Nombre del consultor/a:	*Ricardo Gonzalo Sanz*
Nombre del PRA:	*Ferrán Prados Carrasco*
Fecha de entrega (mm/aaaa):	06/2020
Titulación:	*Máster Universitario en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Área 3: Subárea 1: Análisis e integración de datos ómicos*
Idioma del trabajo:	*Español*
Palabras clave	*Datos multi-ómicos, Biomedicina, selección de variables*

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

Hoy en día resulta imprescindible el uso de datos biológicos de alta resolución en un ámbito Biomédico. Esta información molecular, expresada en diferentes enfoques ómicos (genómica, transcriptómica, proteómica, metabolómica), explica diferentes parcelas de la variabilidad biológica humana. La integración de capas ómicas es un procedimiento que se viene realizando habitualmente mediante protocolos diversos. Sin embargo, la conjunción de información de alta resolución molecular con datos no-ómicos resulta más problemática. En este trabajo se pretende realizar un análisis sistemático de las metodologías enfocadas hoy en día a la integración de capas ómicas con la correspondiente información clínica, epidemiológica o demográfica de los pacientes. Este proceso se lleva a cabo mediante el uso de aproximaciones bioestadísticas alternativas (métodos multivariantes, de regresión o de redes de similaridad) que son evaluadas en términos de su rendimiento. En cuanto al *set* de datos empleado, hemos escogido un proyecto de *The Cancer Genome Atlas* (TCGA) compuesto por alrededor de 180 muestras analizadas para varias capas ómicas y con una rica información clínica disponible. En concreto, nos hemos centrado en las vertientes de expresión génica (transcriptómica) y metilación (epigenómica). Nuestros resultados muestran una elevada heterogeneidad entre los diferentes métodos empleados, en términos del procedimiento de integración, de selección de variables (genes/variables clínicas) y de explotación de los datos. La naturaleza y comportamiento de la variable respuesta escogida ha podido afectar a los resultados obtenidos en nuestro estudio. Posteriores estudios, basados en diferentes *datasets* o en variables respuesta

alternativas, podrían conseguir la construcción de modelos predictivos más robustos.

Abstract (in English, 250 words or less):

Nowadays, using high-resolution biological data in a Biomedical context is essential. This molecular information, showed in different omics approaches (genomics, transcriptomics, proteomics, metabolomics), explains different sides of the human biological variability. Omics data integration is a procedure usually performed by different methods. However, the inclusion of high-resolution molecular information with non-omics data is more troubled. The aim of the present study is to perform a systematic analysis of the methodologies focused to omics and clinical/epidemiological/demographic information from patients, by the use of alternative biostatistics tests (multivariate, regression or similarity network methods) that are evaluated in terms of their performance. We chose a *The Cancer Genome Atlas*- TCGA project, composed by around 180 samples analyzed for several omics layers and with a rich clinical information available. Specifically, we focused on gene expression (transcriptomics) and methylation (epigenomics). Our results showed a high heterogeneity among the different methods used, regarding integration process, feature selection (genes/clinical variables) and data mining. The nature and behaviour of the selected outcome could affect the results drawn from the present work. Further studies, based on different datasets or alternative outcome variables, could reach the development of stronger predictive models.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

En este trabajo se pretende realizar un análisis sistemático de las metodologías enfocadas hoy en día a la integración de diferentes capas ómicas con la correspondiente información clínica, epidemiológica o demográfica de los pacientes. Este proceso se lleva a cabo mediante el uso de aproximaciones bioestadísticas alternativas, y se valoran los métodos que ofrecen los resultados más óptimos para el análisis de un determinado *set* de datos. En concreto, se busca la elaboración de modelos predictivos para poder clasificar a los individuos conjugando información ómica y no-ómica. Se estudian las estrategias analíticas más usadas y las más recientemente descritas mediante una exhaustiva revisión bibliográfica.

Hoy en día resulta imprescindible el uso de datos biológicos de alta resolución en un ámbito Biomédico. Esta información molecular, expresada en diferentes enfoques ómicos (genómica, transcriptómica, proteómica, metabolómica), explicaría diferentes parcelas de la variabilidad biológica humana y tiene grandes aplicaciones en cuanto al estudio de diversas enfermedades, ya que la mayoría de los desórdenes humanos son complejos, y suceden por la interacción de múltiples factores tanto genéticos como ambientales.

La integración de capas ómicas es un procedimiento que se viene realizando habitualmente mediante protocolos diversos. Sin embargo, la conjunción de la información de alta resolución molecular con los datos no-ómicos resulta más problemática, y es esta integración la que supondría un sustancial avance al conocimiento, ya que nos permitiría abordar conjuntamente los fenotipos de las enfermedades con las bases moleculares subyacentes.

Hay diversas dificultades que se asocian a este procedimiento integrador, ligadas tanto a la propia naturaleza de los datos no-ómicos (datos de naturaleza compleja, difícilmente estandarizables debido a que están sujetos a cada diseño experimental, exceso de *missing data*, etc.), como a la interacción entre datos ómicos y no-ómicos (donde surgen problemas como el *ascertainment bias*). Desafíos adicionales se fundamentan en la diversidad de tipos de datos existentes y sus

formatos. Sin embargo, la problemática central de la integración de datos es la selección de variables a considerar y este paso se aborda mediante diferentes metodologías que pueden ofrecer resultados no simétricos entre ellas.

Por otro lado, existe una gran variedad de estrategias en cuanto al proceso intrínseco de la integración, en función de si los modelos se construyen utilizando los datos ómicos y no-ómicos de manera independiente o condicional.

La temática abordada en el presente trabajo, por tanto, reviste un especial interés para la comprensión de diferentes desórdenes humanos. Asimismo, la integración de datos, junto con novedosos enfoques concretos, como por ejemplo la farmacogenómica o la nutrigenómica, representan puertas abiertas para la investigación científica del presente y de los próximos años, nos abocan a la eliminación de las fronteras entre Biología y Medicina, y nos conducen al desarrollo de una Medicina personalizada y de enfoques terapéuticos individualizados.

1.2 Objetivos del Trabajo

Este Trabajo tiene los siguientes objetivos generales:

1. Realizar un análisis sistemático de los diferentes enfoques utilizados hoy en día para la integración de datos ómicos y no-ómicos, centrándonos en los más usados.
2. Implementar modelos predictivos basados en diversos métodos bioestadísticos que nos permitan llevar a cabo clasificaciones de individuos teniendo en cuenta conjuntamente las vertientes ómicas y no-ómicas de la información disponible para ellos.
3. Valorar la eficacia y rendimiento de cada modelo y evaluar las problemáticas asociadas a su uso.

Dichos puntos serán abordados por una serie de objetivos específicos que los desarrollan:

1. Realizar un análisis sistemático de los diferentes enfoques utilizados hoy en día para la integración de datos ómicos y no-ómicos.
    a. Buscar exhaustivamente en la literatura científica la información relacionada con este tema.

b. Caracterizar los diferentes enfoques con la información recopilada en el objetivo anterior. En concreto, nos centraremos en el estudio de métodos de regresión logística, métodos multivariados y en la aplicación de enfoques de similaridad o redes.

c. Comprender las bases bioestadísticas de cada uno de los métodos identificados y las ventajas y limitaciones de los mismos.

2. Implementar modelos predictivos basados en diversos métodos bioestadísticos que nos permitan llevar a cabo clasificaciones de individuos teniendo en cuenta conjuntamente las vertientes ómicas y no-ómicas de la información disponible para ellos.

a. Construir los modelos predictivos correspondientes.

b. Valorar las diferencias encontradas entre modelos e identificar los enfoques asociados a la selección de variables como etapa clave.

c. Aplicar dichos modelos a un *dataset* concreto.

3. Valorar la eficacia y rendimiento de cada modelo y evaluar las problemáticas asociadas.

a. Estudiar, mediante parámetros objetivos, la optimización de cada modelo con el set de datos seleccionado. Estos parámetros incluyen la estimación de errores (*root-mean-square error*), de exactitud (*accuracy*) y utilización de curvas ROC (*Receiver Operating Characteristic*).

b. Identificar de manera práctica las limitaciones de cada metodología.

c. Estandarizar el protocolo o flujo de trabajo para la aplicación de los modelos a cualquier *dataset* que cumpla con determinadas especificaciones.

1.3 Enfoque y método seguido

El enfoque propuesto para la consecución de los objetivos mostrados en el ***Apartado 2*** comienza con una búsqueda bibliográfica actualizada de palabras clave adecuadas (*omics data integration, omics and clinical data,* etc.) en la base de datos PubMed de los NCBI. Las publicaciones científicas seleccionadas se basan en la temática específica e incluyen las aproximaciones más novedosas, lo que resulta fundamental para poder abordar de manera realista y con buen rendimiento la problemática que se pretende valorar aquí.

A continuación, se escoge un set de datos que resulte adecuado y relevante por contener información tanto ómica como no-ómica de interés para la valoración de alguna enfermedad. Este *dataset* se recopila a partir de repositorios públicos de datos moleculares.

Tras ello, se procede a realizar la integración de dichos datos con metodologías que se basan en diferentes enfoques bioestadísticos, y se valora el rendimiento de los mismos. Mediante esta estrategia podemos conjugar una buena representación de la variabilidad de métodos existentes para realizar el mismo procedimiento integrador.

1.4 Planificación del Trabajo

1.4.1. Tareas

Las tareas para la consecución de los objetivos específicos tendrán un marco temporal y una duración determinada, para asegurar el éxito del proyecto. Las tareas planteadas serán las siguientes:

- **Desarrollo del *Plan de Trabajo***. Elaboración del presente documento. Duración: 2 semanas.
- **Búsqueda bibliográfica de literatura científica** (artículos publicados y *preprints*) relacionados con métodos de integración de datos ómicos y no-ómicos. La revisión se realiza utilizando la base de datos PubMed-NCBI. Duración: 2 semanas.
- **Revisión bibliográfica del contexto teórico del trabajo**. Se realiza en paralelo a la tarea anterior. Duración: 2 semanas.

- **Selección del *dataset*** adecuado para testar los diferentes modelos. La búsqueda se centra en los repositorios *GEO Omnibus* y *TCGA*. Desarrollo del análisis exploratorio de las capas ómica y clínica, así como estudio de la expresión diferencial de genes para transcriptómica. Duración: 3 semanas.
- **Identificación de los diferentes enfoques bioestadísticos** para la integración de datos y la creación de modelos predictivos. Duración: 1 semana y media.
- **Análisis de las variables** contenidas en el *dataset* y procesamiento de las mismas para ser incluidas en los modelos. Duración: 2 semanas.
- **Construcción de los modelos predictivos** y análisis de las diferencias resultantes entre ellos. Duración: 1 semana y media.
- **Valoración del rendimiento** de cada modelo e identificación de limitaciones. Duración: 1 semana.
- **Estandarización del *pipeline*** asociado a cada modelo para su posible uso en *datasets* alternativos. Duración: 1 semana y media
- **Redacción de la *Memoria***. Este proceso se plantea de manera horizontal a lo largo de todo el semestre, ya que comprende desde la redacción de los capítulos iniciales (*Introducción* y *Material y Métodos*) como la exposición de los principales resultados y conclusiones.
- **Preparación de la presentación virtual**. Duración: 1 semana.

1.4.2. Calendario

El cronograma con las tareas contextualizadas en un marco temporal se puede consultar en la **Figura 1**. Este calendario se muestra en forma de diagrama de Gantt creado con el programa *SmartSheet*. Se muestran conjuntamente las tareas (***Apartado 4.1***) y los hitos (***Apartado 4.3***).

1.4.3. Hitos

Los diferentes hitos del presente proyecto, basados en el *Plan Docente* de la asignatura, se muestran en la ***Tabla 1***. Estos hitos se han integrado junto con las tareas en el calendario o cronograma (ver ***Apartado 4.2***).

Tabla 1. Principales hitos del presente trabajo y fecha de consecución de cada uno. *Nota:* debido a la ampliación de fechas de entrega, los hitos PEC3-PEC5b fueron atrasados con respecto a lo inicialmente previsto.

Hito	Fecha de consecución
Plan de trabajo (*PEC1*)	16/03/2020
Desarrollo del trabajo – Fase 1 (*PEC2*)	22/04/2020
Desarrollo del trabajo – Fase 2 (*PEC3*)	01/06/2020
Cierre de la memoria (*PEC4*)	24/06/2020
Elaboración de la presentación (*PEC5a*)	28/06/2020
Defensa pública (*PEC5b*)	01/07/2020

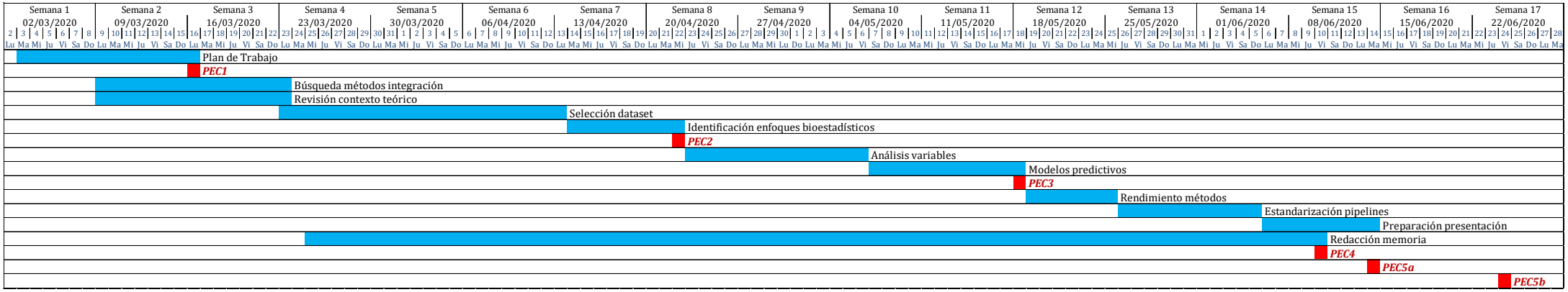

Figura 1. Cronograma con la planificación temporal del trabajo. *Nota:* se presenta el calendario propuesto inicialmente en el *Plan de Trabajo*. Las fechas de las últimas tareas fueron ampliadas pero la estructura y planificación de las mismas ha seguido invariable.

1.4.4. Análisis de riesgos

Existen una serie de factores que pueden afectar al desarrollo y seguimiento del *Plan de Trabajo* presentado aquí:

- Dificultades en encontrar un *dataset* adecuado que englobe datos ómicos y no-ómicos relevantes y que permita la implementación de modelos predictivos.
- Dificultades en el acceso público al *dataset* de interés. Este problema será fácilmente solventable ya que tenderemos a escoger aquellos repositorios de información de alta resolución que sean públicos (por ejemplo, *The Cancer Genome Atlas*, *Gene Expression Omnibus, GEO*).
- Problemas en la preparación de los datos concretos seleccionados, por ejemplo, respecto a que exista una identidad entre los nombres de pacientes entre *datasets* o que haya diferentes tipos de variables en la integración. Para evitar esta problemática, se debe ser sistemático en el procesamiento de datos, y estar atento a la posible ocurrencia de incidencias, siempre realizando comprobaciones de los pasos realizados mediante *feedback* (por ejemplo, ir revisando si el número de variables o número de pacientes coincide con lo esperado tras un determinado procesamiento, etc.).
- Otro posible problema que puede surgir se basa en la capacidad de computación que puede requerir el *set* de datos escogido, para realizar los análisis bioinformáticos en un ordenador personal. Esto puede ser solventado en parte procurando no escoger un *dataset* de gran tamaño muestral.
- Incidencias en cuanto a que exista una buena representación de ambas vertientes de los datos: ómica y no-ómica, en cuanto al número de variables seleccionadas finalmente. Se debe valorar este aspecto de manera especial y sus implicaciones.

1.5. Breve sumario de productos obtenidos

Los resultados esperados del presente trabajo se irán estructurando en función de los documentos requeridos en el *Plan Docente*, y comprenden los siguientes ítems o entregables:

1.5.1. Plan de trabajo

Este documento es la expresión y planificación del presente trabajo y engloba los principales aspectos que se van a desarrollar en el mismo. Su redacción permite realizar una autoevaluación del cumplimiento de todos los puntos recogidos cuando se finalice con el desarrollo del trabajo y la redacción de la memoria.

1.5.2. Memoria

La *Memoria* refleja tanto el contexto teórico de la temática escogida y del área de estudio como el desarrollo del trabajo en sí, incluyendo todo el material empleado, los resultados obtenidos, y un análisis sintético del conjunto del documento (*Conclusiones*).

1.5.3. Producto

El principal producto obtenido en este trabajo serán los modelos predictivos generados, y, de manera más concreta, el protocolo o *pipeline* (*workflow*) para ser implementado en diferentes *datasets.*

1.5.4. Presentación virtual

La presentación virtual va a resumir de manera sintetizada el conjunto del proyecto desarrollado.

1.5.5. Autoevaluación del proyecto

Por último, se realiza un análisis crítico de la planificación inicial de objetivos, tareas e hitos y de si se han podido llevar a cabo en el marco temporal previsto. Esta sección será integrada en el capítulo de *Conclusiones*. Asimismo, se plantean aquellas dificultades que han surgido a lo largo del desarrollo del trabajo y que no habían sido inicialmente previstas.

1.6. Breve descripción de los otros capítulos de la memoria

Este *Trabajo Fin de Máster* conlleva redacción del *Plan de Trabajo* y, a continuación, del resto de la *Memoria*. La *Memoria* viene estructurada, como cualquier trabajo científico-técnico, en los siguientes apartados:

- *Introducción*. Aquí se desglosa el contexto teórico que se requiere para comprender la problemática que se plantea en el trabajo.
- *Material y Métodos*. Este apartado comprende la exposición de los recursos utilizados para el desarrollo del trabajo, tanto respecto a los datos usados, como a las fuentes de información de origen. También, muestra las metodologías que han sido empleadas y los detalles técnicos asociados. Esta sección es fundamental para asegurar la reproducibilidad de la investigación.
- *Resultados*. Empleando los recursos presentados en el anterior apartado, se muestran los principales resultados del trabajo. Asimismo, se van a detallar las posibles dificultades técnicas que surgen en su desarrollo.
- *Discusión*. Se realiza una discusión expositiva de los resultados obtenidos en el marco teórico presentado, mostrando reflexiones acerca del alcance del trabajo.
- *Conclusiones*. Se presentan los puntos fundamentales que emergen del trabajo y resumen los principales hitos alcanzados en el mismo. Además, incluye una reflexión crítica acerca de la consecución de los objetivos planteados en el *Plan de Trabajo* y de la planificación temporal y metodología propuesta. Por último, se muestran las posibles líneas futuras de trabajo que no han podido ser valoradas en el presente trabajo.
- *Glosario*. Se presentan los principales términos empleados en la *Memoria* y su definición.
- *Bibliografía.* Se incluyen las referencias bibliográficas más relevantes respecto a la temática abordada.

2. Desarrollo del trabajo

2.1. Introducción

La gran implantación en numerosos ámbitos, junto con el abaratamiento de los costes, de las técnicas de análisis de nueva generación está permitiendo que se genere un elevado número de datos (*big data*) basados en diferentes enfoques "ómicos". Este tipo de información puede obtenerse en población sana, pero resulta muy interesante su aplicación en pacientes de diversas enfermedades, ya que la mayoría de los desórdenes humanos son complejos, y suceden por la interacción de múltiples factores tanto genéticos como ambientales.

El término *ómico* se puede definir como el campo de estudio de las Ciencias de la Vida que se centra en información o datos de gran escala para comprender la vida. [1]. Más concretamente, se conceptualiza como una serie de métodos globales para la cuantificación de familias de moléculas celulares –ADN, ARN, proteínas y metabolitos intermedios [2]. Estas moléculas también se pueden denominar *biomarcadores*, denotando que describen el normal o anormal funcionamiento de un organismo y que pueden ser analizados y detectados en circulación (sangre, linfa), en tejido o en fluidos corporales (orina, saliva…) [3].

Las diferentes *capas* o abordajes ómicos han sido definidas secuencialmente en función de los descubrimientos técnicos y científicos que se han ido desarrollando, comenzando por la *genómica*, concepto acuñado en 1986 por el genetista Thomas H. Roderick. Posteriormente, el término *proteómica* fue propuesto inicialmente por Marc Wilkins en 1995 para describir el complemento proteínico completo de un organismo.

Más recientemente, el cambio de paradigma tecnológico de la pasada década se ha fundamentado en la adopción y uso global de las tecnologías de alta resolución, que ha permitido a genetistas, biólogos y bioestadísticos unir el *gap* entre genotipo y fenotipo a una escala inconcebible hace solo unos cuantos años [4].

2.1.1. Las ciencias ómicas

Para mostrar la importancia y el desarrollo de estas aproximaciones técnicas, podemos realizar un análisis exploratorio de su impacto en las investigaciones

científicas internacionales, mediante el estudio del número de publicaciones dedicadas a estas temáticas.

La **Figura 2** muestra una definición gráfica de la importancia y la evolución que ha tenido el estudio de las ciencias multi-ómicas a lo largo de los últimos 30 años. Se puede observar la prevalencia de la genómica frente al resto de abordajes experimentales, y la tendencia exponencial que ha experimentado hasta la actualidad en término de número de publicaciones. El segundo enfoque en interés en cuanto a aplicabilidad es la proteómica. El resto de términos tendrían un alcance más restringido en la comunidad científica, aunque es interesante observar la tendencia del estudio del *microbioma*, que es reflejo del gran interés que está alcanzando especialmente en los últimos años.

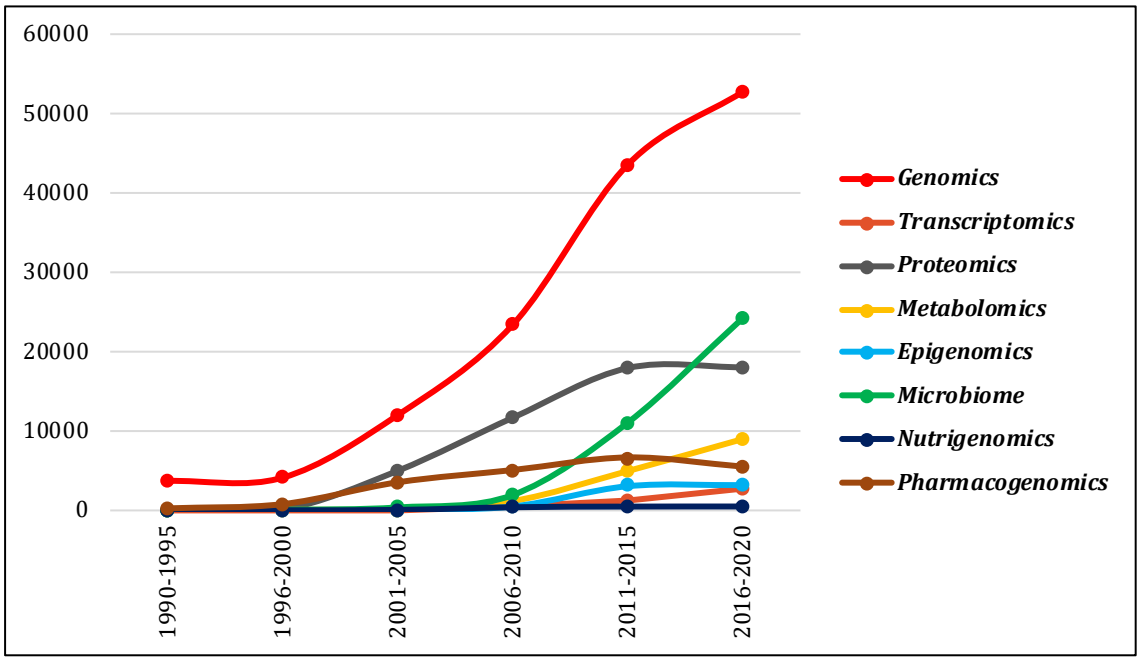

Figura 2. Número de publicaciones indexadas para los términos de las diferentes ciencias ómicas. Para todas las búsquedas se ha empleado la expresión "*human AND genomics*", etc. Fuente: NCBI-PubMed.

El desarrollo de estas tecnologías puede ser expresado en términos del clásico **dogma central de la Biología Molecular**, lo que nos ayuda a comprender las relaciones entre las diferentes capas ómicas [2]. La **Figura 3** muestra estas interacciones y la definición de las ómicas en función de las familias de biomoléculas que estudian.

