UNIVERSIDAD NACIONAL DE CÓRDOBA

MAESTRÍA EN ESTADÍSTICA APLICADA

INTEGRACIÓN DE DATOS DE EXPRESIÓN GÉNICA ASOCIADOS A LÍNEAS CELULARES DE CÁNCERES: UN ENFOQUE UTILIZANDO LA METODOLOGÍA STATIS-ACT, MÉTODOS BIPLOT Y MINERÍA DE TEXTO

Prof. María Laura Zingaretti
Junio-2016

Director:
Prof. Dr. Jhonny Rafael Demey
Co-Director:
Prof. Dr. Julio Alejandro Di Rienzo

Agradecimientos

Al director de este trabajo, el Prof. Dr. Jhonny Demey por su confianza y guía constantes, por su paciencia y su infinita generosidad para enseñarme y especialmente, por transmitirme la pasión por esta disciplina.
Al Dr. Julio Di Rienzo por su confianza y por sus orientaciones, tanto en la realización de este trabajo como en los cursos de la maestría.
Al Dr. Cristóbal Fresno porque siempre ha estado para ayudarme.
A todo el cuerpo de profesores y trabajadores no docentes de la Maestría en Estadística Aplicada por su dedicación constante.
A la Universidad Nacional de Villa María, por haber financiado parte de mis estudios y por la formación que he recibido durante años en esta institución, tanto como estudiante como en mi tarea docente.
A las amigas que encontré en la maestría: Jime, Vale y Belu. Sin duda, ha sido una de las cosas más lindas de este camino.
A todos mis amigos, ¡gracias por estar siempre a mi lado! y por todo lo compartido, pues "¿de qué me sirve a mi la virtud, si no tengo un canto hermoso?"

Resumen

Los datos ómicos, donde se analiza la expresión de miles de genes, proteínas, metabolitos en distintas muestras biológicas, comprenden una gran cantidad de áreas de aplicación y su estudio ha crecido de manera exponencial en los últimos años. La integración de datos es uno de los principales objetivos actuales en bio-ciencias, dado que es común medir expresión simultánea de genes, proteínas y metabolitos o bien tener mediciones de la expresión genética de las mismas muestras en diferentes plataformas de análisis, generando, por lo tanto, distintas fuentes de información. La integración implica el meta-análisis de sus resultados o el análisis simultáneos de los datos originales.

En el marco de este trabajo, se realiza una propuesta metodológica para abordar el problema de estudio y comparación de datos genéticos provenientes de distintas plataformas de microarreglos y el problema de selección de genes basada en los métodos estadísticos de k-tablas y Biplot. Se realiza una aplicación de la metodología propuesta a datos de expresión genética de 60 líneas celulares de 9 tipos diferentes de cáncer provenientes de cuatro plataformas de análisis del panel NCI-60.

Palabras clave

Biplot; Datos ómicos; Microarreglos, k-tablas, STATIS; STATIS-Dual.

Índice general

Índice de figuras

Índice de cuadros

Introducción

Los métodos estadísticos clásicos permiten estudiar problemas en dos modos: individuos y variables. No obstante, numerosos problemas de diversas áreas de conocimiento poseen una naturaleza más compleja y requieren del estudio simultáneo de tres modos (individuos, variables y diferentes condiciones experimentales o temporales). Para su abordaje, se han desarrollado distintas propuestas metodológicas, entre ellas, las comprendidas bajo el nombre de "Métodos de análisis para matrices de tres vías"(des Plantes, 1976).

Estas técnicas se aplican cuando al menos uno de los modos es común a todos los estudios (individuos o variables). Es decir, posibilitan la integración de conjuntos de datos que desde el punto de vista de los métodos clásicos no son comparables, y por lo tanto, permiten establecer relaciones entre individuo-individuo, individuo-variable y entre variables a lo largo de las distintas condiciones (des Plantes, 1976; Lavit *et al.*, 1994).