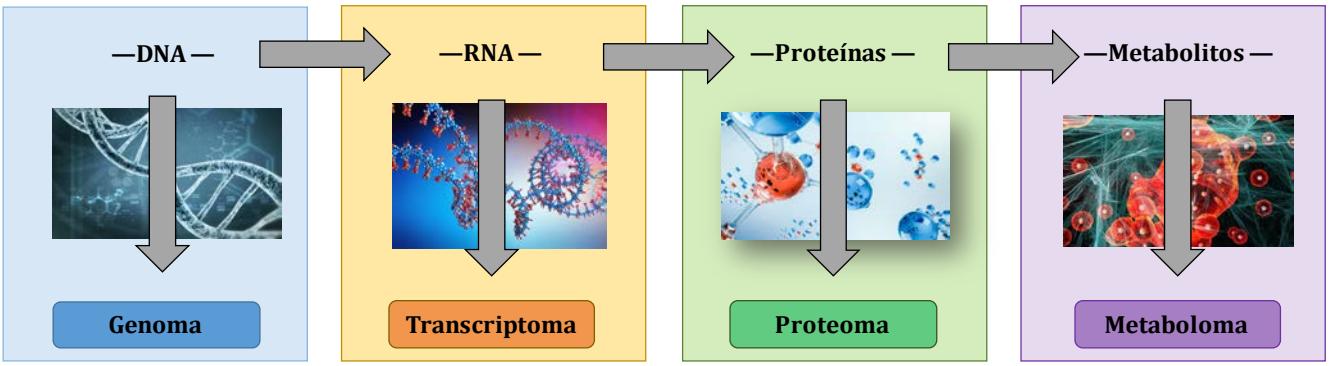

Figura 3. Dogma central de la Biología Molecular en la era de las ciencias ómicas. Se representan la jerarquía o flujo entre los enfoques ómicos y su interacción. Fuente: adaptado de (Debnath et al. 2010).

A continuación, se presenta una breve reseña de estas tecnologías ómicas, las más frecuentemente analizadas [2,3,5]. Al margen de estas aproximaciones, existen otras interesantes aplicaciones de las ciencias ómicas que cabe destacar aquí: *i) toxicogenómica*, como el estudio de la relación entre la estructura y actividad del genoma y los efectos biológicos adversos de agentes exógenos y *ii) nutrigenómica* o la aplicación de las ciencias ómicas en la nutrición humana, especialmente respecto a la relación entre nutrición y salud. Esta última idea es de gran actualidad y nos ayuda a afrontar retos derivados de los contextos alimentarios extremos en los que se sitúan las sociedades occidentales (por el exceso y ocurrencia de sobrepeso y obesidad ligados al síndrome metabólico) y los países en desarrollo (por desnutrición y hambrunas).

Genómica

Se trata del análisis de la secuencia completa de ADN de una célula u organismo. La información hereditaria en el hombre se encuentra ubicada tanto en el núcleo como en otros orgánulos (la mitocondria), e incluye tanto los genes como secuencias no codificantes del ADN. Las secuencias parciales o completas de ADN pueden ser aisladas utilizando diferentes plataformas experimentales, fundamentalmente expresadas en *chips* o *arrays* de *Single Nucleotide Polymorphisms,* SNPs (uso simultáneo de miles de sondas oligonucleotídicas que hibridan con secuencias de ADN específicas en las cuales se conoce que aparecen variantes nucleotídicas) o bien mediante técnicas de secuenciación de nueva generación de ADN (*whole genome sequencing, exome sequencing,* etc.).

El desarrollo de la genómica fue posible en inicio gracias al descubrimiento de la tecnología de clonaje de genes y la habilidad de amplificarlos (reacción en cadena de la polimerasa, PCR) y secuenciarlos (secuenciación Sanger), y fue notablemente potenciado de la mano de grandes proyectos internacionales, fundamentalmente el Proyecto Genoma Humano (*Human Genome Project*). La genómica, en definitiva, ha abierto una nueva era respecto a las Ciencias Biomédicas, permitiendo la identificación de todas las variantes genéticas de un individuo, y por tanto suponiendo el descubrimiento de variantes que causan enfermedades complejas y proporcionando dianas terapéuticas.

Transcriptómica

Estudio del *set* completo de transcritos de ARN generados a partir del ADN en una célula o tejido. En el hombre, en torno al 2% del genoma se representa en el transcriptoma como genes codificantes de proteínas. Incluiría tanto el ARN ribosómico, mensajero, de transferencia o micro ARN. Cada célula expresa, en su desarrollo, diferentes genes en diferentes momentos y bajo diferentes condiciones fisiológicas. En cuanto a las estrategias experimentales para la caracterización de estos transcritos, la situación es paralela al caso de la genómica, utilizándose tanto *microarrays* como secuenciación de ARN (RNAseq). Es interesante destacar aquí que la potencialidad o poder de la transcriptómica radica, entre otras vertientes, en la habilidad de sub-clasificar desórdenes que aparentemente pudieran ser similares.

Proteómica

Análisis del *set* completo de proteínas expresado en una célula, tejido u organismo. Esta dimensión de por sí es muy compleja, ya que las proteínas sufren diferentes modificaciones post-traduccionales (por ejemplo, fosforilación o ubiquitinización), muestran diversas configuraciones espaciales y localizaciones intracelulares. Además, experimentan interacciones con otras proteínas o moléculas. El análisis del proteoma se realiza mediante espectrometría de masas y técnicas de *microarray*.

En esencia, se puede considerar que se trata de una capa ómica que complementa y no reemplaza a la genómica. Su estudio englobaría el análisis de funciones y procesos biológicos muy diversos y con grandes repercusiones en el ámbito de la

salud y enfermedad humana, como por ejemplo la producción de hormonas o la modulación de la respuesta inmunitaria.

Existen diferentes consorcios que procuran realizar un catálogo de todas las proteínas humanas (*Human Proteome Organization*, HUPO). Uno de los enfoques más prometedores en la intersección entre el análisis de genes y proteínas es el desarrollo de nuevos potenciales fármacos para el tratamiento de la enfermedad, para lo cual es necesario comprender la estructura y función de cada proteína y la complejidad de las interacciones entre los componentes del proteoma.

Metabolómica

Evaluación del *set* completo de las pequeñas moléculas de metabolitos que se encuentran en una muestra biológica. Incluiría intermediarios de vías de carbohidratos, lípidos, aminoácidos, y hormonas y otras moléculas de señalización, así como sustancias exógenas (fármacos). El metaboloma es dinámico y puede variar dentro del mismo organismo y entre organismos de la misma especie debido a varios factores (cambios en la dieta, estrés, actividad física, efectos farmacológicos y enfermedad) y puede ser evaluado mediante espectrometría de masas. Por tanto, el estado de los metabolitos refleja una codificación por parte del genoma, pero también una modificación en respuesta a factores ambientales. Los metabolitos, como reguladores clave de la homeostasis del sistema, pueden ser usados para detectar cambios fisiológicos causados por tóxicos o químicos o relacionados con síndromes específicos. El metaboloma proporciona una gran potencialidad en cuanto al desarrollo de marcadores diagnósticos para el estado de una enfermedad, la sub-clasificación de la misma y el estudio de los mecanismos que subyacen en estos diferentes fenotipos.

Relacionando el proteoma y metaboloma, podemos avanzar el concepto de *interactoma*, como la ocurrencia de redes complejas de interacciones entre macromoléculas [6]. Las redes celulares resultantes son las que conforman la mayoría de las relaciones genotipo-fenotipo y, por tanto, son básicas para comprender la enfermedad humana.

Epigenómica

Por último, queremos brevemente resaltar una capa que se situaría en una posición jerárquica superior a la genómica: la epigenómica, representada por las modificaciones químicas reversibles del ADN o de las histonas que se unen al ADN y que producen cambios en la expresión de los genes sin alterar la secuencia de bases. Estas modificaciones pueden ser el resultado de un contexto celular o tisular determinado, de una respuesta a factores ambientales o también del desarrollo de diferentes estados de una enfermedad y pueden persistir a través de las generaciones.

2.1.2. Desafíos y oportunidades de las aproximaciones multi-ómicas

Hemos desgranado la dimensión y potencialidad de los datos multi-ómicos en el campo de la Biomedicina. Sin embargo, su implementación práctica supone enfrentarse a numerosos retos, englobados en diferentes vertientes.

La aplicación de las ciencias ómicas en Medicina

Para conseguir que los enfoques ómicos puedan permear al campo clínico [7], se deben superar determinados aspectos relacionados con:

- La diseminación, gestión e interpretación de los datos ómicos en un contexto clínico. La democratización de los datos multi-ómicos es un factor clave para su aplicación en Medicina. Las barreras físicas al acceso, gestión, y transferencia de los datos se eliminan mediante la digitalización de los archivos, sin embargo, la utilidad clínica de los datos de investigación se limita por la privacidad y por otras barreras, para evitar el abuso de información protegida y sensible. Transformar e incorporar datos provenientes de aproximaciones ómicas a un contexto clínico es esencial, pero a su vez complejo y problemático.
- La garantía de que los resultados ómicos puedan suponer un valor añadido en cuanto a los existentes paradigmas del cuidado al paciente o a las decisiones o prácticas clínicas actuales.

Desafíos informáticos y tecnológicos en la era ómica

Existen otra serie de desafíos ligados al procesamiento de las grandes cantidades de datos generados por estas tecnologías. De manera global, estos retos comprenden

problemas en cuanto a colección, análisis, minería, transferencia, visualización y archivo de estos *big data* [4].

A continuación, se exponen otros aspectos más concretos y algunas soluciones y abordajes para enfrentarse a ellos [7].

En primer lugar, aquellos referidos al **almacenamiento de datos y organización de *datasets*.** Este tipo de *set* de datos suele ser heterogéneo por naturaleza, tanto desde una perspectiva intra-ómica (p.e., datos genómicos provenientes de diferentes plataformas) como inter-ómica (proteómica vs. genómica). Esto genera una falta de reproducibilidad de los datos resultantes de análisis no estandarizados. Algunas recomendaciones para abordar estas problemáticas se centran en el establecimiento de estándares para todos los tipos de datos ómicos (mediante, por ejemplo, la creación de consorcios) y también, en paralelo, proporcionar incentivos para que los investigadores que depositen sus resultados en repositorios públicos.

El empleo de diferentes bases de datos y repositorios es básico para el intercambio de información [8]. Algunas de estas iniciativas relacionadas con datos ómicos serían *Human Genome Project, Ensembl, 1000 Genomes Project, International Cancer Genome Consortium* (genómica), *Gene Expression Omnibus, GEO* (proteómica), *Cancer Genome Project* (fenómica) [9].

Asimismo, el procesamiento de datos ómicos **requiere una formación multidisciplinar.** El tratamiento de este tipo de datos debe realizarlo personal altamente especializado, con conocimientos de diversos campos (Biología, Medicina, Informática, Estadística), además requiriendo una constante actualización de los conocimientos mediante el aprendizaje de nuevas técnicas tanto de laboratorio como de tratamiento de datos. Esta situación puede solventarse de dos maneras: *i)* contando con personal con una formación muy amplia en estos campos o *ii)* formando un equipo multidisciplinar compuesto por clínicos, científicos experimentales y bioinformáticos, que permitan que todas las fases de un experimento, desde el muestreo, diseño experimental y bio-interpretación, así como los pasos translacionales clínicos (ensayos clínicos e implementación práctica) queden cubiertas.

Del mismo modo, **el procesamiento bioinformático de los datos ómicos** (*pipeline* o *workflow*) es un proceso de pasos múltiples sobre el que se debe ser

sistemático para garantizar que las inferencias finales sean sólidas y se sustentes sobre un análisis consistente. Aquí queremos detenernos de manera más concreta para desgranar las etapas básicas que debe comprender este procesamiento [3]:

1. *Control de calidad de los datos (QC)*. Dado que los *dataset* ómicos están compuestos por miles o millones de medidas, este primer paso es crucial. Dentro de esta etapa, deberíamos eliminar aquella información (p.e., genes), que se expresen por debajo de un determinado umbral, lo que va a eliminar ruido de fondo para el análisis, así como agilizar la computación. Del mismo modo, se deben evaluar aquí posibles *batch effects* (factores no-biológicos que pueden influir en el experimento, por ejemplo, debidos a reactivos, instrumental, etc.) o una posible heterogeneidad entre las muestras.
2. *Desarrollo de modelos computacionales.* Tras este primer paso, se deben diseñar y desarrollar diferentes *tests* candidatos que permitan asociar determinados sub-grupos biológicos, o resultados clínicos de cierto fenotipo de interés, con las medidas ómicas. Se pueden escoger un número ilimitado de herramientas estadísticas con este objetivo. El proceso comenzaría realizando una selección de variables (*feature selection*), que escogerá una serie de variables (p.e. genes) que parezcan biológicamente relevantes para dicho fenotipo. Desde ese momento, se escogería ese *subset* de variables para desarrollar un modelo computacional específico que permita predecir el resultado clínico sobre la base de las medidas ómicas.

   En este punto, debemos tener en cuenta un fenómeno relevante en cuanto a la dimensión ómica de los datos (*overfitting*). Este problema (*large p, small n*), supone que se analicen grandes cantidades de variables o factores ómicos (en ocasiones por encima de cientos de miles *p* o número de medidas por muestra) en tamaños muestrales más discretos (valores de *n* de cientos o por debajo de 100 muestras independientes). La estrategia más adecuada para evitar esta problemática en nuestros modelos es utilizar una aproximación de *sets* de análisis diferenciales. Una partición de los datos (*training*) será empleado para la construcción, ajuste o entrenamiento del modelo, mientras que otra diferente y no redundante (*test*) será empleado para evaluar el rendimiento del modelo.

3. *Confirmación del modelo computacional en un dataset independiente.* En este punto, se implementaría el modelo construido en un set de datos independiente, generado por diferentes instituciones, instrumental, formado por diferentes muestras del *dataset* sobre el cual se construyó el modelo.
4. *Lanzamiento o publicación de datos código y procedimientos computacionales a la comunidad científica.* Los resultados deben ser puestos a disposición de la comunidad internacional en repositorios públicos *online*, tanto referidos a los procedimientos (por ejemplo, *GitHub*) como a los propios resultados ómicos (*GEO, European Genome-Phenome Archive*, etc.). Este paso va a asegurar la reproducibilidad de la investigación.
5. *Traslación a un contexto clínico.* Este último paso, el más translacional, es también el más complejo y que va a requerir un importante proceso de *feedback.* Los métodos usados para obtener las medidas ómicas deben ser modificados para establecer un ensayo que sea clínicamente y económicamente asumible y suficientemente robusto para ser implementado en la práctica clínica.

Al margen de los puntos arriba tratados, existe un enfoque con gran potencialidad ya que permite un análisis simultáneo de las diferentes capas ómicas y un notable incremento de la información que puede ser obtenida en relación con la salud y enfermedad humana: la integración multi-ómica (integrÓmica).

2.1.3. La integración de datos ómicos

Para poder estudiar procesos biológicos complejos de manera holística, se requiere tener en cuenta un enfoque integrador que combine datos multi-ómicos para resaltar las interrelaciones de las biomoléculas y sus funciones. Además, yendo más allá, la integración de estas capas ómicas con la información clínica es crucial. Sin embargo, este proceso requiere el uso de herramientas, métodos y plataformas que permitan el análisis, visualización e interpretación de los datos multi-ómicos [10].

Estrategias de integración

El término "integración de datos" se refiere a la situación en la que, frente a un mismo sistema, existen diferentes fuentes de datos disponibles y se pretende un estudio

conjunto de los mismos [8]. La integración puede ser concebida también, además de integrar diferentes tipos de datos ómicos para la misma cohorte de pacientes, como la integración de un mismo tipo de dato ómico a través de diferentes estudios. Aquí se han puesto de manifiesto, por tanto, las dos estrategias de integración: *i)* horizontal, donde se combinan capas ómicas de la misma naturaleza, y *ii)* vertical, donde se integran diferentes dimensiones ómicas (ver **Figura 4**).

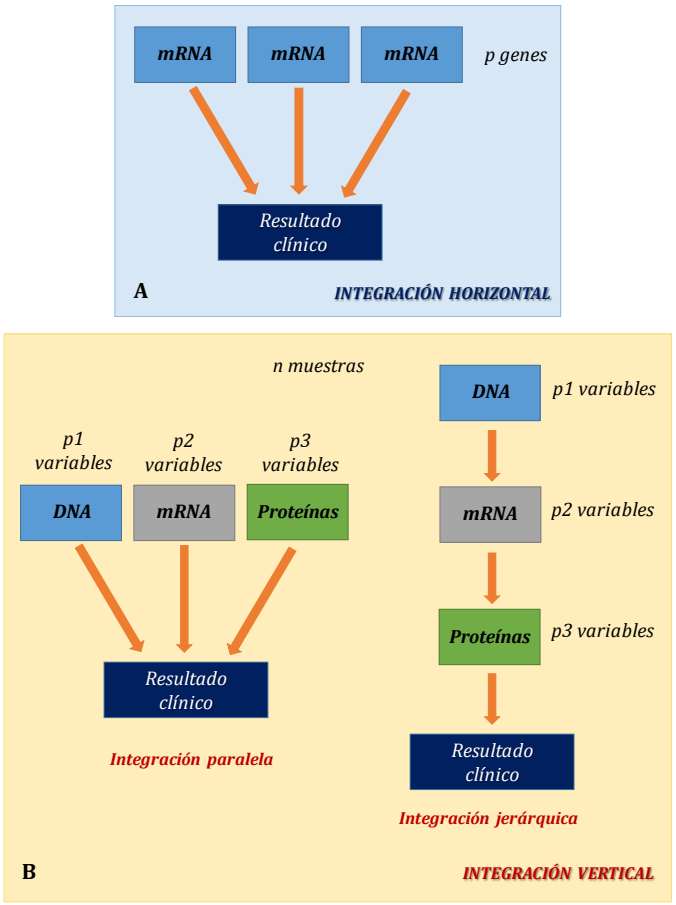

Figura 4. Diferentes estrategias de integración: **A:** horizontal y **B:** vertical. Fuente: adaptado de [11].

La mayoría de métodos estadísticos de integración vertical de datos ómicos multi-dimensionales se basan en modelos y pueden ser caracterizados por análisis de regresión (métodos supervisados) o exploratorios (métodos no supervisados), dependiendo de si el objetivo del estudio es la predicción bajo rasgos fenotípicos, como por ejemplo un fenotipo de enfermedad, estatus de cáncer o no.

Dentro de los enfoques verticales, existen dos abordajes diferenciales [11]:

- Integración paralela, que supone el tratamiento de cada capa ómica de manera equiparable.
- Integración jerárquica, que incorpora el conocimiento previo de la relación entre diferentes plataformas de datos ómicos.

Para todas las estrategias, es crucial tener en cuenta el carácter complejo e los datos ómicos, en dos vertientes: *i)* cada tipo de medida ómica es multi-dimensional por sí misma, por lo que cuando se realiza un análisis integrativo, los datos agregados de diferentes niveles son de mayor dimensión aún y *ii)* dentro de las variables ómicas de alta dimensión, solo un pequeño *subset* tendrá importantes implicaciones y repercusiones en los fenotipos de interés. Por tanto, la *selección de variables* juega un importante papel en estos procedimientos.

Integración de la ómica y la clínica

Existe una gran multitud de metodologías para la integración de diferentes capas ómicas. Sin embargo, pocos algoritmos muestran una verdadera habilidad de ser implementados en los campos clínicos y de salud pública, pese a que resulta crucial poder integrarlas en el mismo modelo. Solo un pequeño número de estudios consiguen alcanzar una verdadera integración de datos ómicos y no-ómicos.

Los principales problemas en cuanto a la integración de los datos ómicos y no-ómicos se fundamentan principalmente en la naturaleza de los datos y la relación entre ellos. Los datos no-ómicos son complejos y definidos de manera heterogénea, en muchos casos, su registro carece de estándares y se dificulta su integración en modelos predictivos y de inferencia. Los datos epidemiológicos dependen del diseño de las encuestas y de su estandarización. También las variables clínicas pueden ser complejas en su definición. Por ejemplo, un estadio tumoral resulta de la combinación de patología e información de imágenes [12].

La integración de datos clínicos, densos en información, y los ómicos han atraído la atención en el campo del modelado predictivo de los resultados clínicos, pese a que, habitualmente los estudios prestan más atención a la modelización de la parte ómcia que clínica. Si las variables clínicas son insuficientemente incorporadas al modelo, el valor predictivo de los datos ómicos puede ser sobreestimado. Es decir, que los datos

ómicos pueden parecer más útiles de lo que son actualmente, como resultado de no explorar completamente el potencial de las variables clínicas [13].

2.2. Material y Métodos

2.1.1. Selección del *dataset*

Para el presente trabajo, se realizó una búsqueda de un *set* de datos adecuado que comprendiese datos tanto de tecnologías ómicas como datos clínicos y epidemiológicos. Con este objetivo, se consultó la base de datos *The Cancer Genome Atlas* (TCGA) a través del portal *Genomic Data Commons Data Portal* (https://portal.gdc.cancer.gov/). Este repositorio incluye una de las mayores colecciones de *datasets* multi-ómicos para un gran número de tipos diferentes de cáncer y nos permite unir, analizar e interpretar los perfiles de ADN, ARN, proteínas y cambios epigenéticos de muestras tumorales junto con datos clínicos, epidemiológicos e histológicos.

Los criterios utilizados para la selección de los datos se centraron en los siguientes puntos:

- **Tamaño muestral**: para la búsqueda de los proyectos se planteó un número de individuos manejable, menor de 200, para minimizar los problemas de computación a la hora de realizar los análisis que pudieran surgir en grandes *datasets*. Por otro lado, este tamaño muestral aproximado también nos va a permitir que los resultados obtenidos sean robustos.
- **Disponibilidad de diferentes capas ómicas y la correspondiente información clínica para cada paciente.** Se buscó un *set* de datos que englobase un buen número de tecnologías ómicas, por tanto, con potencialidad para desarrollar diferentes análisis. Además, se comprobó que toda la información estuviera disponible para todos los pacientes.
- **Disponibilidad de variables *target* con varios estados.** En este caso, y fundamentalmente con el objetivo de realizar el estudio de expresión diferencial y, posteriormente, el desarrollo de modelos predictivos de clasificación de casos, se procuró que existieran varias categorías en

ciertos factores clave (localización tisular de tumores y coexistencia de muestras control y muestras tumorales en el mismo *dataset*).

Todos estos criterios se cumplían en el caso del *dataset* escogido, el proyecto de cáncer de esófago (TCGA-ESCA, *dbGaP Study Accession* phs000178). La **Figura 5** muestra la información referida a dicho proyecto. Se puede observar cómo existe información de ~185 casos, para diferentes tecnologías multi-ómicas (lecturas de secuenciación, perfil de transcriptómica, datos de *Single Nucleotide Polymorphisms,* SNPs, datos de *Copy Number Variation,* CNVs, de metilación de ADN, así como información clínica). Por tanto, tenemos un buen espectro de datos para poder explotar.

Figura 5. Detalles acerca del proyecto seleccionado del repositorio TCGA, incluyendo el número de individuos para cada categoría de datos y el formato o estrategias experimentales utilizadas. Fuente: https://portal.gdc.cancer.gov/projects/TCGA-ESCA.

Por otro lado, si observamos la información por categoría de la enfermedad, podemos comprobar cómo existen dos localizaciones en las que se suceden los tumores (sitios primarios): i) esófago y ii) estómago. Esta diferenciación va a poder ser explotada también en términos de expresión diferencial de genes (ver **Figura 6**).

Primary Site	Disease Type	Cases	Available Cases per Data Category							Files
			Seq	Exp	SNV	CNV	Meth	Clinical	Bio	
Esophagus	• adenomas and adenocarcinomas • cystic, mucinous and serous neoplasms • squamous cell neoplasms ▲ collapse	183	183	182	182	183	183	183	183	5596
Stomach	• adenomas and adenocarcinomas • squamous cell neoplasms ▲ collapse	2	2	2	2	2	2	2	2	88

Figura 6. Información acerca de los sitios primarios tumorales que comprende el *dataset* TCGA-ESCA y el número de casos de cada uno.

Este *dataset* contiene información acerca de cáncer de esófago, clasificado por histología como adenocarcinoma o carcinoma celular escamoso [14]

2.1.2. Descarga y tratamiento de los datos multi-ómicos y clínicos

Acceso a la información desde repositorios públicos

La información depositada en TCGA puede ser accedida a través del portal *Genomic Data Commons* (GDC), y está organizada de una manera muy estandarizada. Cada muestra depositada en TCGA se encuentra definida por un código de barras (*barcode*) que indica información referente al proyecto, tipo de muestra, tejido, etc., lo que permite que las búsquedas y la extracción de estas muestras con diferentes propósitos pueda ser realizada de manera automática con las herramientas bioinformáticas adecuadas (ver **Figura 7**). El tipo de muestra es un parámetro muy relevante. Está representada por números: de 01-09 (tipos de tumores), 10-19 (muestras normales) y de 20-29 (muestras controles).

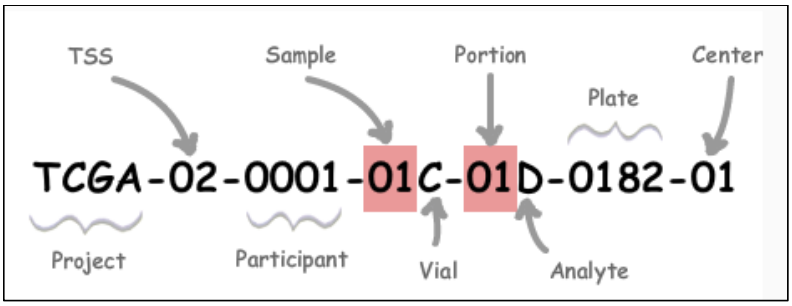

Figura 7. Ejemplo de un *barcode* del repositorio TCGA y sus componentes. Fuente: https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/.

Además, el portal recoge diferentes categorías de datos ómicos y clínicos para cada proyecto, que pueden ser filtrados a la hora de escoger los datos que se desean para un determinado propósito. Estas categorías también vienen reflejadas en una serie de estrategias experimentales que del mismo modo se pueden filtrar (ver **Figura 8**).

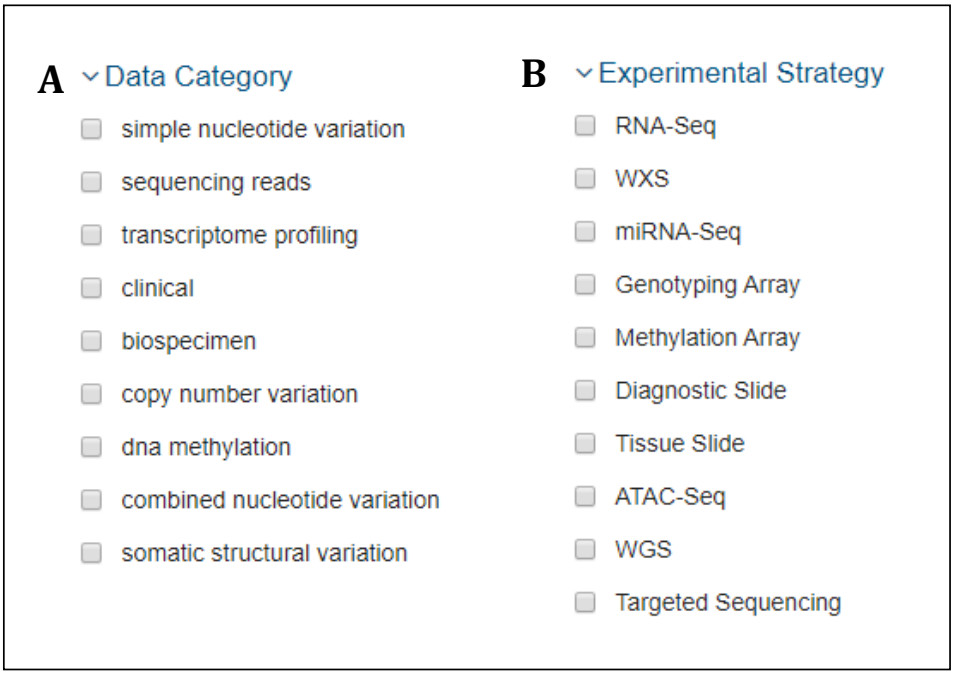

Figura 8. Tipo de información disponible en los proyectos accesibles a través del GDC portal, en cuanto a **A)** categoría de datos y **B)** estrategias experimentales.

Para poder realizar un acceso eficiente de la información contenida en el proyecto TCGA-ESCA, se ha empleado un paquete de R, dentro del entorno *Bioconductor* que permite un análisis y selección del proyecto TCGA de interés. Este paquete es *TCGAbiolinks* [15]. Es interesante indicar que el entorno *Bioconductor* proporciona un set de datos específico que se denomina *SummarizedExperiment*. Este fichero contiene una gran cantidad de información, entre la que se encuentran metadatos (datos clínicos y demográficos básicos), datos ómicos (por ejemplo, valores de metilación o de conteos de transcritos) y metadatos de las capas ómicas (por ejemplo, cromosoma, posiciones de comienzo y fin de genes, símbolos de genes, etc.). Toda la información puede ser accedida mediante el paquete *SummarizedExperiment.*

- Para acceder a la matriz de las muestras (metadatos), se emplea la función *colData*.
- Los datos moleculares pueden consultarse mediante la función *assays*.

- Los metadatos ómicos (la matriz de genes) se pueden consultar gracias a *rowRanges.*

La **Figura 9** muestra la estructura del objeto *SummarizedExperiment*.

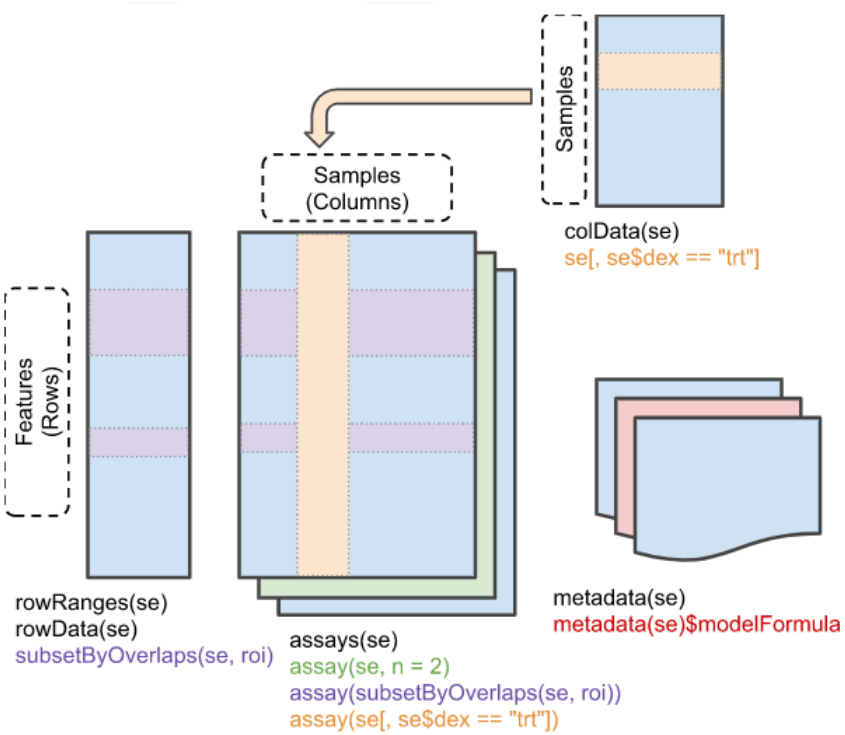

Figura 9. Estructura de un objeto *SummarizedExperiment*. Fuente: http://bioconductor.org/packages/SummarizedExperiment/.

Tratamiento exploratorio de los datos ómicos y clínicos

Hemos realizado un estudio cuantitativo de las capas ómicas en el entorno del paquete *TCGAbiolinks*, que funciona como una *suite* muy completa, incluyendo diferentes funciones recientemente desarrolladas [16], que permiten el análisis exploratorio de los datos y la obtención de información clave como el análisis de expresión diferencial (*Differential Expression Analysis, DEA*) (ver **Figura 10**). Para consultar un tutorial muy completo acerca del procesamiento de datos ómicos (análisis de expresión, metilación, etc.), consultar el siguiente *link*, que ha servido de base para el desarrollo de los análisis del presente trabajo:
http://bioinformaticsfmrp.github.io/TCGAWorkflow/

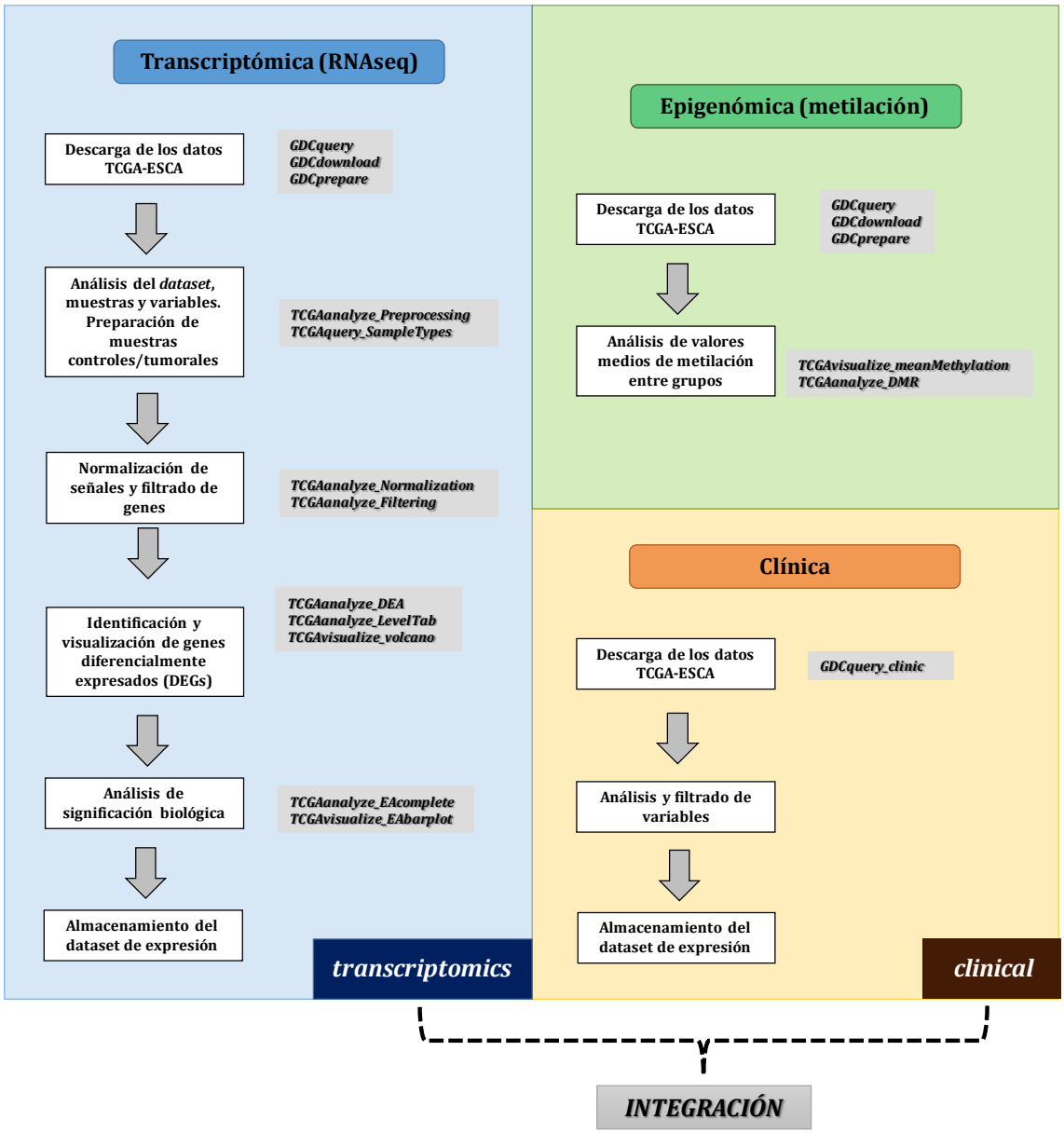

Figura 10. *Pipeline* para la descarga y el análisis exploratorio de los *datasets* utilizados en el presente estudio.

Mediante este flujo de trabajo, hemos descargado la información biológica y clínica, así como desarrollado sucesivos análisis. El análisis DEA se realizó enfrentando a los grupos control (*Solid Tissue Normal*) frente a las muestras tumorales (*Primary solid Tumor*). Gracias a esta aproximación conseguiremos filtrar los genes para que los requisitos computacionales posteriores sean más reducidos a la hora de ajustar el modelo.

2.1.3. Métodos y modelos estadísticos utilizados

En el presente trabajo se van a realizar dos aproximaciones para el tratamiento de datos.

En primer lugar, respecto al método de selección de variables (*feature selection*) se va a valorar el uso de la regresión penalizada. Nuestro objetivo es reducir la dimensionalidad de los datos. El modelo lineal estándar funciona deficientemente cuanto se posee un gran *dataset* multivariante con un número de variables superior al número de muestras. Una alternativa más adecuada es la regresión penalizada, que permite crear un modelo lineal de regresión con una penalización por tener tantas variables mediante la adición de un parámetro de restricción en la ecuación. También se denominan métodos de *shrinkage* (contracción) o regularización. La consecuencia de esta penalización es reducir los valores de los coeficientes hacia 0, permitiendo que haya menos variables con un coeficiente cerca de 0 o 0. Los métodos requieren la selección de un parámetro de afinación (*tuning*) que determine la cantidad de contracción. Los más frecuentes son *ridge regression, Least Absolute Shrinkage and Selection Operator* (LASSO) y *Elastic Net* (ENET). Para valorar el ajuste o rendimiento de cada modelo a los datos, se utiliza el parámetro RMSE (*root mean square error*) o la suma de los cuadrados de las diferencias entre los valores actuales y predichos de la variable dependiente [17].

Por otro lado, el proceso de integración de las capas ómicas con la clínica se va a realizar mediante dos métodos con bases estadísticas diferenciales, que son implementados en paquetes de R. En primer lugar, se empleará el paquete SNFtools, que se basa en la aproximación *Similarity Network Fusion* (SNF), metodología que radica en el estudio de *networks* o similaridad/redes. Este método se basa en construir redes de muestras (pacientes) para cada tipo de datos disponibles, y fusionar estas redes en una única que represente todo el espectro de datos [18].

Por último, emplearemos el paquete *mixOmics* [19], dedicado al análisis multivariante de *datasets* biológicos con una especial atención a la exploración de datos, reducción de dimensiones mediante selección de variables y visualización. Este método permite incorporar un amplio espectro de capas ómicas.

2.3. Resultados

2.3.1. Análisis del *dataset* transcriptómico (RNAseq)

Comenzamos nuestro estudio con el análisis del *dataset* de expresión génica (transcriptómica) del proyecto TCGA-ESCA.

Preparación del entorno, descarga de los datos y análisis exploratorio

En primer lugar, cargaremos los paquetes necesarios para el estudio.

```
library(TCGAbiolinks)
library(SummarizedExperiment)
library(dplyr)
library(tidyverse)
library(caret)
library(knitr)
library(glmnet)
library(ROCR)
library(edgeR)
library(ggplot2)
library(cowplot)
library(SNFtools)
library(rgl)
library(mixOmics)
library(dummies)
```

Ahora mostraremos información detallada acerca de las características y los datos que contiene el proyecto, en cuanto a número de archivos y de muestras en cada una de sus aproximaciones experimentales mediante el entorno *TCGAbiolinks*.

```
TCGAbiolinks:::getProjectSummary("TCGA-ESCA")
```

```
## $file_count
## [1] 5657
##
## $case_count
## [1] 185
##
## $data_categories
##   file_count case_count               data_category
## 1        919        184     Transcriptome Profiling
## 2       1486        184 Simple Nucleotide Variation
## 3        962        185                  Biospecimen
## 4        207        185                     Clinical
## 5        202        185             DNA Methylation
## 6       1115        185       Copy Number Variation
## 7        766        185            Sequencing Reads
##
```

```
## $file_size
## [1] 8.198261e+12
```

A continuación, procederemos a preparar la descarga de los datos.

```
query_ESCA <- GDCquery(project = "TCGA-ESCA",
                       data.category = "Gene expression",
                       data.type = "Gene expression quantification",
                       experimental.strategy = "RNA-Seq",
                       platform = "Illumina HiSeq",
                       file.type = "results",
                       legacy = TRUE)
```

Una visualización global del proyecto nos permite observar las diferentes variables que comprende el *dataset*.

```
esca <- getResults(query_ESCA)

colnames(esca)
```

```
##  [1] "file_name"            "type"                 "state_comment"
##  [4] "updated_datetime"     "cases"                "data_category"
##  [7] "id"                   "file_id"              "experimental_strategy"
## [10] "access"               "data_release"         "data_type"
## [13] "submitter_id"         "tags"                 "md5sum"
## [16] "state"                "file_state"           "file_size"
## [19] "version"              "data_format"          "platform"
## [22] "project"              "short_name"           "center_center_type"
## [25] "center_name"          "center_code"          "center_center_id"
## [28] "center_namespace"     "tissue.definition"
```

Ahora mostraremos la información de las muestras definidas por una variable clave, que es el tipo de tejido. Como podemos observar aquí, hay una categoría (*Metastatic*) que engloba solo una observación. Por tanto, evitaremos descargar esta categoría.

```
table(esca$tissue.definition)
```

```
##
##                          Additional - New Primary
##                                                 0
##                             Additional Metastatic
##                                                 0
##                              Blood Derived Normal
##                                                 0
##                                Bone Marrow Normal
##                                                 0
##                                Buccal Cell Normal
##                                                 0
##                Cell Line Derived Xenograft Tissue
##                                                 0
##                                       Cell Lines
##                                                 0
##                                   Control Analyte
##                                                 0
##                            EBV Immortalized Normal
##                                                 0
```

```
##                       Human Tumor Original Cells
##                                                0
##                                      Metastatic
##                                                1
##        Primary Blood Derived Cancer - Bone Marrow
##                                                0
##   Primary Blood Derived Cancer - Peripheral Blood
##                                                0
##                              Primary solid Tumor
##                                              184
##                          Primary Xenograft Tissue
##                                                0
##      Recurrent Blood Derived Cancer - Bone Marrow
##                                                0
## Recurrent Blood Derived Cancer - Peripheral Blood
##                                                0
##                            Recurrent Solid Tumor
##                                                0
##                              Solid Tissue Normal
##                                               11
```

Vamos a seguir preparando tanto la descarga de datos como los análisis *downstream*.

```
# Preparación de la descarga de los datos con las muestras filtradas
query_final <- GDCquery(project = "TCGA-ESCA",
                    data.category = "Gene expression",
                    data.type = "Gene expression quantification",
                    experimental.strategy = "RNA-Seq",
                    platform = "Illumina HiSeq",
                    file.type = "results",
                    legacy = TRUE,
                    # Escogeremos solamente dos categorías para la descarga,
                    # muestras control ("NT") y tumorales ("TP")
                    sample.type = c("Primary solid Tumor", "Solid Tissue
Normal"))

#Descargamos los datos
GDCdownload(query_final)

# Cargamos el dataset a nuestro entorno y activamos la opción para guar
darlo
esca.expr <- GDCprepare(query_final, save = TRUE, save.filename = "esca
Expr.rda")

# Obtenemos la información de los subtipos
dataSubt <- TCGAquery_subtype(tumor = "ESCA")

# Obtenemos la información clínica
dataClin <- GDCquery_clinic(project = "TCGA-ESCA", "clinical")

# Realizamos el subset de las muestras tumorales "TP" ("Primary Solid T
umor")
dataSmTP <- TCGAquery_SampleTypes(getResults(query_final, cols = "cases
"), "TP")

# Realizamos el subset de las muestras normales o controles "NT" ("Soli
d Tissue Normal)
```

```
dataSmNT <- TCGAquery_SampleTypes(getResults(query_final, cols = "cases
"), "NT")

class(esca.expr)

## [1] "RangedSummarizedExperiment"
## attr(,"package")
## [1] "SummarizedExperiment"
```

El tipo de objeto es *SummarizedExperiment*, objeto complejo que comprende varias capas de información, entre las que se encuentran los metadatos, la información de los conteos y los nombres de los genes y de las muestras.

```
head(esca.expr)

## class: RangedSummarizedExperiment
## dim: 6 195
## metadata(1): data_release
## assays(2): raw_count scaled_estimate
## rownames(6): A1BG|1 A2M|2 ... SERPINA3|12 AADAC|13
## rowData names(4): gene_id entrezgene ensembl_gene_id
##   transcript_id.transcript_id_TCGA-KH-A6WC-01A-11R-A336-31
## colnames(195): TCGA-KH-A6WC-01A-11R-A336-31
##   TCGA-LN-A49M-01A-21R-A260-31 ... TCGA-R6-A6KZ-01A-11R-A31P-31
##   TCGA-L5-A8NQ-01A-11R-A36D-31
## colData names(144): sample patient ...
##   subtype_GEA.CIN.Integrated.Cluster...MKL.KNN.4
##   subtype_GEA.CIN.Integrated.Cluster...MKL.KNN.7
```

Veremos las dimensiones del objeto, en concreto, de la matriz de los conteos génicos, que se encuentra en el *slot assays*, y nos revela que el *dataset* contiene 195 muestras y 19947 variables o genes listados.

```
dim(assay(esca.expr))

## [1] 19947   195
```

Procesamiento del* dataset *para análisis downstream mediante TCGAbiolinks

Tras este tratamiento inicial del *dataset* de expresión génica, vamos a retomar el *pipeline* de análisis y seguir preparando los archivos necesarios para correr el estudio de expresión diferencial en el entorno del paquete *TCGAbiolinks*. Desarrollaremos en primer lugar un pre-procesamiento de los datos (ver **Figura 11**), lo que nos permite localizar posibles *outliers*.

```
dataPrep <- TCGAanalyze_Preprocessing(object = esca.expr, cor.cut = 0.6
)
```

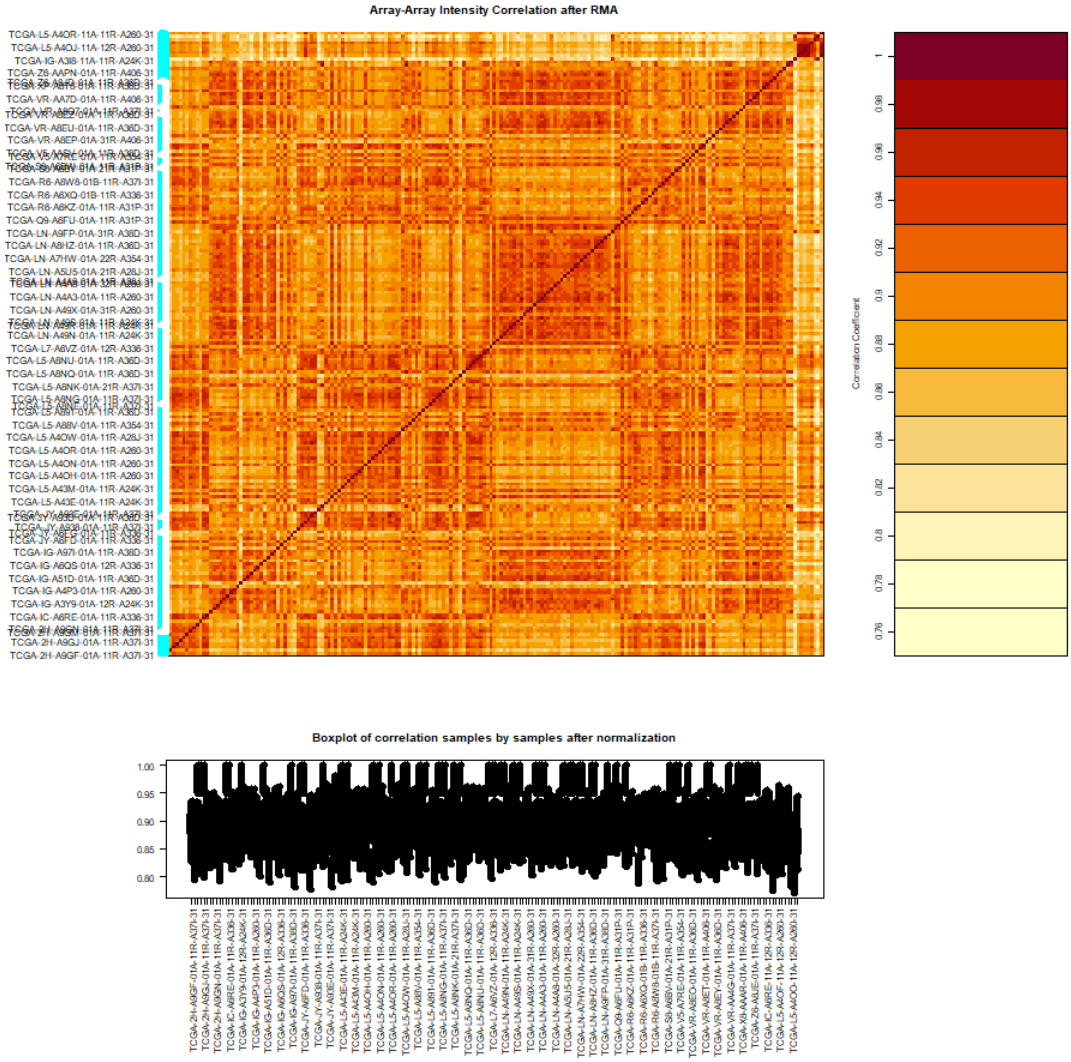

Figura 11. Matriz simétrica de correlación de Pearson entre las muestras (*Array Array Intensity correlation plot*) y *boxplot* de la correlación.