Una de las áreas de aplicación de estas metodologías, que no ha sido suficientemente explotada, es en el estudio de datos de expresión génica derivados del estudio simultáneo de la información obtenidas a través de tecnologías de alto rendimiento, como los microarreglos de ADN, espectroscopía de masas, chips de proteínas y secuenciación masiva, entre otras, bajo diferentes tipos

de plataformas, tratamientos, condiciones experimentales, ensayos clínicos y grupos de investigación (Acharjee, 2013).

El abordaje clásico del estudio de datos de expresión está basado en el análisis de un ensayo, típicamente consiste en la identificación de genes candidatos a través del ajuste de modelos lineales gen a gen (Cui y Churchill, 2003) o el agrupamiento de genes por su perfil de expresión mediante análisis multivariantes tradicionales (Belacel *et al.*, 2006).

En este sentido, buscar la mejor estrategia para integrar la información provista por diferentes estudios es un desafío, no solo desde el punto de vista médico y/o biológico, sino desde el punto de vista de la metodología estadística a emplear para su análisis. Los métodos de k-tablas pueden ser una alternativa metodológica que permita la identificación de los genes asociados a diferentes tipos de enfermedades, como por ejemplo el cáncer. Donde entre otros, es importante explorar y determinar relaciones entre distintos tipos de tumores, encontrar nuevos subtipos de acuerdo a su comportamiento en la propensión a metástasis o respuesta a terapias o buscar genes candidatos que posibiliten la construcción de terapias universales. Es así como, la integración de información de diversos estudios sobre expresión génica es una herramienta potencialmente de gran utilidad.

El presente trabajo abordará la problemática estadística de la meta-experimentación, es decir de la integración de información experimental generada por diferentes fuentes, la cual será ilustrada con el estudio del panel de datos de Cáncer NCI-60, que disponen de información sobre respuesta a drogas, expresión génica, entre otras, de líneas celulares de 9 sub-tipos de cánceres humanos (Shankavaram *et al.*, 2009).

Objetivo general

Proponer un abordaje de análisis que posibilite la integración de datos provenientes de distintas fuentes de expresión génica, de líneas celulares de cánceres. El enfoque incluye metodología STATIS-ACT, métodos Biplot y herramientas de minería de textos, que permitan mejorar la interpretación y contribuyan a la comprensión de las relaciones existentes entre tipos de tumores, genes-tumores, y entre genes y las funciones que se activan o deprimen ante la presencia de la enfermedad.

Objetivos específicos

1. Realizar una revisión exhaustiva sobre el estado actual del arte de los métodos para medir expresión génica y los análisis estadísticos más utilizados en estudios del cáncer en humanos.
2. Proponer la aplicación del método STATIS-ACT como herramienta de integración de datos provenientes de diferentes fuentes de información generadas de estudios de expresión génica.
3. Proponer la aplicación del Biplot lineal para la proyección de los genes sobre el espacio consenso generado por el método STATIS-ACT.
4. Proponer una metodología para cuantificar la sensibilidad del método STATIS-ACT a través del estudio de la variabilidad y la calidad de representación de individuos y grupos.
5. Ilustrar el enfoque de análisis propuesto a través del estudio del panel de datos de cáncer NCI-60 y comparar los resultados obtenidos con los provistos por la literatura.

Para realizar los cálculos y representaciones gráficas, se desarrolló una librería en el software **R** (R Core Team, 2015), denominada **kimod** (Zingaretti *et al.*, 2015). Además, se utilizó el software **InfoStat** (Di Rienzo *et al.*, 2011).