De acuerdo a esta matriz, no observamos muestras con baja correlación (cor.cut = 0.6, ver *boxplot* en la parte inferior) que puedan ser identificadas como posibles *outliers*, por lo que continuamos el análisis manteniendo todas ellas.

Por otro lado, el *boxplot* de las señales de intensidad brutas de los *array* nos muestra ciertas variaciones entre individuos que son las que cabría esperar (ver **Figura 12**).

```
boxplot(dataPrep, outline = FALSE, xaxt = "n")
```

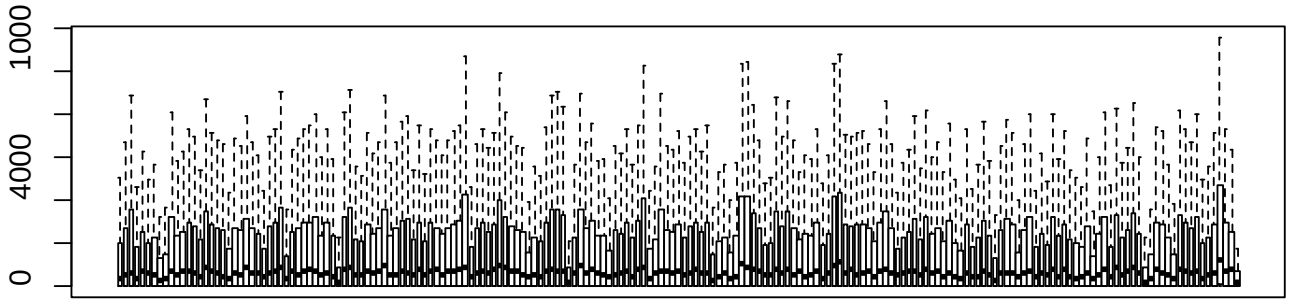

Figura 12. Representación de las señales de intensidad de los datos brutos del *dataset* de expresión génica.

Posteriormente, normalizaremos los conteos génicos, mediante el ajuste del contenido de GC (u otros efectos a nivel génico) en las lecturas, así como la normalización de cuantiles.

```
# Normalización de los genes
dataNorm <- TCGAanalyze_Normalization(tabDF = dataPrep,
                                      geneInfo = geneInfo,
                                      method = "gcContent")

boxplot(dataNorm, outline = FALSE, xaxt = "n")
```

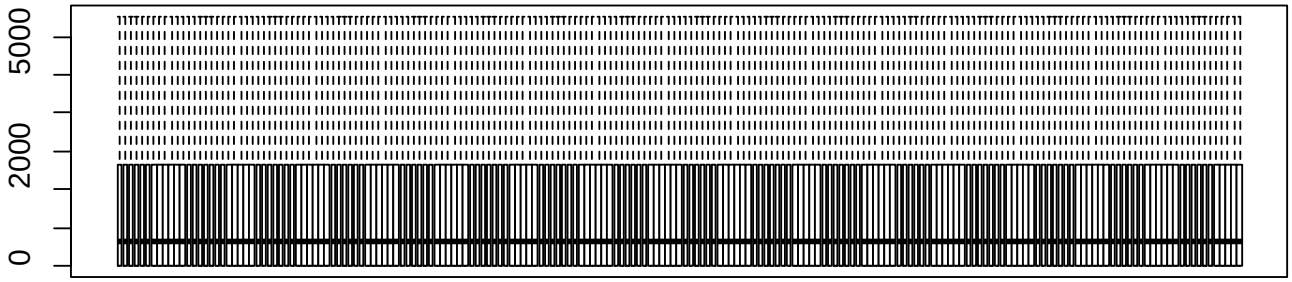

Figura 13. Representación de las señales de intensidad de los datos normalizados del *dataset* de expresión génica.

Aquí podemos observar que las señales de intensidad de los *array* entre las muestras para los datos normalizados se han igualado (ver **Figura 13**).

Después, eliminaremos aquellos genes que muestren una baja señal a través de las muestras.

```
# Filtrado por cuantiles de los genes
dataFilt <- TCGAanalyze_Filtering(tabDF = dataNorm,
                                  method = "quantile",
                                  qnt.cut =  0.25)
```

Podemos comprobar el número de genes que, tras este procesamiento, de los 19.947 que teníamos inicialmente, nos quedamos con un primer *subset* de 14.899.

```
nrow(dataFilt)

## [1] 14899
```

Análisis de expresión diferencial (DE)

A continuación, desarrollaremos un análisis de expresión diferencial entre las muestras tumorales y normales. Nuestro principal objetivo será, además de mostrar aquellos genes que son responsables de la manifestación de la enfermedad, realizar un notable filtrado del conjunto de genes que contiene nuestro *dataset* (19.947) para poder implementar *a posteriori* la integración de los datos ómicos y clínicos y el desarrollo de los diferentes modelos de predicción (regresión) con mayor eficacia computacional.

```
# Propondremos unos cut.offs restrictivos para lograr un número manejable de g
enes
dataDEGs <- TCGAanalyze_DEA(mat1 = dataFilt[,dataSmNT],
                            mat2 = dataFilt[,dataSmTP],
                            Cond1type = "Normal",
                            Cond2type = "Tumor",
                            fdr.cut = 0.001 ,
                            logFC.cut = 2,
                            method = "glmLRT")
```

Ahora, filtramos el output de *dataDEGs* por el valor absoluto de LogFC>=2 y usamos la función *TCGAanalyze_LevelTab* para crear una tabla que muestre los DEGs (*differentially expressed genes*), *log fold change*, *false discovery rate* (FDR), el valor de expresión génica para las muestras normales y tumorales y el valor delta (la diferencia entre los valores de expresión génica entre las dos condiciones multiplicado por logFC). Esta tabla se encuentra ordenada por este valor delta.

```
# Mostramos la tabla de DEG con los valores de expresión en muestras normales
y tumorales
dataDEGsFiltLevel <- TCGAanalyze_LevelTab(dataDEGs,"Tumor","Normal",
dataFilt[,dataSmTP],dataFilt[,dataSmNT])
head(dataDEGsFiltLevel)

##          mRNA     logFC          FDR     Tumor       Normal     Delta
## COL1A1 COL1A1  2.984585 9.611930e-05 458118.79 57830.272727 1367294.6
## JUP       JUP  2.066963 2.671430e-04 199053.54 47494.454545  411436.4
## LCN2     LCN2  3.749081 6.230298e-04  85295.20  6342.727273  319778.6
## S100A7 S100A7  6.418305 3.961789e-04  31477.01   367.818182  202029.0
## MUC2     MUC2 10.651755 1.437988e-04  12639.84     7.727273  134636.5
## LAMC2   LAMC2  2.805883 9.037585e-06  37423.93  5349.727273  105007.2

nrow(dataDEGsFiltLevel)
```

```
## [1] 995
```

Hemos identificado 995 genes diferencialmente expresados entre las 11 muestras normales y las 184 tumorales, con un *log fold change* >=2 y un FDR <0.1%.

Podemos exportar los genes seleccionados en forma de lista para posteriores pasos:

```
genesDEA <- rownames(dataDEGs)
```

Vamos a representar los resultados del análisis de DE mediante un *Volcano Plot* (ver **Figura 14**). Dada la complejidad de la representación (debido al gran número de DEGs), la **Figura 15** muestra los 10 top-genes que se encuentran regulados al alza o a la baja.

```
TCGAVisualize_volcano(x = dataDEGs$logFC,
                      y = dataDEGs$FDR,
                      filename = "Volcano.png",
                      x.cut = 3,
                      y.cut = 10^-5,
                      names = rownames(dataDEGs),
                      color = c("black","red","darkgreen"),
                      names.size = 2,
                      xlab = " Gene expression fold change (Log2)",
                      legend = "State",
                      title = "Volcano plot (Normal vs Tumor)",
                      width = 10)
```

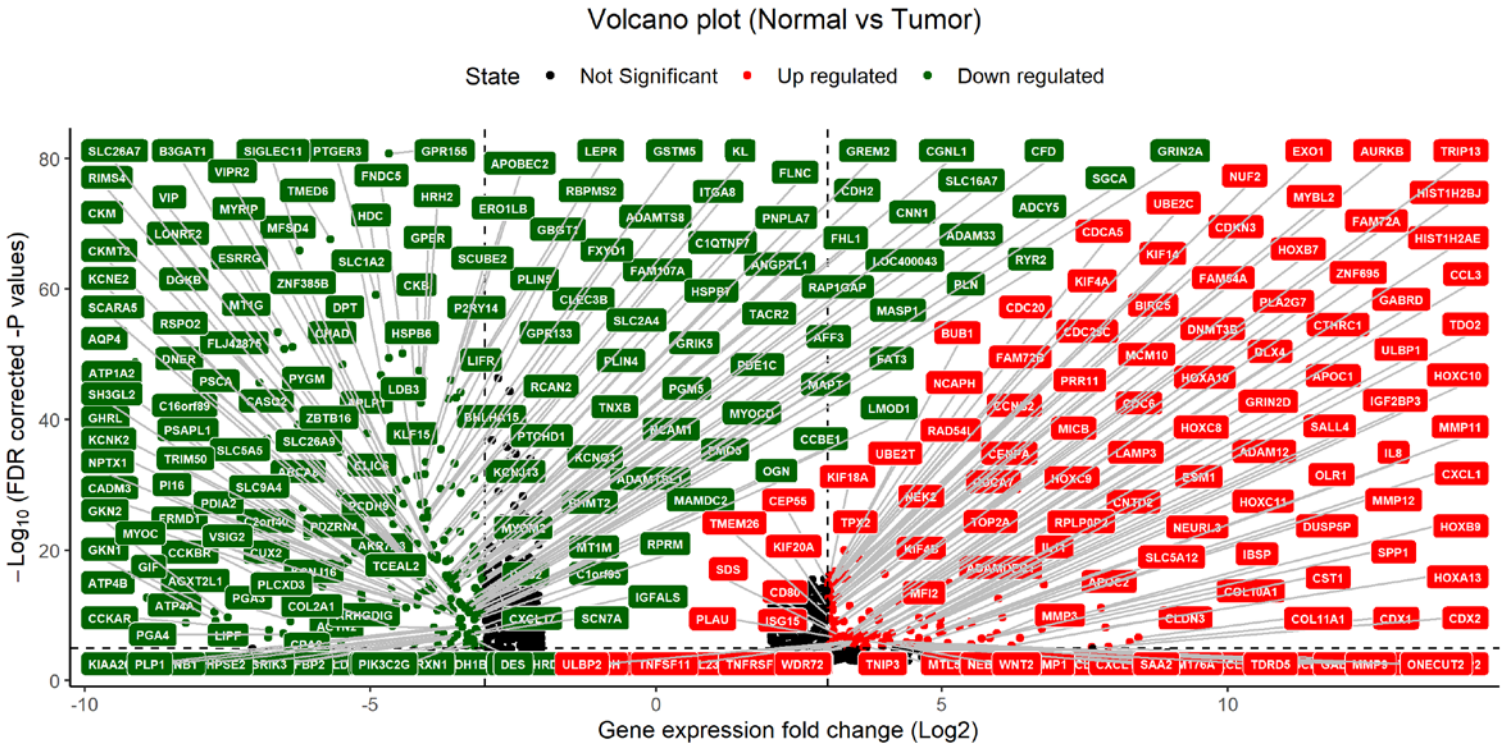

Figura 14. *Volcano plot* para los resultados de expresión génica diferencial entre muestras normales y tumorales para el *dataset* TCGA-ESCA.

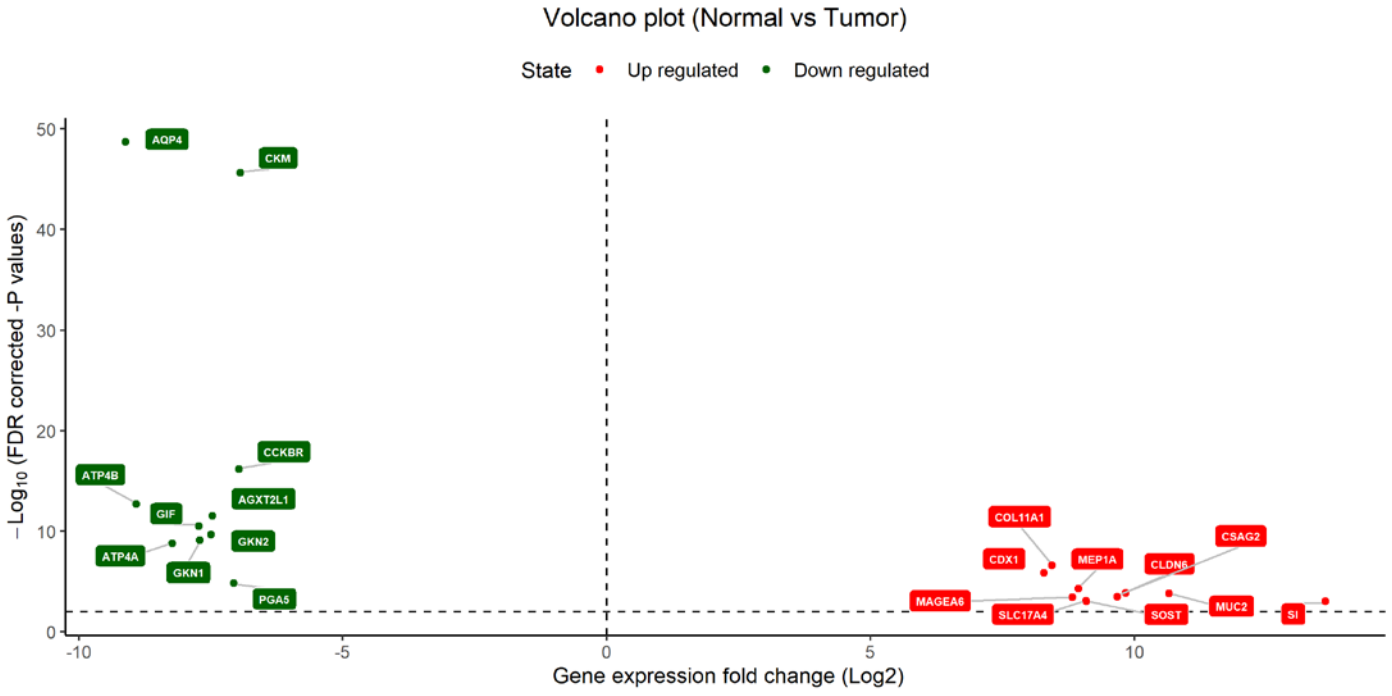

Figura 15. *Volcano plot* para los resultados de expresión génica diferencial entre muestras normales y tumorales para el *dataset* TCGA-ESCA con respecto a los 20 top-genes.

Estos *plots* plasman los nombres de los genes que se encuentran altamente expresados (o sobre-regulados, en rojo) o bajamente expresados (infra-regulados, en verde) en las muestras tumorales con respecto a las normales.

Análisis de enriquecimiento o significación biológica

Para comprender los procesos biológicos subyacentes a los genes que se han mostrado diferencialmente expresados, realizaremos un análisis de enriquecimiento. Vamos a visualizar los genes que han resultado candidatos del proceso de DE en términos de significación biológica respecto a *Gene Ontology*.

```
ansEA <- TCGAanalyze_EAcomplete(TFname="DEA genes Normal Vs Tumor", Reg
ulonList = rownames(dataDEGs))

# En este paso, vamos a realizar la representación mediante barplots
TCGAvisualize_EAbarplot(tf = rownames(ansEA$ResBP),
                        GOBPTab = ansEA$ResBP,
                        GOCCTab = ansEA$ResCC,
                        GOMFTab = ansEA$ResMF,
                        PathTab = ansEA$ResPat,
                        nRGTab = rownames(dataDEGs),
                        nBar = 10)
```

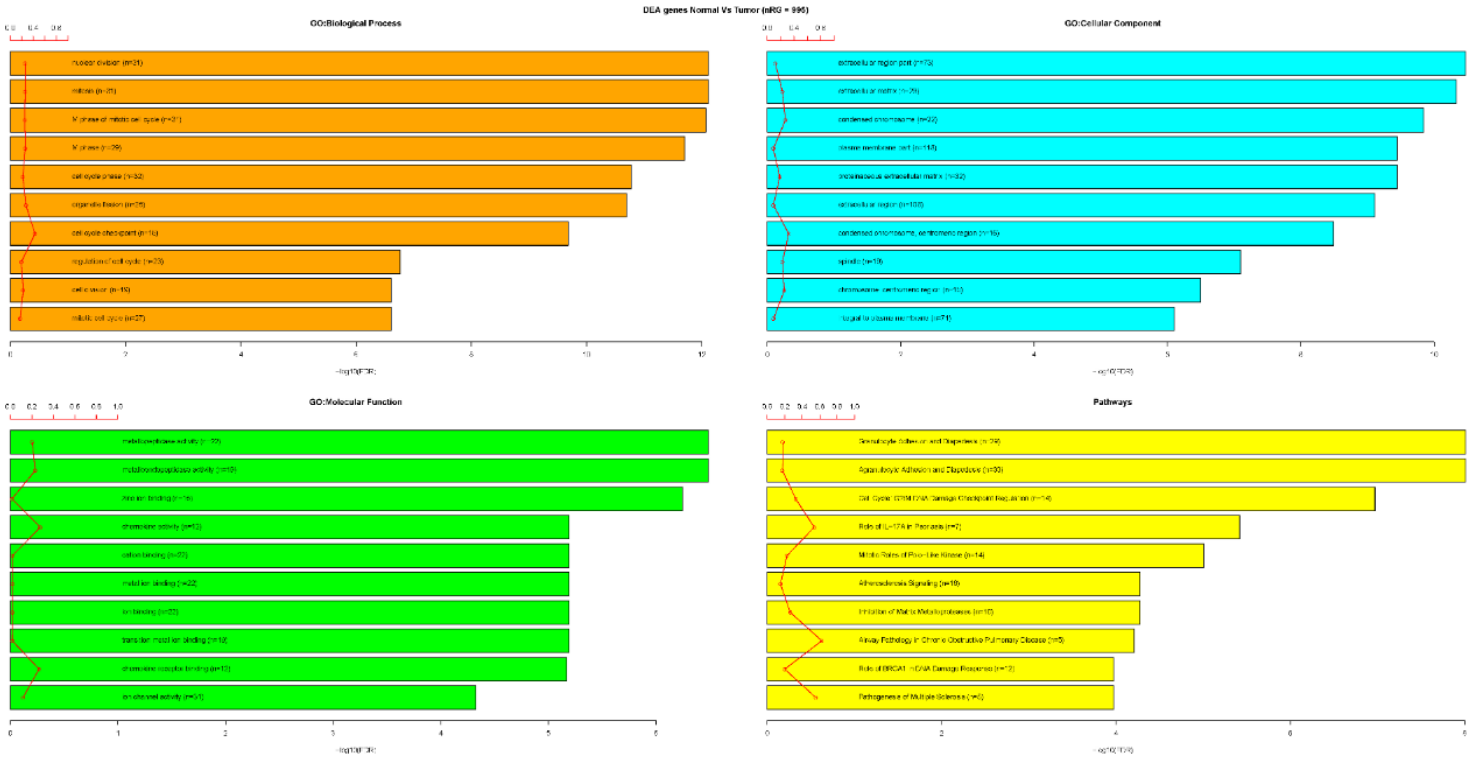

Figura 16. Vías canónicas significativamente sobre-representadas o enriquecidas por los genes diferencialmente expresados (DEGs).

La **Figura 16** representa la información requerida, mostrando las vías canónicas que significativamente están más representadas o enriquecidas por los genes DEGs (diferencialmente expresados). Las vías están listadas de acuerdo al p-valor corregido FDR (-log) y la ratio de genes encontrados en cada vía entre el número total de genes en esa vía (línea roja).

En cuanto a proceso biológico (*GO Biological Process*), los genes DE están relacionados básicamente con procesos de división nuclear y celular, así como otros mecanismos de regulación del ciclo celular y de desarrollo de orgánulos. Con respecto a componente celular (*GO Cellular Component*), se encuentran principalmente en la región plasmática de la membrana (n=118) y extracelularmente (n=108). Por otro lado, la función biológica (*GO Molecular Function*) de los genes se relaciona con actividad enzimática y de receptores y canales iónicos. Por último, en amarillo se resaltan las vías (*Pathways*) que están involucradas, entre las que podemos observar adhesiones celulares y diferentes cascadas de señalización.

Preparación del dataset ómico a partir de los genes diferencialmente expresados

Como hemos dicho, uno de los objetivos de realizar en análisis de DE es la reducción selectiva del número de variables (genes) que van a ser considerados para los análisis *downstream* de integración de capas. Comenzaremos por tanto por extraer los datos de expresión génica del objeto *dataFilt* exclusivamente para los genes que han sido seleccionados por expresión diferencial (contenidos en el vector *genesDEA*).

```
dataFiltDEA <- dataFilt[row.names(dataFilt) %in% genesDEA,]
# Comprobamos si se muestra el mismo número de genes diferencialmente e
xpresados en nuestra selección
dim(dataFiltDEA)

## [1] 995 195
```

Finalmente, reorganizaremos el *dataset* para mostrar en filas las muestras y en columnas los genes, como se requiere en las posteriores aproximaciones:

```
transcriptomics <- t(dataFiltDEA)
```

Ahora debemos ser cuidadosos a la hora de mostrar el *barcode* de las muestras, ya que en este *dataset* aparece de manera completa, mientras que en el *dataset* clínico está abreviado (eliminándose información de placa, centro, etc.). Por tanto, debemos

eliminar de nuestra matriz de transcriptómica la información no contemplada en la clínica (los últimos 4 sets numéricos o 16 caracteres).

```
row.names(transcriptomics) <- substr(rownames(transcriptomics), 1, 12)
# Sustituiremos caracteres para eliminar los guiones
row.names(transcriptomics) <- gsub('-', '.', row.names(transcriptomics)
)
# Ahora ordenamos por el nombre de la muestra
transcriptomics <- transcriptomics[order(row.names(transcriptomics)),]
```

2.3.2. Análisis del *dataset* epigenómico

Además de la capa ómica de expresión génica, vamos a analizar datos epigenómicos mediante el estudio de la información de metilación génica. Para ello, comenzamos con la preparación de la descarga de los datos. Del mismo modo que para la transcriptómica, seleccionaremos exclusivamente las muestras "NT" y "TP". Obtendremos información de 201 muestras y 485577 sondas.

```
query.met <- GDCquery(project = "TCGA-ESCA",
                      legacy = TRUE,
                      data.category = "DNA methylation",
                      platform = "Illumina Human Methylation 450",
                      sample.type = c("Primary solid Tumor", "Solid Tis
sue Normal"))

# Descargamos los datos
GDCdownload(query.met)

# Cargamos el dataset a nuestro entorno
esca.met <- GDCprepare(query = query.met,
                      save = TRUE,
                      save.filename = "escaDNAmet.rda",
                      summarizedExperiment = TRUE)

dim(esca.met)

## [1] 485577    201
```

Análisis de metilación diferencial

Desarrollaremos, también en paralelo a lo implementado para el *dataset* de transcriptómica, un estudio comparativo de los niveles medios de metilación entre los dos grupos, control y tumoral.

```
TCGAvisualize_meanMethylation(esca.met,"definition")
```

```
=================== DATA Summary ====================
```

```
               groups      Mean    Median       Max
1 Primary solid Tumor 0.4469903 0.4526487 0.5507217
2 Solid Tissue Normal 0.4427186 0.4453486 0.4646647
        Min
1 0.3072065
2 0.3935468
==================== END DATA Summary ====================
==================== T test results ====================
                     Primary solid Tumor
Primary solid Tumor                   NA
Solid Tissue Normal            0.4244046
                     Solid Tissue Normal
Primary solid Tumor            0.4244046
Solid Tissue Normal                   NA
```

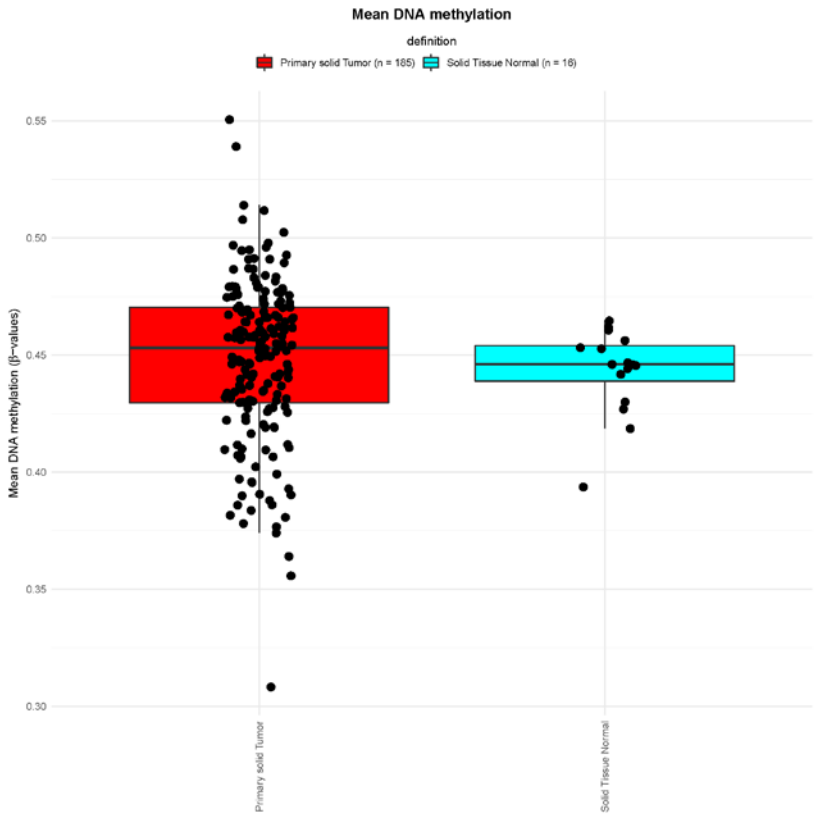

Figura 17. Niveles medios de metilación en muestras tumorales (rojo) y normales (azul).

Los resultados nos indican que no existen diferencias significativas en los datos de metilación entre los dos grupos que hemos definido (ver **Figura 17**). Podemos ver también los valores medios, máximos y mínimos de los beta-valores de metilación (que oscilan entre 0 y 1). Por tanto, no se observa un proceso de metilación diferencial de los genes entre ambas categorías.

Podemos visualizar esta información también mediante un *Volcano Plot*.

```
# Eliminamos los missing data
esca.met2 <- subset(esca.met,subset = (rowSums(is.na(assay(esca.met)))
== 0))

# Volcano plot
```

```
esca.met2 <- TCGAanalyze_DMR(esca.met2, groupCol = "definition",
                             group1 = "Solid Tissue Normal",
                             group2="Primary solid Tumor",
                             p.cut = 10^-5,
                             diffmean.cut = 0.25,
                             legend = "State",
                             plot.filename = "Volcano plot Metilación.png
")
```

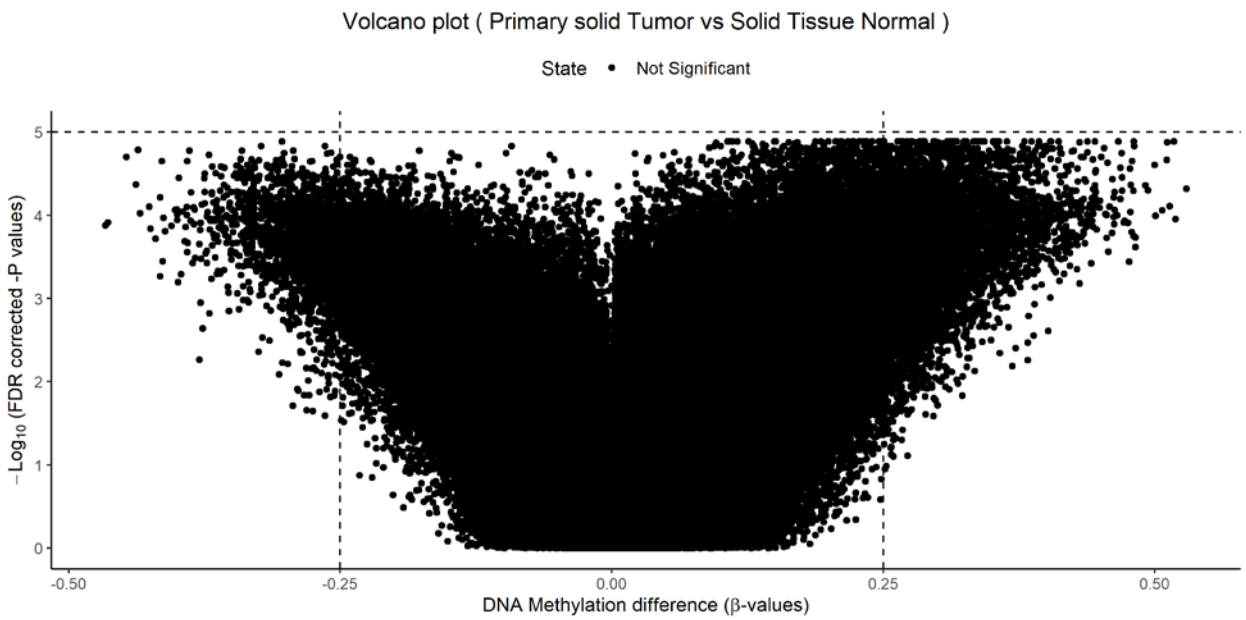

Figura 18. *Volcano plot* para los resultados de niveles de metilación diferencial entre muestras normales y tumorales para el *dataset* TCGA-ESCA.

Este *plot* vuelve a mostrar cómo no existe ningún gen que muestre metilación diferencial. A la vista de estos resultados, hemos decidido no utilizar el *dataset* de metilación para los posteriores análisis de integración de capas.

2.3.3. Análisis y tratamiento de los datos clínicos

Con anterioridad habíamos podido descargar los datos clínicos asociados a nuestro proyecto (objeto *dataClin*).

Vamos a realizar un análisis descriptivo de las principales variables que describen el *dataset* (ver **Figura 19**).

```
dataClinvarcont <- dataClin[,c("age_at_diagnosis", "bmi")]
# Transformaremos la variable edad en años (desde días) para una mejor
interpretación
dataClinvarcont$age_at_diagnosis <- dataClinvarcont$age_at_diagnosis/36
5

summary(dataClinvarcont)
```

```
##  age_at_diagnosis      bmi
##  Min.   :28.01    Min.   :13.40
##  1st Qu.:54.36    1st Qu.:21.30
##  Median :61.13    Median :24.26
##  Mean   :62.99    Mean   :25.31
##  3rd Qu.:72.19    3rd Qu.:27.67
##  Max.   :90.06    Max.   :60.24
##                   NA's   :10
```

```
age <- ggplot(dataClinvarcont, aes(x=age_at_diagnosis)) +
  geom_histogram(binwidth=3, fill="#69b3a2", color="#e9ecef")
bmi<- ggplot(dataClinvarcont, aes(x=bmi)) +
  geom_histogram(binwidth=3, fill="#b3697a", color="#e9ecef")
plot_grid(age, bmi,
          labels = c("A", "B"),
          ncol = 2, nrow = 1)
```

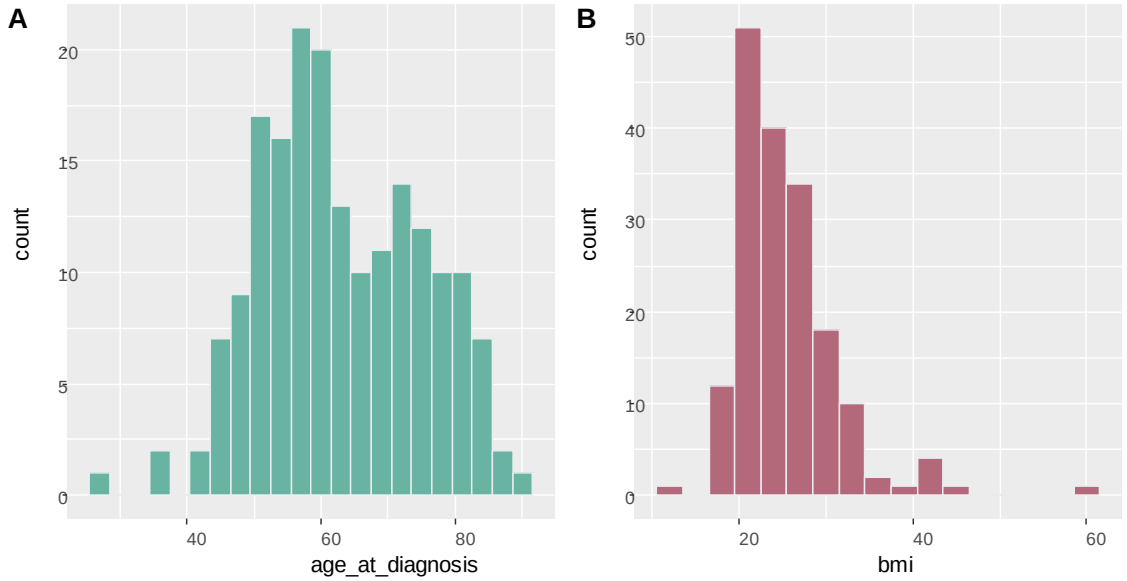

Figura 19. Histogramas que muestran la distribución de los datos de las variables edad al diagnóstico (**A**) y BMI – *Body Mass Index* (**B**).

Para obtener los datos estadísticos de las variables categóricas, en primer lugar, obtenemos tablas de porcentajes (ver **Tablas 2-5**), y, posteriormente, visualizaremos la información de manera gráfica (ver **Figura 20**).

```
kable(prop.table(table(dataClin$tissue_or_organ_of_origin)))
```

Tabla 2. Frecuencias relativas de las diferentes categorías de la variable "tejido".

	Freq
Cardia, NOS	0.0108108
Esophagus, NOS	0.0540541
Lower third of esophagus	0.6594595

Middle third of esophagus	0.2378378
Thoracic esophagus	0.0108108
Upper third of esophagus	0.0270270

```
kable(prop.table(table(dataClin$gender)))
```

Tabla 3. Frecuencias relativas de las diferentes categorías de la variable "sexo".

	Freq
female	0.1459459
male	0.8540541

```
kable(prop.table(table(dataClin$vital_status)))
```

Tabla 4. Frecuencias relativas de las diferentes categorías de la variable "estatus vital".

	Freq
Alive	0.5837838
Dead	0.4162162

```
kable(prop.table(table(dataClin$race)))
```

Tabla 5. Frecuencias relativas de las diferentes categorías de la variable "población".

	Freq
asian	0.2486486
black or african american	0.0270270
not reported	0.1081081
white	0.6162162

```
tej <- ggplot(dataClin, aes(x = `tissue_or_organ_of_origin`)) +
        geom_bar(fill = "#FF6666") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
gend <- ggplot(dataClin, aes(x = `gender`)) +
        geom_bar(fill = "#66B3FF") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
vital <- ggplot(dataClin, aes(x = `vital_status`)) +
        geom_bar(fill = "#FFFF66") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
race <- ggplot(dataClin, aes(x = `race`)) +
        geom_bar(fill = "#66FFB2") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
plot_grid(tej, gend, vital, race,
          labels = c("A", "B", "C", "D"),
          ncol = 2, nrow = 2)
```

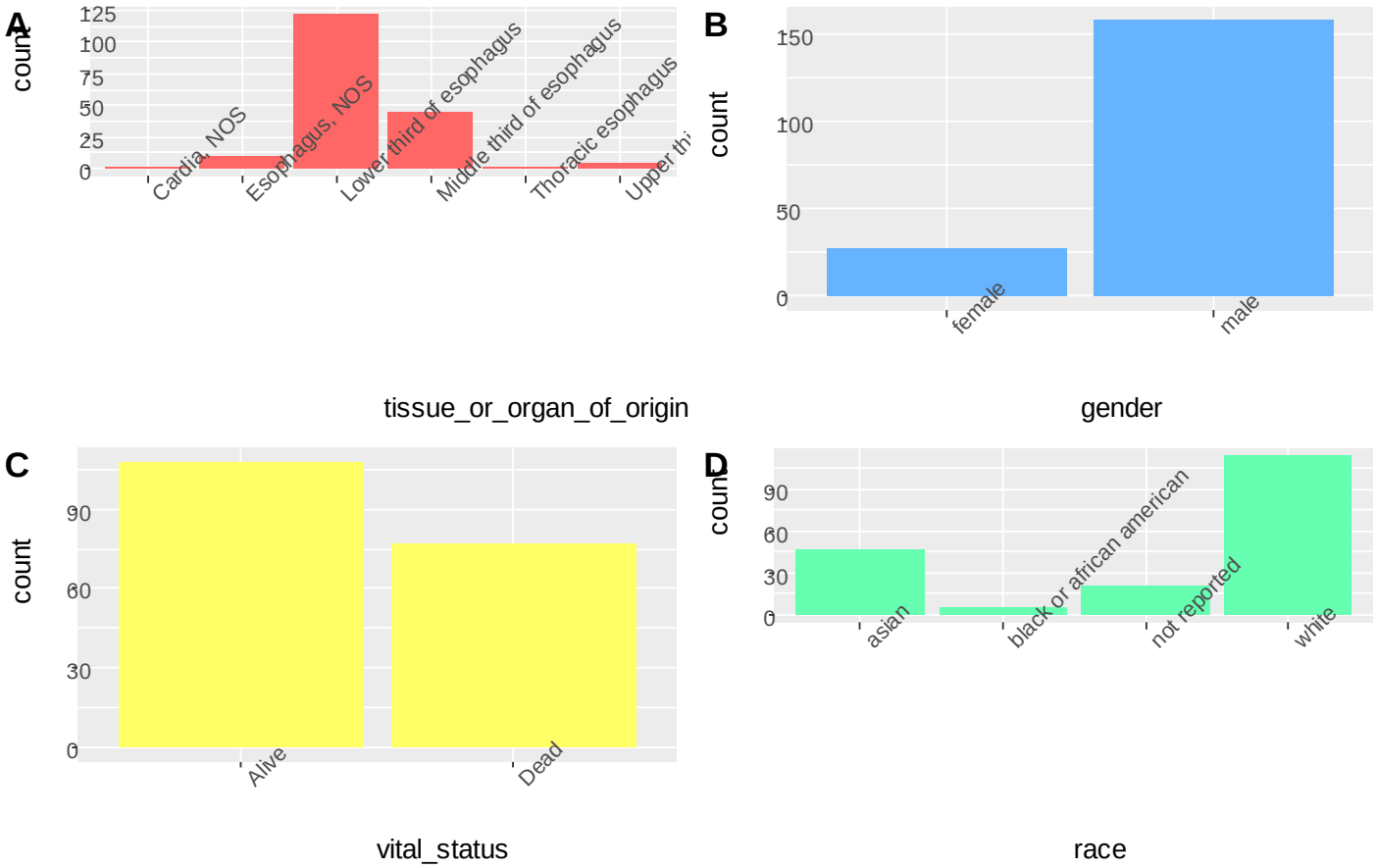

Figura 20. Gráficos de barras que representan las categorías de 4 variables seleccionadas de nuestro *dataset*: tejido (**A**), sexo (**B**), estatus vital o supervivencia (**C**) y población (**D**).

De las variables que contiene este *dataset,* vamos a escoger "*vital_status*" como variable objetivo o respuesta (*outcome*), a partir de la cual vamos a realizar nuestros modelos predictivos a través de diferentes variables predictoras (tanto ómicas como clínicas). Tras los análisis de DE basados en la comparación entre muestras tumorales y muestras normales, consideramos que esta variable respuesta puede ser una buena aproximación a la gravedad del cáncer ya que, obviando factores como por ejemplo la edad de los individuos u otros condicionantes, "*vital_status*" representa la supervivencia de los individuos frente a la enfermedad.

Análisis de supervivencia

Ya que vamos a escoger la variable "*vital_status*" como respuesta, podemos realizar un análisis de supervivencia en función de diferentes variables, gracias al entorno *TCGAbiolinks.*

```
# Variable sexo
TCGAanalyze_survival(dataClin,
                     "gender",
                     main = "TCGA Set\n ESCA",height = 10, width=10)

# Variable tejido
TCGAanalyze_survival(dataClin,
                     "tissue_or_organ_of_origin",
                     main = "TCGA Set\n ESCA",height = 10, width=10)
```

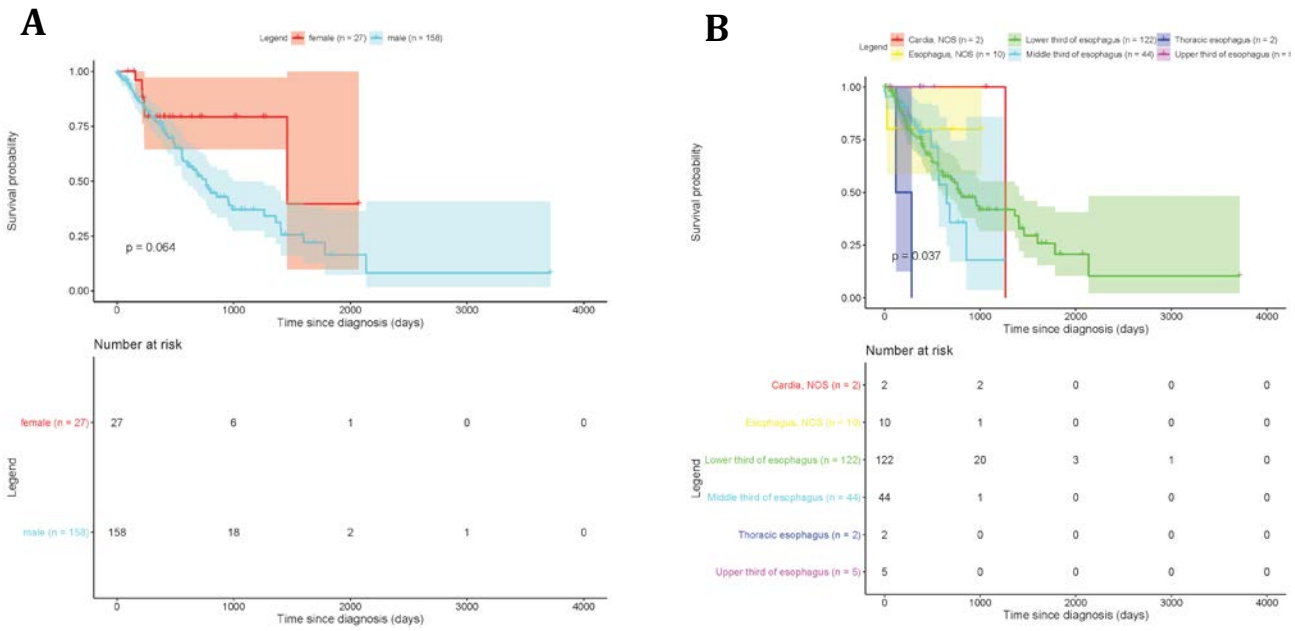

Figura 21. Análisis de supervivencia para las variables "sexo" (**A**) y "tejido" (**B**).

Podemos ver cómo no existen diferencias significativas entre los sexos para esta variable objetivo, aunque sí sería así en el caso del tipo de tejido (ver **Figura 21**).

Preparación del dataset *clínico para su posterior tratamiento conjunto con las ómicas*

Previamente a realizar la integración con la capa ómica pre-procesada, debemos asegurarnos de dos aspectos: *i)* de que los nombres de las muestras sean idénticos en ambos casos. Esto ya lo hemos abordado para el *dataset* de expresión; *ii)* de que todas las variables sean representativas. En el caso del *dataset* ómico, ya hemos realizado un filtrado inicial de dichas variables mediante análisis de DE. Ahora, estudiaremos las variables clínicas y eliminaremos las no procedentes o que no tengan información.

```
# Sustituiremos caracteres
dataClin$submitter_id <- gsub('-', '.', dataClin$submitter_id)
# Ahora convertiremos la primera columna del dataset clínico (nombre mu
```

```
estra) a row.name para lograr una correspondencia con el dataset ómico
dataClin2 <- data.frame(dataClin[,-1], row.names=dataClin[,1])
```

Ahora vamos a estudiar las variables para filtrar las no representativas.

```
# Eliminamos las variables que tienen missing data
dataClin2 <- dataClin2[ , colSums(is.na(dataClin2)) == 0]

# Descartaremos, por falta de información, aquellas variables que prese
ntan el mismo valor para todos los casos.

dataClin2 <- dataClin2[vapply(dataClin2, function(x) length(unique(x))
> 1, logical(1L))]

# Eliminaremos aquellas que representan información accesoria, que no s
erán a priori variables explicativas o bien aquellas que sean dependien
tes de otras

dataClin2 <- subset(dataClin2, select=-c(updated_datetime, diagnosis_id
, exposure_id, demographic_id, days_to_birth, treatments_pharmaceutical
_treatment_id, treatments_radiation_treatment_id, bcr_patient_barcode))
```

Finalmente, trabajaremos con la información de 185 pacientes y 18 variables clínicas. La naturaleza de las variables es diferente: binarias, categóricas y numéricas. Nuestra variable *outcome* o respuesta ("vital_status") tiene carácter binario.

2.3.4. Preparación de los *datasets* ómico y clínico para la integración y la implementación de modelos

Con carácter previo, habíamos llevado a cabo la homogeneización entre los nombres de las muestras contenidos en ambos *datasets*. Además, lógicamente, el número e identificador de cada muestra deben ser los mismos, por lo que extraeremos del *dataset* de transcriptómica las muestras que no aparecen en el clínico.

```
transcriptomics <- transcriptomics[-c(14, 16, 18, 52, 60, 62, 66, 68, 7
1, 74, 76),]
# También eliminaremos la muestra sobrante en el dataset clínico
dataClin2 <- dataClin2[rownames(dataClin2) %in% row.names(transcriptomi
cs),]
```

Por último, comprobaremos que todas las operaciones han sido correctas y que el número de filas es idéntico en ambos objetos, así como que los nombres son idénticos:

```
nrow(transcriptomics)

## [1] 184

nrow(dataClin2)
```

```
## [1] 184
```

```
all(row.names(transcriptomics) == row.names(dataClin2))
```

```
## [1] TRUE
```

Ahora ya podemos pasar a los diferentes abordajes que permiten combinar estas capas.

2.3.5. Selección de variables mediante regresión logística penalizada

En primer lugar, y teniendo en cuenta los requerimientos en cuanto a la naturaleza de las variables que poseen los métodos de regresión penalizada, vamos a transformar aquellas variables del *dataset* clínico que no sean numéricas primero a factor y luego a numéricas. Esto nos va a evitar problemas en el procesamiento y desarrollo de los modelos. El *dataset* de transcriptómica puede ser mantenido tal cual, ya que todas sus variables son cuantitativas.

```
clinical <- dataClin2 %>% mutate_if(is.character,as.factor)
clinical <- clinical %>% mutate_if(is.factor,as.integer)
```

A continuación, crearemos un objeto conjunto que incluya información tanto de expresión como clínica, formado por 184 muestras y 1013 variables.

```
join <- cbind(transcriptomics, clinical)
```

```
dim(join)
```

```
## [1]  184 1013
```

Vamos a realizar una partición aleatoria del *set* de datos con el fin de poder obtener con posterioridad los parámetros de rendimiento de los modelos.

```
# Fijaremos una semilla aleatoria para asegurar la reproducibilidad
set.seed(12345)
# Creamos la partición entre el set de entrenamiento (80%, para construir un modelo predictivo) y el de test (20%, para evaluar el modelo)
training.samples <- join$vital_status %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- join[training.samples, ]
test.data <- join[-training.samples, ]
```

Implementando la regresión penalizada

Ahora se requiere crear dos objetos, *i)* y, para almacenar la variable objetivo o resultado (en nuestro caso, "*vital_status*") y *ii)* x, que contiene las variables

predictoras. Este objeto debe ser creado con la función *model.matrix()*, la cual permite que cualquier variable que no sea cuantitativa sea automáticamente transformada a variable *dummy*, lo cual va a permitir la construcción de los modelos, ya que *glmnet()* solo admite variables numéricas.

```
# Variables predictoras
x <- model.matrix(vital_status~., train.data)[,-1]
# Variable respuesta
y <- train.data$vital_status
```

La función que vamos a escoger es *glmnet()* del paquete *glmnet* para computar estos modelos de regresión lineal penalizados, cuya fórmula es:

```
glmnet(x, y, alpha = X, lambda = NULL)
```

Siendo x, la matriz de variables predictoras, y, la variable respuesta, alpha, que es el parámetro de mezcla *elasticnet*, que toma valores de 1 (regresión *lasso*), 0 (regresión *ridge*) e intermedio para *elasticnet*, y lambda, un valor numérico que indica la cantidad de constricciones, que puede ser optimizado.

El mejor valor de lambda para los datos será aquel que minimice la tasa de error de predicción por *cross-validation*, y puede ser determinada por la función *cv.glmnet()*.

Regresión ridge

```
# Buscamos el mejor valor de lambda por cross-validation
set.seed(12345)
# Dado que "gaussian" es la opción por defecto en cuanto a la familia d
e modelos, tenemos que especificar que en nuestro caso tenemos una vari
able binomial
cv1 <- cv.glmnet(x, y, alpha = 0, family = "binomial", type.measure = "
class")
# El mejor valor de lambda
cv1$lambda.min
## [1] 1.804904
```

```
# Ahora ajustaremos el modelo con datos training
model1 <- glmnet(x, y, alpha = 0, family = "binomial", lambda =
```

```
cv1$lambda.min)
```

Este método de regresión indica que todas las variables (las 1013) serían significativamente predictoras de la variable resultado ("*vital_status*"). Por tanto, aquí no hemos conseguido una reducción de la dimensionalidad del dataset.

A continuación, valoraremos parámetros de rendimiento del modelo.

```
# Ahora realizamos predicciones en los datos test
x.test <- model.matrix(vital_status~., test.data)[,-1]
predictions1 <- model1 %>% predict(x.test) %>% as.vector()
# Estimación del rendimiento
ridge <- data.frame(
  RMSE = RMSE(predictions1, test.data$vital_status),
  Rsquare = R2(predictions1, test.data$vital_status))
ridge

##       RMSE    Rsquare
## 1 2.221089 0.03630012
```

Podemos representar la curva ROC (*receiver operating characteristic*) del modelo y calcular el área bajo la curva (AUC).

```
prob_std1 <- predict(model1, x.test, type="response", s=cv1$lambda.min)

y.test <- test.data$vital_status

pred_std1 <- prediction(prob_std1, y.test)
perf_std1 <- performance(pred_std1, measure = "tpr", x.measure = "fpr")

# Tasa de verdaderos positivos
tpr.points1rid <- attr(perf_std1, "y.values")[[1]]

# Tasa de falsos positivos
fpr.points1rid <- attr(perf_std1,"x.values")[[1]]

# AUC
auc1rid <- attr(performance(pred_std1, "auc"), "y.values")[[1]]
formatted_auc1rid <- signif(auc1rid, digits=3)

roc.data1rid <- data.frame(fpr=fpr.points1rid, tpr=tpr.points1rid, mode
l="GLM")

# Representamos la curva ROC
ggplot(roc.data1rid, aes(x=fpr, ymin=0, ymax=tpr)) +
    geom_ribbon(alpha=0.2) +
    geom_line(aes(y=tpr)) +
```

```
    ggtitle(paste0("Curva ROC regresión Ridge con AUC=", formatted_auc1
rid))
```

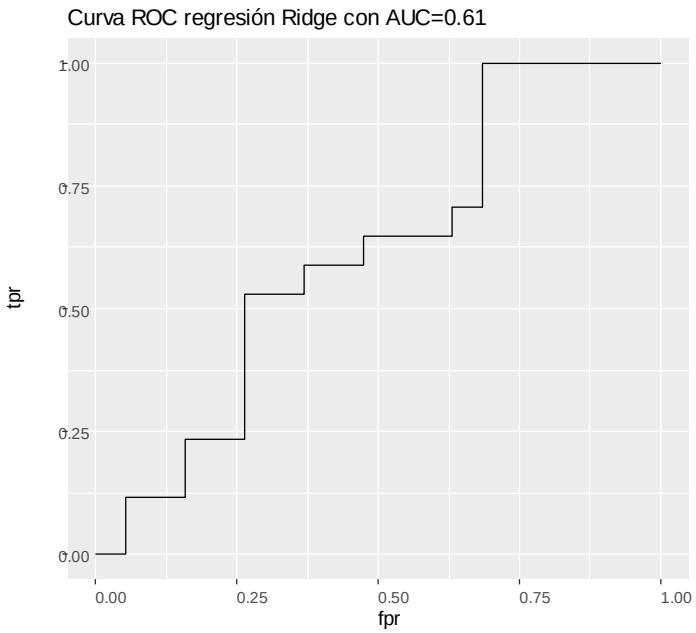

Figura 22. Curva ROC para el modelo de regresión penalizada *ridge*.

Para poder valorar estos valores de rendimiento, realizaremos la implementación de los otros modelos y la posterior comparación, aunque ya podemos ver en la **Figura 22** cómo el valor de AUC es muy deficiente (<0.7) y el ajuste de la curva ROC no es muy bueno.

Regresión lasso

```
# Optimización de lambda
set.seed(12345)
cv2 <- cv.glmnet(x, y, alpha = 1, family = "binomial", type.measure = "
class")
cv2$lambda.min
## [1] 0.08983051
```

```
model2 <- glmnet(x, y, alpha = 1, family = "binomial", lambda =
cv2$lambda.min)
lasso.coef <- coef(model2)
```

En este caso hemos logrado reducir notablemente las variables predictoras con respecto al caso anterior (ver siguiente apartado). A continuación, realizaremos la evaluación de los parámetros de rendimiento y la construcción de la curva ROC (ver **Figura 23**), utilizando el mismo código que para el método *ridge*.

```
lasso

##       RMSE    Rsquare
## 1 1.962708 0.06405086
```

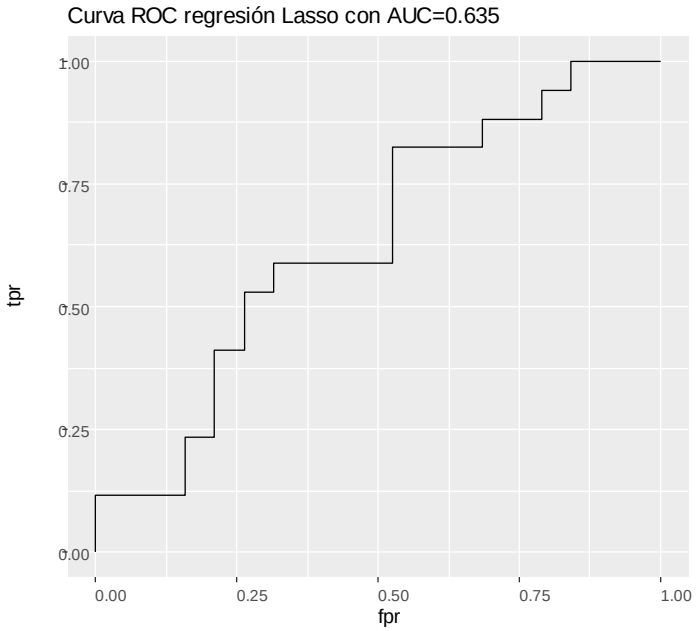

Figura 23. Curva ROC para el modelo de regresión penalizada *lasso.*

Regresión elastic net

Para desarrollar esta regresión, utilizaremos el paquete *caret*, para seleccionar automáticamente los mejores parámetros del modelo en función de nuestros datos. Siempre serán aquellos que minimicen el error de *cross-validation*.

```
# Construimos el modelo usando el set de entrenamiento
set.seed(12345)
# Utilizaremos la combinación de 10 diferentes valores de alpha y lambd
a para lograr los óptimos (tuneLength)
model3 <- train(vital_status ~., data = train.data, method = "glmnet",
trControl = trainControl("cv", number = 10), tuneLength = 10)

# Mostramos los mejores parámetros con la opción bestTune
model3$bestTune

##     alpha    lambda
## 100     1 0.1637004

# Por último, incluiremos estos parámetros optimizados al modelo final
Enet
coef(model3$finalModel, model3$bestTune$lambda)
```

El modelo *elastic net* muestra que ninguna de las variables que hemos introducido serían explicativas de nuestra variable objetivo.

Vamos a evaluar este modelo, realizando el mismo proceso que en casos anteriores.

```
enet
```

```
##       RMSE    Rsquare
## 1 1.962708 0.06405086
```

Los resultados son idénticos, en cuanto a rendimiento, no en selección de variables, que en el caso de la regresión *lasso*, debido a que el valor óptimo de alpha seleccionado por la optimización de parámetros es también 1.

Comparación del rendimiento de los diferentes modelos

El mejor modelo será aquel que muestre un menor error de predicción, RMSE (*root-mean-square error*).

```
mod.comp <- as.data.frame(predict(ridge, lasso, enet)
data.frame(mod.comp) <- c("Ridge", "Lasso", "ElasticNet")
```

Tabla 6. Parámetros de rendimiento para los 3 modelos de regresión logística penalizada testados.

	RMSE	Rsquare
Ridge	2.221090	0.0362996
Lasso	1.962708	0.0640509
ElasticNet	1.962708	0.0640509

Podemos observar cómo el menor valor de RMSE lo muestra el método *ridge* (ver **Tabla 6**). Sin embargo, como hemos visto más arriba, este método selecciona todas las variables como predictoras, lo cual no responde a nuestros objetivos. En el caso de las regresiones *lasso* y *elastic net*, los errores son los mismos, dado que el parámetro alpha usado en ambos casos es 1. Ante esta situación de igualdad, escogeríamos el modelo *lasso* ya que es el que permite seleccionar algunas de las variables que hemos introducido.

Vamos a extraer el nombre de las variables que resultan predictoras, que representan 4 genes y 2 variables clínicas.

```
lasso.coef
```

```
## COL10A1     -1.087282e-05
## GAD1         2.033050e-04
```

```
## PACSIN1                                          3.225438e-04
## PMFBP1                                           2.582856e-04
## gender                                           2.049760e-01
## treatments_radiation_treatment_or_therapy       -1.164954e-01
```

Análisis de significación biológica para los genes predictores de la variable "supervivencia"

Ahora podemos analizar la significación biológica de los genes que han resultado predictores de nuestra variable respuesta, es decir, que tienen potencialmente un papel en el desarrollo o gravedad de la enfermedad. Como hemos visto anteriormente, uno de los recursos más utilizados con este objetivo son las bases de datos de *Gene Ontology* (GO), que anotan los genes de acuerdo a un diccionario de términos.

Vamos a procesar el nombre de los 4 genes seleccionados (COL10A1, GAD1, PACSIN1, PMFBP1) en la herramienta web https://tools.dice-database.org/GOnet/ , donde podemos introducirlos para obtener información acerca de las vías metabólicas y biológicas que se pueden ver afectadas.

El análisis revela que no hay ningún término GO que se encuentre enriquecido significativamente. La siguiente *network* nos muestra las conexiones en términos de GO de dichos genes (ver **Figura 24**).

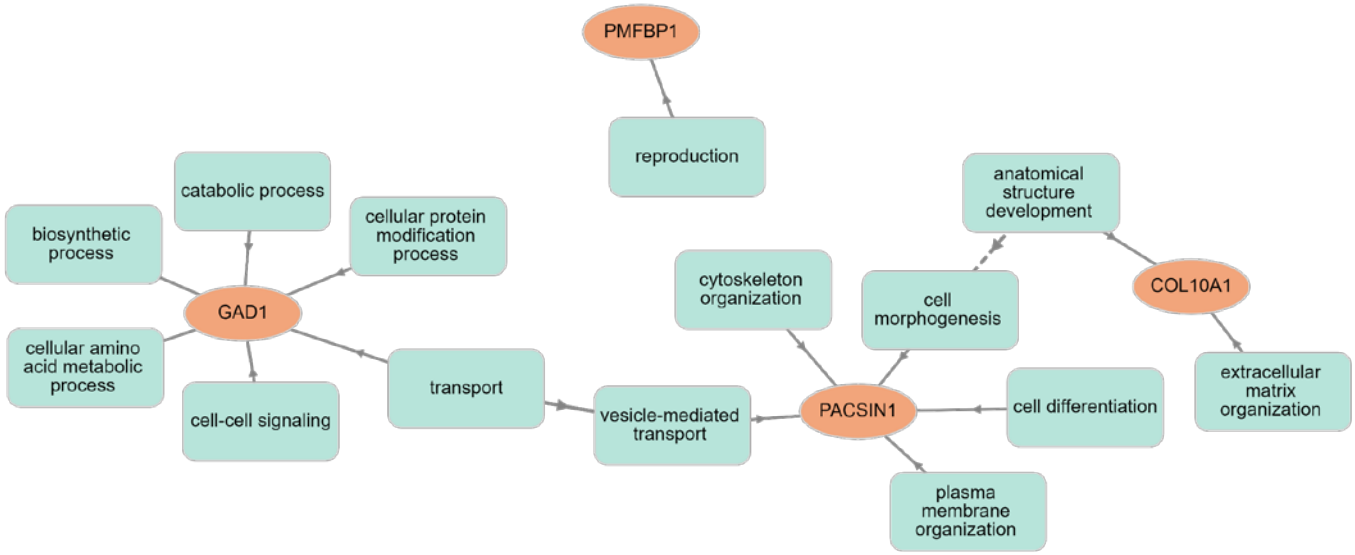

Figura 24. Red de términos y funciones GO en relación con los genes seleccionados por regresión logística penalizada.

2.3.6. Integración de *datasets* mediante *SNFtools*

Ahora vamos a testar herramientas concretas de integración de datos ómicos y datos clínicos, comenzando con el paquete *SNFtools*. Dado que el análisis de selección de variables ya nos ha mostrado qué genes y variables clínicas serían explicativos de la variable respuesta "*vital_status*" realizaremos dos aproximaciones*: i)* Utilizando las 6 variables predictoras resultado del paso anterior. *ii)* Utilizando todas las variables ómicas (los 995 genes resultantes del análisis de DE) y clínicas.

Esta estrategia nos va a permitir, mediante la comparación de parámetros de optimización, si el proceso de *feature selection* ha sido correcto, e inferir por tanto si las variables seleccionadas serían buenas predictoras.

Dataset filtrado por selección de variables

Tenemos dos *datasets* de información clínica y expresión génica, con las filas mostrando pacientes y las columnas las variables. No sería necesario normalizar los datos, ya que nuestro pre-procesamiento de los datos de transcriptómica ya incluye un proceso de normalizado de los conteos génicos.

```
# En primer lugar, seleccionaremos solo las variables clave
clinical2 <- subset(clinical, select=c(gender, treatments_radiation_tre
atment_or_therapy))
transcriptomics2 <- subset(transcriptomics, select=c(COL10A1, GAD1, PAC
SIN1, PMFBP1))
```

Ahora por requerimientos del paquete vamos a necesitar la información del "*cluster label*", es decir, de la etiqueta de cluster de cada muestra. Esto corresponderá al valor de la variable respuesta (*vital_status*) en cada muestra:

```
truelabel <- clinical$vital_status
```

Ahora calcularemos las matrices de distancias para ambos *datasets*, usando por defecto la distancia Euclídea.

```
dist1 <- as.matrix(dist(clinical2))
# En el caso de la matriz de expresión génica, y dado que los datos son
continuos, vamos a utilizar una expresión ligeramente distinta, siguien
do las especificaciones del manual
dist2 <- dist2(as.matrix(transcriptomics2),as.matrix(transcriptomics2))
```

Para convertir las matrices de distancias en matrices de similaridad, se debe usar la función *affinityMatrix*.

```
# Establecemos los parámetros
K <- 20 # número de vecinos, normalmente 10~30
alpha <- 0.5 # hiperparámetro, normalmente 0.3~0.8
T <- 10 # número de iteraciones, normalmente 10~20
C <- 2 # Se establecen el número de clusters (en nuestro caso 2, los pr
ocedentes de la variable respuesta vital_status)

# Ahora construimos los gráficos de similaridad
W1 <- affinityMatrix(dist1, K, alpha)
W2 <- affinityMatrix(dist2, K, alpha)
```

Tras calcular estas matrices de similaridad, podemos intentar el *clustering* de cada matriz individualmente (ver **Figura 25**):

```
displayClustersWithHeatmap(W1, spectralClustering(W1, C))
displayClustersWithHeatmap(W2, spectralClustering(W2, C))
```

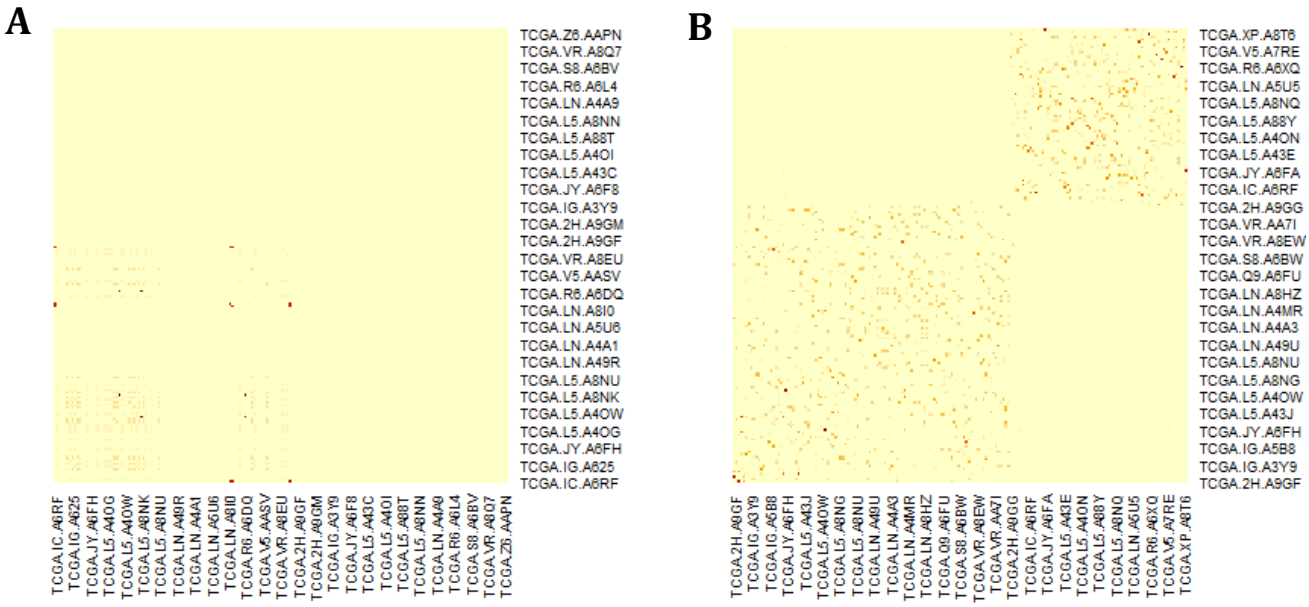

Figura 25. Visualización de los *clusters* presentes en las matrices de similaridad basadas en variables clínicas (**A**) y de expresión génica (**B**) seleccionadas.

Como se puede ver, ambas matrices muestran algo de información sobre clusters internamente. La matriz basada en datos ómicos (génicos) parece mostrar información de 2 clusters, mientras que la representación de las variables clínicas es algo más difusa. Podemos integrar los dos datasets en uno para lograr un *clustering* más rico y sólido mediante *similarity network fusion* (SNF):

```
# Ahora crearemos la red fusionada (fused network)
W = SNF(list(W1, W2), K, T)
```

Con este gráfico W unficado se puede realizar un *clustering* espectral (ver **Figura 26**):

```
# Establecemos el agrupamiento
group <- spectralClustering(W, C)
displayClustersWithHeatmap(W, group)
```

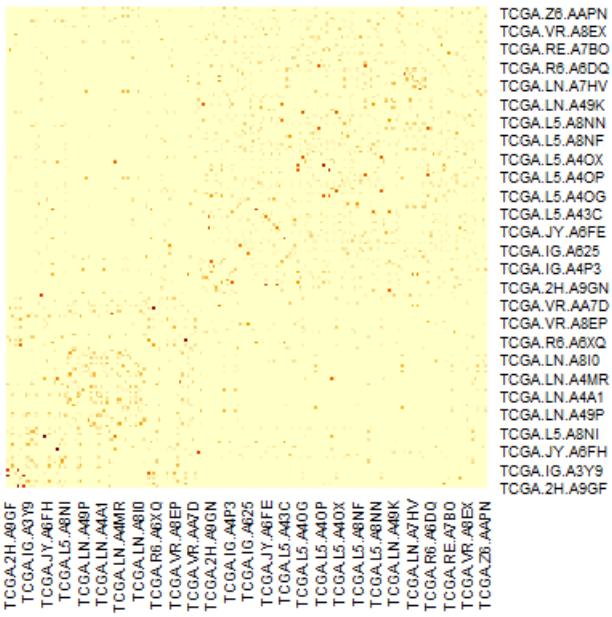

Figura 26. Visualización de los *clusters* presentes en la red fusionada de datos de expresión génica y clínicos.

Para evaluar la bondad de ajuste de los resultados del *clustering* se puede buscar la concordancia entre cada individuo de la red y la red fusionada:

```
ConcordanceMatrix <- concordanceNetworkNMI(list(W, W1, W2), C)
colnames(ConcordanceMatrix) <- c("global", "clínica", "transcriptómica"
)
rownames(ConcordanceMatrix) <- c("global", "clínica", "transcriptómica"
)
ConcordanceMatrix
```

```
##                    global     clínica
## global          1.00000000 0.044647106
## clínica         0.04464711 1.000000000
## transcriptómica 0.08231501 0.006923368
##                 transcriptómica
## global              0.082315013
## clínica             0.006923368
## transcriptómica     1.000000000
```

Dada una lista de matrices, la combinada (W), la debida a factores clínicos (W1) y la de expresión génica (W2), esta función devuelve una matriz que contiene los valores NMI (*normalized mutual info score*) entre las asignaciones de los clusters realizadas por *clustering* espectral. Esta matriz de afinidades revela la cercanía de las estimaciones de asignaciones entre las 3 matrices. En este caso concreto se puede ver una mayor afinidad entre las asignaciones de la matriz global y la transcriptómica, de

lo que se puede deducir que la variable respuesta (o etiqueta) estaría mejor explicada por las variables de expresión génica que por las variables clínicas.

Otra manera de testar la eficacia del método es realizar un ejemplo de la predicción de nuevas etiquetas con propagación de etiquetas.

```
# En primer lugar, creamos un objeto lista combinado con los dos datase
ts
join_SNF <- list(clinical2, transcriptomics2)
# A continuación, creamos datasets de entrenamiento y prueba
n <- floor(0.8*length(truelabel)) # número de casos de training
set.seed(12345)
trainSample <- sample.int(length(truelabel), n)
train <- lapply(join_SNF, function(x) x[trainSample, ])
test <- lapply(join_SNF, function(x) x[-trainSample, ])
groups2 <- truelabel[trainSample]
```

Ahora aplicaremos la función de predicción a los datos.

```
method <- TRUE
newLabel <- groupPredict(train,test,groups2,K,alpha,T,method)

# Compararemos la exactitud de la predicción
accuracy <- sum(truelabel[-trainSample] == newLabel[-c(1:n)])/(length(t
ruelabel) - n)
accuracy

## [1] 0.5675676
```

La exactitud emanada de las variables pre-filtradas por *feature selection* es de 0.57.

Dataset sin filtrar

Debemos eliminar del *dataset* clínico inicial la variable respuesta que en este paquete introducimos como etiqueta (*truelabel*), que es *vital_status*, y después utilizaremos el mismo código que en el apartado anterior.

```
clinical_SNFglob <- clinical[, -15]
```

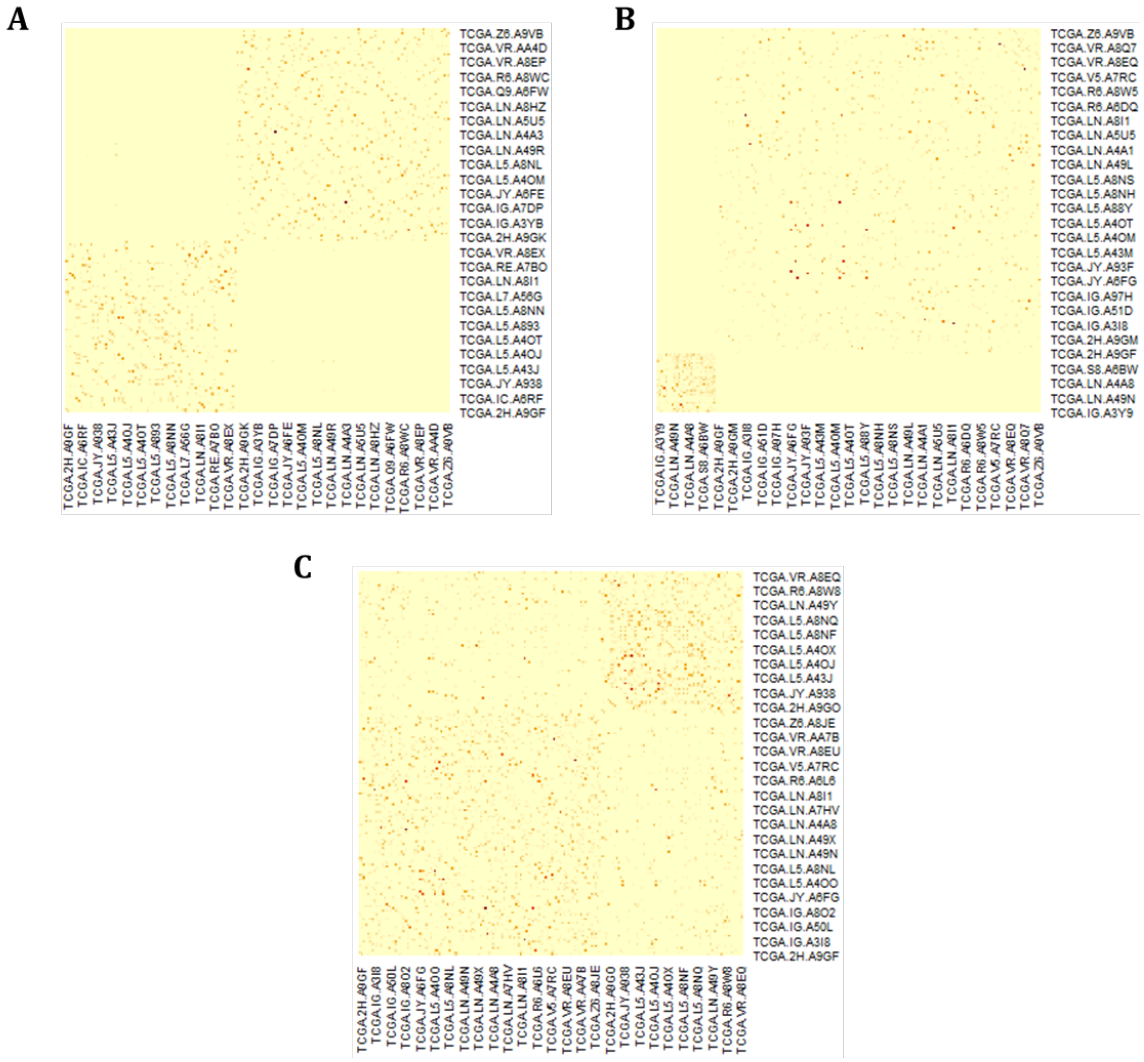

Figura 27. Visualización de los *clusters* inferidos para los datos clínicos (**A**), de expresión génica (**B**) y la red fusionada (**C**).

La concordancia entre redes muestra los siguientes resultados:

```
ConcordanceMatrix2

##                       global     clínica
## global          1.00000000 0.152076078
## clínica         0.15207608 1.000000000
## transcriptómica 0.07918927 0.005339008
##                 transcriptómica
## global              0.079189269
## clínica             0.005339008
## transcriptómica     1.000000000
```

Aquí, sin embargo, en contraposición con el caso anterior en el que realizábamos un filtrado previo de variables, la asignación derivada de la matriz de variables clínicas sería la más próxima a la que se obtiene de la matriz global, por lo que se

puede considerar que el conjunto de variables clínicas explicaría de manera más adecuada la diferente asignación de las etiquetas (variable respuesta "vital_status").

Ahora aplicaremos la función de predicción a los datos mediante partición del *dataset* y cálculo de la exactitud.

```
accuracy.2
```

```
## [1] 0.6216216
```

Aquí vemos que se obtiene un mayor valor de *accuracy* que lo observado cuando utilizábamos las variables pre-seleccionadas por regresión logística (0.57). Sin embargo, la diferencia entre estos parámetros de rendimiento no es tan grande, y, por tanto, frente a una situación en la que nos enfrentemos a complejidades computacionales por el uso de numerosas variables (como puede ser nuestro caso, dado que se escogían 995 variables ómicas más las clínicas), podríamos realizar la implementación del método de redes de similaridad directamente con las variables pre-seleccionadas por regresión.

2.3.7. Integración de *datasets* mediante *mixOmics*

Mediante este paquete, vamos a aplicar el método sPLS (*sparse Partial Least Squares*), lo que nos va a permitir realizar simultáneamente tanto la integración como la selección de variables de dos *datasets*.

Por tanto, vamos a partir de los *datasets* completos para poder comparar si el proceso de selección de variables que se produce internamente en el paquete tiene unos resultados comparables a lo obtenido por regresión logística penalizada.

En primer lugar, estableceremos nuestros *sets* de trabajo: *i)* Expresión génica: *transcriptomics*; *ii)* Datos clínicos: podemos usar los datos clínicos filtrados por la variable respuesta "*vital_status*", usados en el paquete anterior. Sin embargo, este paquete está pensado para ser usado con datos cuantitativos, por lo que para que pueda acoger datos categóricos, se requiere una transformación con la creación de variables *dummy*. Para ello se puede emplear la función *dummy* del paquete *dummies* para la transformación; *iii)* Variable respuesta: *vital_status*.

```
# Extraeremos del dataset las variables cuantitativas para que no se so
metan a la transformación siguiente
keep <- clinical_SNFglob[c(10,13)]
# Eliminamos estas variables del dataset
```

```
clinical_SNFglob2 <- clinical_SNFglob[-c(10,13)]
# A continuación, efectuaremos la transformación a variables dummy del
resto del dataset
cat.vars <- apply(clinical_SNFglob2, 2, function(x) {
  return(dummy(x, sep = "."))
})
```

Ya podemos establecer nuestros *datasets*.

```
X <- transcriptomics
Y <- clinical_SNFglob
```

Analizaremos estos dos sets de datos usando sPLS con un modelo de regresión intentando predecir o explicar las variables clínicas con respecto a los niveles de expresión génica.

Análisis preliminar con PCA

```
pca.gene <- pca(X, ncomp = 10, center = TRUE, scale = TRUE)
pca.gene

## Cumulative proportion explained variance for the first 10 principal
components, see object$cum.var:
##       PC1       PC2       PC3       PC4
## 0.2298004 0.3446345 0.4091645 0.4524401
##       PC5       PC6       PC7       PC8
## 0.4809116 0.5028883 0.5189515 0.5339800
##       PC9      PC10
## 0.5484977 0.5615268


pca.clinical <- pca(Y, ncomp = 10, center = TRUE, scale = TRUE)
pca.clinical


## Cumulative proportion explained variance for the first 10 principal
components, see object$cum.var:
##       PC1       PC2       PC3       PC4
## 0.2666873 0.4111922 0.5231714 0.6055731
##       PC5       PC6       PC7       PC8
## 0.6845478 0.7533955 0.8118451 0.8624417
##       PC9      PC10
## 0.9050338 0.9441774
```

En ambos casos, los 3 primeros componentes engloban un 40-50% de variación de los datos.

Representación de las muestras

Para mostrar la ubicación de las muestras, podemos utilizar información de variables que *a priori* sepamos, por análisis previos, que pueden tener influencia en los datos:

```
# Representaremos los puntos de las muestras en función de la variable
sexo, tanto para el dataset de expresión como para el clínico
plotIndiv(pca.gene, comp = c(1, 2), group = dataClin2$vital_status,
ind.names = dataClin2$gender,
legend = TRUE, title = 'PCA1&2 Expresión génica')
plotIndiv(pca.clinical, comp = c(1, 2), group = dataClin2$vital_status,
ind.names = dataClin2$gender,
legend = TRUE,title = 'PCA1&2 Clínica')
```

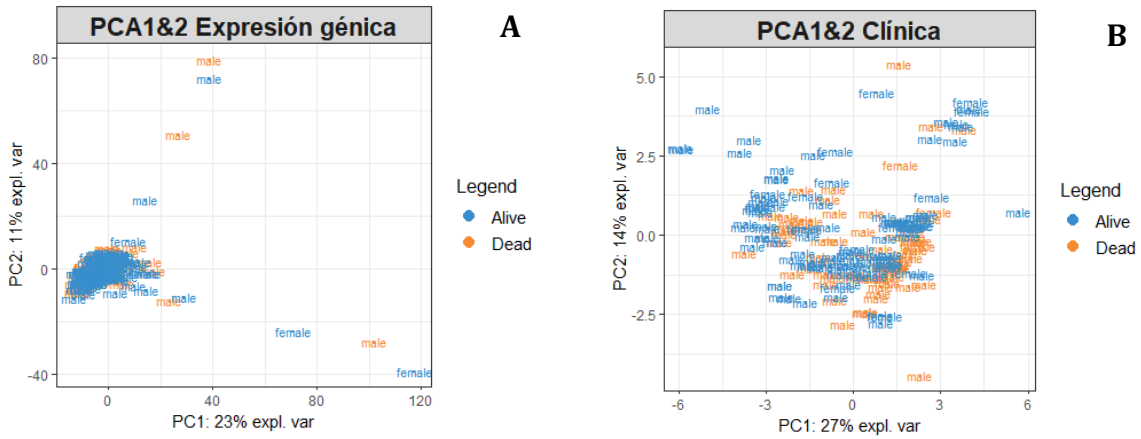

Figura 28. Posición de las muestras en un *plot* en función de los dos primeros componentes del PCA para los datos de expresión gènica (**A**) y clínicos (**B**). Las leyendas y código de color representan el estatus vital del individuo y su sexo.

El *plot* de las muestras debido a la información de expresión génica revela una posición muy *clusterizada* de la mayoría de ellas, con algunos *outliers*. Del mismo modo, la posición de las muestras según la dimensión clínica parece bastante independiente del estatus de nuestra variable respuesta.

Análisis sPLS

A continuación, teniendo en cuenta la información emergida del análisis del PCA, en cuanto al número de componentes, vamos a indicar un número máximo de variables a seleccionar con los argumentos *keepX* y *keepY*. Teniendo en cuenta los valores que hemos obtenido en el análisis por regresión logística penalizada, usaremos un umbral

de 10 variables por componente. *A priori* retendremos 10 componentes para cada componente.

```
esca.spls <- spls(X, Y, ncomp = 10, keepX = c(10,10,10), keepY = c(5,5,
5), mode = "regression")

tune.spls <- perf(esca.spls, validation = "Mfold", folds = 10, progress
Bar = FALSE, nrepeat = 50)
```

Q2.total puede ser usado para tunear el número de componentes. El número óptimo de componentes a elegir se indica una vez que el Q2 baja del umbral de 0.0975. En nuestro caso, 2 componentes serían suficientes para el modelo.

```
tune.spls$Q2.total

##             Q2.total
## 1 comp   0.129268337
## 2 comp   0.017953778
## 3 comp  -0.004685593
## 4 comp  -0.010927788
## 5 comp  -0.002524121
## 6 comp  -0.025398039
## 7 comp   0.009783121
## 8 comp  -0.010406286
## 9 comp  -0.034963726
## 10 comp -0.011427935
```

Ahora representamos las muestras proyectadas en los componentes sPLS usando la función plotIndiv() (ver **Figura 29**).

```
# Gráfico para el espacio Y (clínica)

plotIndiv(esca.spls, comp = 1:2, rep.space= 'Y-variate', group = dataCl
in2$vital_status,
ind.names = dataClin2$gender,
legend = TRUE, legend.title = 'Vital_status', title = 'ESCA sPLS comp 1
- 2, Y-space')

# Gráfico para el espacio X (expresión)

plotIndiv(esca.spls, comp = 1:2, rep.space= 'X-variate', group = dataCl
in2$vital_status,
ind.names = dataClin2$gender,
legend = TRUE, legend.title = 'Vital_status', title = 'ESCA sPLS comp 1
- 2, X-space')

# Gráfico combinado (X-Y)

plotIndiv(esca.spls, comp = 1:2, rep.space= 'XY-variate', group = dataC
lin2$vital_status,
ind.names = dataClin2$gender,
```

```
legend = TRUE, legend.title = 'Vital_status', title = 'ESCA, sPLS comp
1 - 2, XY-space')
```

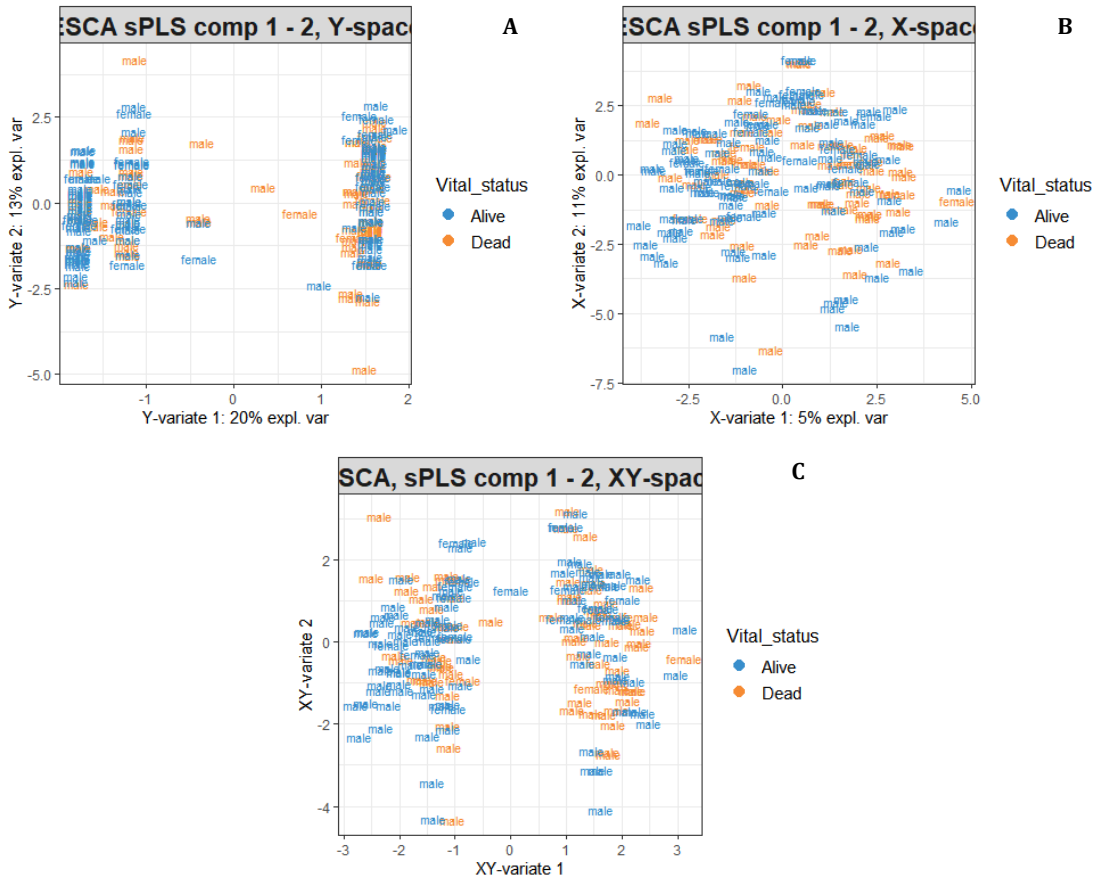

Figura 29. Posición de las muestras en un *plot* en función de los dos primeros componentes del sPLS para los datos clínicos (**A**), de expresión génica (**B**) y combinados (**C**). Las leyendas y código de color representan el estatus vital del individuo y su sexo.

La **Figura 29C** representa la distribución de las muestras dimensionadas en el subespacio medio, en el cual las coordenadas están promediadas desde los subespacios X e Y. Visualmente, parece que ninguna de las dos variables, *"vital_status"* (nuestra variable respuesta) o sexo, tienen un efecto en la distribución de las muestras.

Las variables seleccionadas por sPLS pueden proyectarse en un *plot* circular de correlación usando la función *plotVar*:

```
plotVar(esca.spls, comp = 1:2,
        var.names = list(X.label = colnames(transcriptomics),
        Y.label = TRUE), cex = c(3, 3))
```

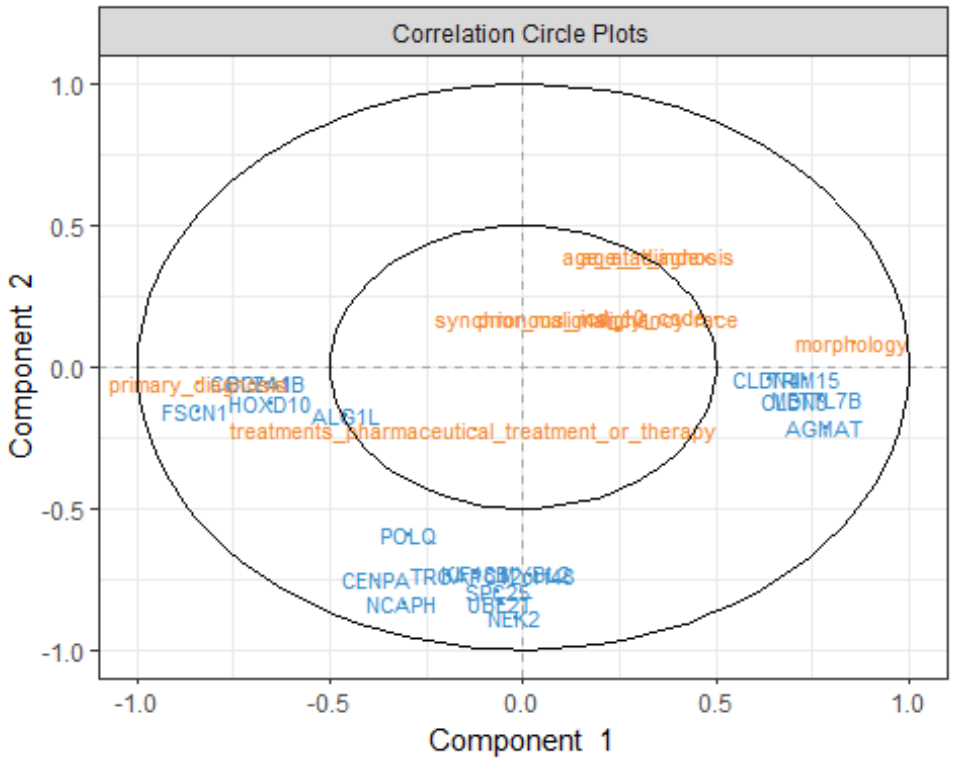

Figura 30. Gráfico de correlación de las variables de expresión (genes) y clínicas. Se muestran dos círculos, de radio 0.5 y radio 1 para denotar la importancia de las variables.

Las variables que están positivamente correlacionadas se encuentran próximas las unas a las otras en el *plot* (ver **Figura 30**). Cuando la correlación es fuertemente negativa, las variables o grupos de variables se proyectan en sitios diametralmente opuestos del círculo de correlación. Cuanto más cercano al centroide del círculo, menos relevantes las variables. En el *plot* podemos observar cómo existen ciertas agrupaciones de variables ómicas (genes) que en principio estarían correlacionadas entre sí. La situación de las variables clínicas es algo más dispersa y no muestra esta *clusterización* preferencial que reflejaría correlación entre ellas.

Otro tipo de representaciones nos pueden ayudar a la interpretación de los resultados. Las *clustered image maps* (CIM) son *heatmaps* que nos reflejan de nuevo la correlación entre variables (ver **Figura 31**).

```
# CIM
cim(esca.spls, comp = 1:2, xlab = "clínica", ylab = "genes", save = 'jp
eg', name.save = 'CIM')
```

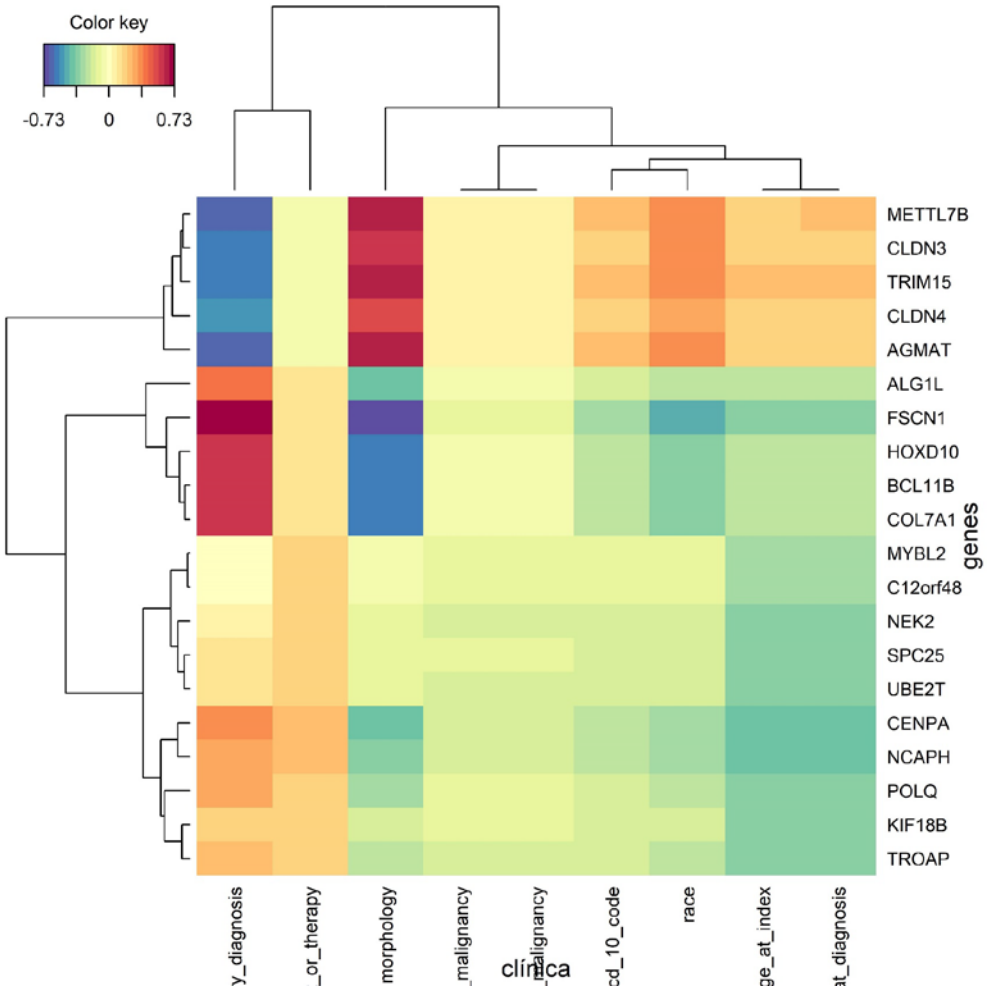

Figura 31. *Heatmap* CIM que representa las correlaciones entre variables de expresión y clínicas.

Las filas y columnas se encuentran reordenadas en función de un *clustering* jerárquico, mostrándose los dendrogramas para filas (X, genes) y columnas (Y, variables clínicas). Los bloques de color representarían asociaciones entre diferentes *subsets* de las variables X e Y. Los colores cálidos muestran correlación positiva los colores fríos, correlación negativa, y los tonos amarillos unos débiles valores de correlación.

Se puede observar una correlación positiva fuerte entre la variable "*primary_diagnosis*" y un set de genes (COL7A1, BCL11B, FSCN1, HOXD10, ALG1L), todos estos agrupados en un *cluster* monofilético. Este mismo *cluster* correlaciona negativamente con la variable "*morphology*", mientras que ésta se encontraría asociada a los genes METTL7B, CLDN3, TRIM15, CLDN4 y AGMAT.

Otra representación interesante son las redes de relevancia *(relevance networks*), mostradas aquí para los dos primeros componentes (ver **Figura 32**).

```
network(esca.spls, comp = 1:2, color.node = c("orange","blue"), cutoff
= 0.60, save = 'jpeg', name.save = 'PLSnetwork')
```

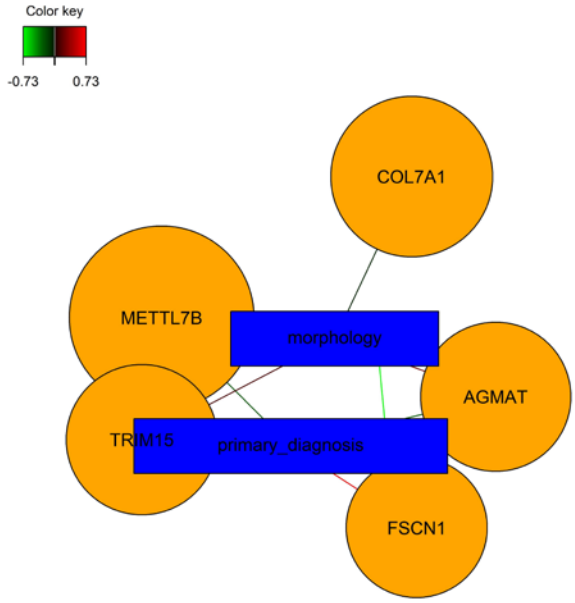

Figura 32. Red de relevancia para los dos primeros componentes. Los vectores verdes y rojos representan correlaciones positicas y negatives, respectivamente. Las variables X se denotan por círculos (genes) y las Y por rectángulos (clínicas).

Si queremos extraer los genes seleccionados:

```
# tenemos que extraer los genes de cada componente seleccionado
select.genes.1 <- selectVar(esca.spls, comp = 1)
select.genes.2 <- selectVar(esca.spls, comp = 2)
select.genes.3 <- selectVar(esca.spls, comp = 3)

select.genes.1$X$name

##  [1] "FSCN1"   "COL7A1"  "BCL11B"  "HOXD10"  "AGMAT"   "TRIM15"
##  [7] "RFC4"    "METTL7B" "CLDN3"   "ALG1L"
```

No hay ninguna coincidencia entre las variables seleccionadas por el método sPLS en mixOmics y aquellas filtradas por regresión logística penalizada. Solamente hemos encontrado una similaridad en cuanto a genes que codifican para cadenas proteínicas de colágeno (COL10A1, seleccionado por *lasso* y COL7A1 por sPLS).

2.3.8. Estandarización del *pipeline*

La **Figura 33** muestra un resumen gráfico del flujo de trabajo realizado en el presente trabajo para el tratamiento conjunto de datos ómicos y clínicos.

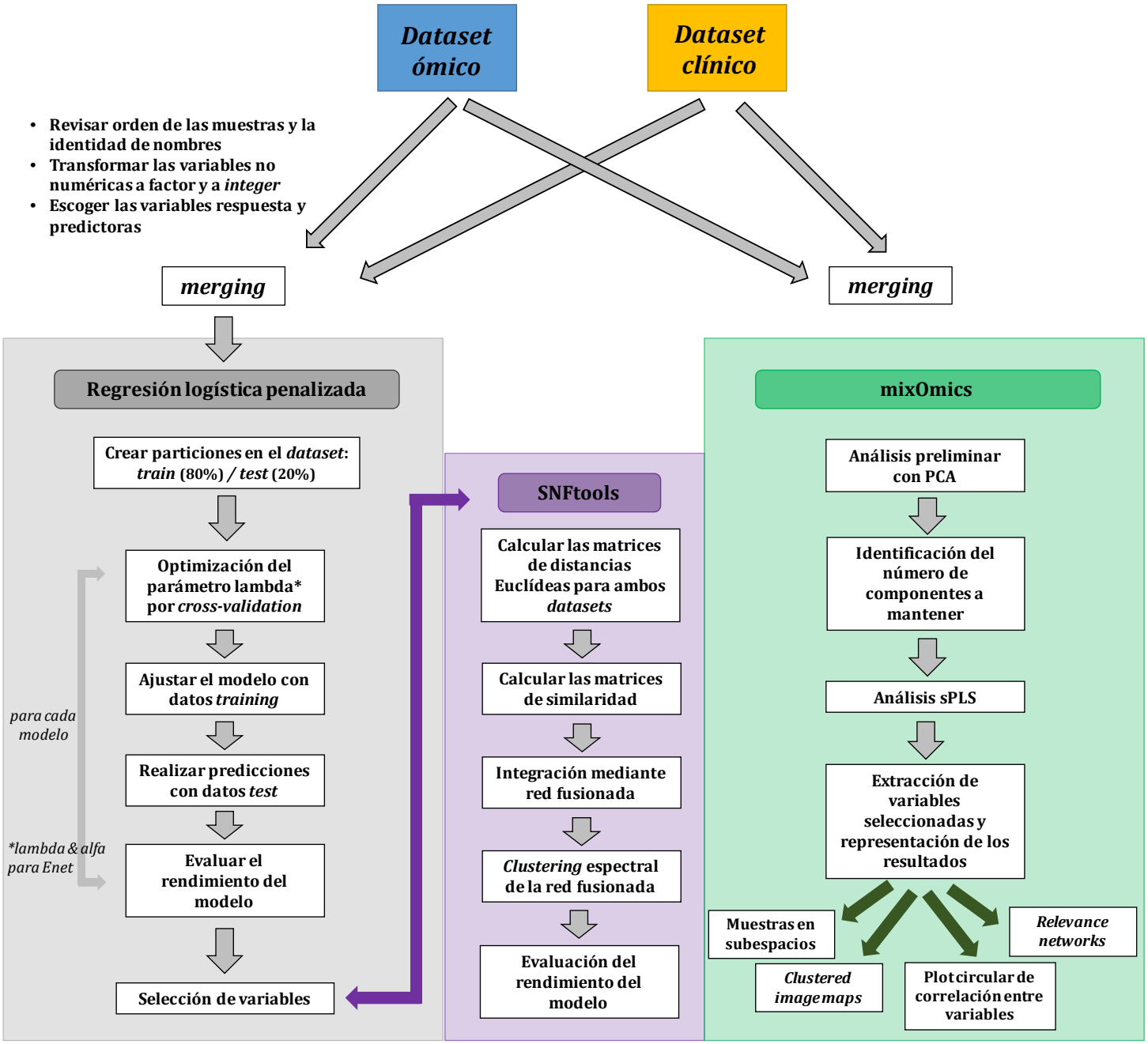

Figura 33. *Pipeline* desarrollado en el presente trabajo que muestra el uso de diferentes recursos para la integración de datos ómicos y clínicos.

2.4. Discusión

El presente estudio ha desarrollado una integración de datos ómicos y clínicos a partir de diferentes metodologías, siguiendo un flujo de trabajo bioinformático sistemático y complejo. En el entorno del programa R, hemos podido evaluar las ventajas y limitaciones de diferentes paquetes y metodologías.

Cuando se proyectó este trabajo, no conocíamos las múltiples funcionalidades del paquete *TCGAbiolinks*, más allá de su uso como plataforma para la descarga de datos del repositorio TCGA. Sin embargo, investigando las *vignettes* publicadas por sus autores, pudimos comprobar cómo este paquete funciona como una suite que permite también realizar pre-procesamientos de datos ómicos (incluyendo normalización y filtrado de variables), análisis de expresión y metilación diferencial, así como análisis de significación biológica. Resulta muy ventajoso poder realizar todos estos procedimientos en cascada, ya que se trata de un método muy fluido de procesar los datos.

A lo largo del trabajo, hemos podido visualizar que la implementación de particiones en los datos es un procedimiento altamente adecuado, ya que nos permite construir, ajustar y entrenar los modelos, así como evaluar su rendimiento. Cuando empleamos modelos alternativos, como es el caso de este estudio, es fundamental conocer estos parámetros, que nos ayudan a poder escoger el modelo óptimo para nuestros datos.

El desarrollo de los tres modelos de regresión logística penalizada (*ridge*, *lasso* y *enet*), ha mostrado diferencias muy notables respecto a la selección de variables. La primera aproximación no permitía realizar ninguna reducción de la dimensionalidad del *dataset*, ya que escogía todas las variables disponibles como predictoras de la variable respuesta ("*vital_status*"), mientras que la última no mostraba la influencia de ninguna de estas variables. Solamente la regresión *lasso* reflejaba un resultado que podía ser utilizado para nuestros propósitos. Los parámetros de rendimiento (tanto RMSE como AUC) de los tres modelos son muy pobres, por lo que nos indican que estos modelos no logran explicar el comportamiento de la variable respuesta, es decir, que no permitían la clasificación de las muestras con respecto a dicha variable.

Una muestra de que estos modelos no son adecuados para explicar la naturaleza de la variable "*vital_status*", que nos estaría reflejando la supervivencia de los

individuos, puede encontrarse en una falta de términos GO enriquecidos en el análisis subsecuente de significación biológica de los genes seleccionados por regresión *lasso*.

El uso del paquete *SNFtools* resulta una interesante y diferente aproximación a la integración de datos. Puede ser utilizado para integrar datos ómicos de diferente naturaleza o datos ómicos con datos clínicos. El punto limitante de este paquete radica en que exclusivamente pueden ser integrados dos datasets, lo que lo diferencia del resto de enfoques metodológicos empleados en este trabajo (regresión logística y paquete *mixOmics*). Tal y como sucedía con la regresión *lasso*, los parámetros de rendimiento de las redes fusionadas resultantes de la integración ómica-clínica son muy limitados.

Por último, las funcionalidades del paquete *mixOmics* han sido exploradas mediante el uso del enfoque *sPLS*. Los resultados pueden ser visualizados mediante representaciones gráficas muy ilustrativas que nos ayudan a comprender la naturaleza de las relaciones entre las variables seleccionadas en términos de correlaciones. Este paquete incluye los dos pasos requeridos respecto a la unión de capas ómicas y clínicas: *i)* reducción de la dimensionalidad de los datos y *ii)* integración de los datasets. Es interesante destacar aquí que no ha habido paralelismo entre las variables (genes) seleccionadas por este método y las que emergían de la regresión *lasso.*

La variable escogida como *outcome* ha condicionado los resultados obtenidos en el presente estudio, debido a que dicha variable no permite una discriminación entre las muestras, ni desde la perspectiva ómica ni la clínica. La **Figura 28** representa esta situación, ya que los individuos no se clusterizan en el PCA en ninguna de ambas capas con respecto a los dos estados de la variable respuesta "*vital_status*" (*alive/dead*). Además, podíamos también indicar que esta variable, contrariamente a nuestro planteamiento inicial, no sería una buena aproximación a la supervivencia de los individuos frente a la enfermedad. Este hecho podemos fundamentarlo en la elevada edad de las muestras (media ~ 63 años), lo que podría indicar que el fallecimiento del individuo no se deba exclusivamente al desarrollo de la enfermedad o su gravedad, sino a procesos de senescencia.

Al margen de la selección de la variable respuesta, pueden existir otros factores que hayan podido influir en que los modelos desarrollados no hayan sido robustos o en otros resultados obtenidos. Al analizar el *set* de datos, hemos podido observar

cómo las dos clases de muestras que hemos contrapuesto para realizar los análisis son muy desiguales en tamaño (muestras tumorales: 184, muestras normales: 11). Este desequilibrio entre los grupos muestrales ha podido afectar también a los resultados. Así, puede que teniendo una mayor representación del grupo control, pudiéramos haber logrado algún resultado significativo respecto al análisis de metilación diferencial.

3. Conclusiones

Conclusiones del presente trabajo

1. La integración de datos clínicos en el marco de las ciencias ómicas es fundamental para contemplar todas las variables que van a afectar a un determinado fenotipo o enfermedad humana.
2. El paquete *TCGAbiolinks* es una poderosa herramienta para las primeras etapas de análisis de datos ómicos y clínicos, proporcionando una eficaz descarga de dichos datos, así como un amplio espectro de funciones analíticas para el procesamiento de los *datasets.*
3. Resulta importante realizar un tratamiento de las capas ómicas de carácter previo a la integración con los datos clínicos, con el objetivo de reducir las dimensiones de esta información molecular, de tal modo que se intente aproximar el número de variables contenidas en cada *set* de datos.
4. Para la implementación de modelos predictivos, usando cualquiera de las metodologías empleadas en el presente trabajo, es fundamental el estudio de los *datasets* de análisis, para seleccionar tanto la variable respuesta como las posibles variables predictoras que van a ser contempladas.
5. El desarrollo de particiones en los datos permite crear los modelos predictivos y evaluar el rendimiento de los mismos mediante los propios datos.
6. En el presente estudio se ha mostrado cómo diferentes enfoques de selección de variables no han mostrado resultados simétricos entre sí, tanto dentro de los tres modelos de regresión logística penalizada, como con el uso de la metodología sPLS, lo que puede ser explicado por la baja discriminación de las muestras respecto a la variable respuesta escogida.

Autoevaluación del cumplimiento de los objetivos y del seguimiento de la planificación

Se han cumplido los objetivos en su totalidad y con el marco temporal inicialmente previsto, salvo pequeñas desviaciones referidas a tareas concretas.

- En la planificación temporal de las tareas (ver **Figura 1**), se planteó inicialmente que la tarea "*Identificación de los diferentes enfoques*

bioestadísticos para la integración de datos y la creación de modelos predictivos" fuera abordada de carácter previo a "*Selección del dataset*". Sin embargo, en la práctica, se pudo comprobar cómo era más adecuado realizar la búsqueda del *set* de datos con la mayor celeridad posible, con el fin de poder comenzar a familiarizarse con el procesamiento del mismo. Por ello, el orden temporal de estas dos tareas fue invertido con respecto a lo inicialmente propuesto.

- Por otro lado, la tarea "*Análisis de las variables contenidas en el dataset y procesamiento de las mismas para ser incluidas en los modelos*", previa al desarrollo de los modelos predictivos, es de alta complejidad, debido a los requerimientos específicos de los paquetes utilizados para su implementación. Por tanto, el tiempo destinado a esta tarea ha sido ligeramente mayor a lo previsto inicialmente. Esto ha sido así ya que los *datasets* utilizados han requerido un intenso trabajo de procesamiento para que las variables que contienen puedan ser asumidas por las metodologías bioinformáticas empleadas. En concreto, se han tenido que modificar la naturaleza de determinadas variables, así como filtrar otras de bajo interés o que no pudieran ser introducidas en los modelos. Asimismo, los requerimientos de los paquetes de R empleados han provocado que la variable objetivo o respuesta escogida tuviera necesariamente que ser, o bien numérica o bien binomial, lo que también ha limitado las posibilidades a la hora de escoger dicha variable. Sin embargo, la variable *outcome* finalmente seleccionada se considera que es muy interesante como indicador de la gravedad del proceso tumoral y con posibles relaciones con otras variables tanto clínicas como ómicas.

También, en inicio se había planteado el uso de otras ómicas al margen de la expresión génica. En la práctica, se ha llevado a cabo la descarga y tratamiento de datos de metilación génica (epigenómica), pero dado que los resultados no mostraban patrones de metilación diferencial entre grupos de muestras, se decidió no seguir adelante con estos datos y, por tanto, no introducir esta capa en los análisis de integración.

En el análisis de riesgos que presentábamos al inicio de la *Memoria*, en el *Plan de Trabajo*, contemplábamos la posibilidad de sufrir problemas en cuanto a la capacidad de computación. Efectivamente, esto ha sido un aspecto que hemos podido experimentar, sobre todo en cuanto al pre-procesamiento de las capas ómicas (transcriptómica y epigenómica), dada la gran cantidad de datos contenidos y el elevado número de muestras utilizado. Además, este pre-procesamiento no fue inicialmente expuesto en el *Plan de Trabajo*, ni el análisis de expresión diferencial de genes, que fue desarrollado al comenzar con el tratamiento de estos datos ómicos. Pese a no haber sido previstas estas tareas, fueron desarrolladas de manera fluida junto con las tareas planificadas.

Futuras líneas de trabajo

Para completar el procesamiento bioinformático de los datos ómicos (y clínicos en el presente caso), deben ser desarrolladas posteriores etapas tras desarrollar y evaluar los modelos computacionales (ver en la *Introducción* la sección "*Desafíos informáticos y tecnológicos en la era ómica*").

Estos pasos serían jerárquicamente los siguientes: *i)* confirmar los modelos en s*ets* de datos independientes, *ii)* hacer públicas las metodologías empleadas en términos de códigos y procedimientos y *iii)* realizar una traslación de los resultados obtenidos a un contexto clínico, con el objetivo de obtener una rentabilidad en términos de salud pública de nuestros estudios.

Para estas etapas, dado que los modelos construidos no son del todo sólidos o robustos, por los motivos que se han desgranado en el capítulo de *Discusión*, se deberían estudiar modelos alternativos basados en otras variables respuesta en el mismo *dataset*, o bien optar por escoger un *dataset* diferente para este procedimiento.

También sería interesante tener la oportunidad de analizar otros enfoques bioestadísticos que se incluyen en diferentes paquetes de R con el objetivo de realizar una integración exitosa de los datos ómicos y la construcción de modelos predictivos de calidad.

4. Glosario

Bioconductor: proyecto que engloba numerosas herramientas en forma de paquetes de R para el análisis de datos de alta resolución genómica.

***Biomarcador*:** *set* de información biológica medida y analizada mediante un test basado en tecnologías ómicas. Biomarcadores pueden ser medidas de macromoléculas (ADN, ARN, proteínas, lípidos), células, procesos que describan el normal o anormal funcionamiento de un organismo. Pueden ser analizados y detectados en circulación (sangre, linfa), en tejido o fluidos corporales (orina, saliva…).

Cross-validation: o validación cruzada. Método que permite valorar los resultados de un determinado *test* en función de las particiones *train*/*test* de los datos.

***Curva ROC*:** *Receiver Operating Characteristic*. *Plot* que permite evaluar la exactitud diagnóstica de los modelos, enfrentando los parámetros de sensibilidad y especificidad de la prueba.

DE: análisis de expresión diferencial.

DEGs: genes diferencialmente expresados entre dos condiciones (p.e. muestras tumorales / normales).

***Enet*:** *Elastic Net*. Modelo de regresión logística que combina las penalizaciones L1 y L2 de los métodos *lasso* y *ridge*.

Feature selection: proceso de selección de variables en un entorno de análisis de datos ómicos.

Lasso*: Least Absolute Shrinkage and Selection Operator*. Tipo de regresión que desarrolla una regularización L1, la cual añade una penalización igual al valor absoluto de la magnitud de los coeficientes. Algunos coeficientes de magnitud 0 pueden ser eliminados del modelo.

***mixOmics*:** paquete del entorno R para el estudio de datos ómicos de diferente naturaleza basado en un amplio rango de análisis multivariantes con un especial interés en la selección de variables.

Omics/ciencias ómicas: disciplinas científicas que comprenden el estudio de sets relacionados de moléculas biológicas (genes, transcritos, proteínas, metabolitos, etc.).

PCA: *Principal Component Analysis* - Análisis de Componentes Principales. Método multivariante que permite la reducción de las dimensiones de los datos mediante la descripción de los mismos gracias a nuevas variables o componentes que son independientes entre sí.

Pipeline: conjunto de pasos para desarrollar un análisis bioinformático plenamente reproducible.

Ridge: tipo de regresión que implementa una regularización L2, implicando una penalización que es igual al cuadrado de la magnitud de los coeficientes.

RMSE: *root-mean-square error*. es una medida de uso frecuente de las diferencias entre los valores (valores de muestra o de población) predichos por un modelo o un estimador y los valores observados.

SNFtools: paquete del entorno R que permite desarrollar un procesamiento de datos multi-ómicos construyendo una red de similaridad integrada por cada capa ómica utilizada.

sPLS: *sparse Partial Least Squares*. Método de regresión lineal multivariante que supone reducción de dimensiones y que permite enfrentarse al problema de un gran número de predictores (*p*), bajo tamaño muestral (*n*) y alta colinealidad entre predictores.

SummarizedExperiment: tipo de objeto, utilizado en el entorno *Bioconductor*, que almacena matrices rectangulares de resultados experimentales.

TCGA: *The Cancer Genome Atlas.* Repositorio de datos de diferente naturaleza (ómicos, clínicos…) para la caracterización de más de 20000 muestras tumorales relacionadas con 33 tipos de cáncer.

TCGAbiolinks: paquete de R del entorno *Bioconductor* que permite la descarga de datos desde TCGA y el procesamiento inicial de los mismos.

Workflow: flujo de trabajo. Ver descripción de "***Pipeline***".

5. Bibliografía

1. Yadav SP. The wholeness in suffix -omics, -omes, and the word om. *J Biomol Tech*. 2007;18: 277.
2. Debnath M, Prasad GBKS, Bisen PS, Debnath M, Prasad GBKS, Bisen PS. Omics Technology. En: Molecular Diagnostics: Promises and Possibilities. 2010; pp. 11–31.
3. Micheel CM, Nass SJ, Omenn G., editors. Evolution of Translational Omics. Evolution of Translational Omics. Washington, D.C.: The National Academies Press; 2012.
4. Hardiman G. An introduction to systems analytics and integration of big omics data. *Genes* (Basel). 2020;11.
5. Mishra N. Science of omics: Perspectives and Prospects for human health care. *Integr Mol Med.* 2016;3: 1–8.
6. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. *Cell*. 2011;144: 986–998.
7. Boja ES, Kinsinger CR, Rodriguez H, Srinivas P. Integration of omics sciences to advance biology and medicine. *Clin Proteomics*. 2014;11: 1–12.
8. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. *BMC Syst Biol*. 2014;8: I1.
9. De Sanctis G, Colombo R, Damiani C, Sacco E, Vanoni M. Omics and Clinical Data Integration. En: Integration of Omics Approaches and Systems Biology for Clinical Applications. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2018. pp. 248–273.
10. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights.* 2020;14: 7–9.
11. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S. A selective review of multi-level omics data integration using variable selection. *High-Throughput.* 2019;8: 1–25.
12. de Maturana EL, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorno L, et al. Challenges in the integration of omics and non-omics data. *Genes* (Basel).

2019;10.

13. Volkmann A, De Bin R, Sauerbrei W, Boulesteix AL. A plea for taking all available clinical information into account when assessing the predictive value of omics data. *BMC Med Res Methodol.* 2019;19: 162.
14. Kim J, Bowlby R, Mungall AJ, Robertson AG, Odze RD, Cherniack AD, et al. Integrated genomic characterization of oesophageal carcinoma. *Nature*. 2017;541: 169–174.
15. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 2016;44: e71.
16. Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, et al. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEX. *PLoS Comput Biol*. 2019;15.
17. Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: Tools, advances and future approaches. *J Mol Endocrinol.* 2019;62: R21–R45.
18. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11: 333–337.
19. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. 2017;13: 1–19.