1. Marco Teórico

1.1 Nuevas perspectivas en la investigación sobre cáncer

Según la Organización Mundial de la Salud (**OMS**) (OMS, 2015), cáncer es un término genérico que se utiliza para denominar un conjunto de enfermedades que se caracterizan por un crecimiento incontrolado de células anómalas en determinados tejidos del cuerpo humano y son susceptibles de invadir otros tejidos, proceso que se conoce con el nombre de metástasis. El cáncer, constituye una seria preocupación para la comunidad científica debido a su alta incidencia en la población, llegando a constituirse en una de las mayores causas de muertes anuales en todo mundo (Siegel *et al.*, 2014). Si bien existe desde hace siglos, el número de casos ha aumentado considerablemente en los últimos años. De acuerdo a la OMS, aproximadamente un 30 % de las muertes por cáncer se deben a cinco factores de riesgo comportamentales y alimentarios: índice de masa corporal elevado, consumo insuficiente de frutas y verduras, falta de actividad física y consumo de tabaco y alcohol, es decir que pueden prevenirse. Otra de las causas que contribuyen al incremento del número de casos es el aumento de la esperanza de vida de

la población, dado que la enfermedad tiene mayor impacto en personas de edad avanzada.

La enfermedad admite dos clasificaciones diferentes: por su lugar de origen o por el tipo de tejido. Según la Clasificación Internacional para Enfermedades Oncológicas (CIE-O), los tumores pueden agruparse en seis categorías: carcinoma, sarcoma, mieloma, leucemia, linfoma y tipos mixtos, a continuación, una breve descripción de las mismas (NCI, 2014c).

Carcinoma: es el tipo de cáncer más común, representando entre un 80 a 90 % del total de cánceres. Se origina en el tejido epitelial, que está presente en la piel y constituye el recubrimiento de algunos órganos internos. Este tumor se divide en dos subtipos: el adenocarcinoma (que se origina en algún órgano o glándula) y el carcinoma de células escamosas. Ambos se originan en muchas áreas del cuerpo, aunque la mayoría de ellos afectan órganos o glándulas que producen secreción y por esta razón tienen una alta frecuencia en pechos, pulmones, próstata, vejiga o colón.

Sarcoma: se origina en tejidos conjuntivos del organismo tales como articulaciones, grasa, huesos, cartílagos, vasos sanguíneos, entre otros. Ocurre con mayor frecuencia en adultos jóvenes. Estos tumores se asemejan al tejido en el cuál se desarrollan y provienen de las células que forman el mesodermo. Si bien pueden aparecer en varias partes del cuerpo, la mayoría se sitúan en las zonas de rodillas y tobillos. En su fase inicial suelen ser asintomáticos, lo que generalmente ocasiona un diagnóstico tardío, comprometiendo la supervivencia. Poseen baja incidencia en la población y son muy heterogéneos. Suelen generar metástasis en aproximadamente el 40 a 60 % de los casos, con un pobre pronóstico de sobrevida, dada la escasa efectividad de los

tratamientos existentes. Estos presentan una gran cantidad de subtipos y son muy variables, por lo que aún hay una escasa comprensión sobre ellos. El análisis de expresión génica de distintos grupos, puede contribuir a una mejor comprensión de la biología subyacente y al desarrollo de tratamientos específicos para cada caso particular.

Mieloma: es una neoplasia maligna de células plasmáticas, que produce una expansión incontrolada y acumulación de células monoclonales en la médula ósea. Si bien puede aparecer en forma de tumor (generalmente en el hueso), en la mayoría de los casos las células anormales no producen masas sólidas. Esta enfermedad representa aproximadamente el 1 % de todos los tipos de cánceres existentes. El único tratamiento disponible es la quimioterapia.

Leucemia: esta enfermedad comienza en los tejidos que forman la sangre, desarrollándose inicialmente en las células madre que residen en la médula ósea, provocando la producción de grandes cantidades de glóbulos blancos anormales que ingresan al torrente sanguíneo. Posee dos grandes clasificaciones, de acuerdo a la rapidez con que se desarrolla; la *leucemia crónica* se genera lentamente y es asintomática, dado que no impide el cumplimiento de las funciones de las células normales. Por otra parte, la *leucemia aguda*, donde el número de células anormales crece muy rápidamente, impidiendo el desarrollo de las funciones de las demás células. Es una enfermedad que tiene muchas opciones de tratamiento, que van desde la quimioterapia, radioterapia, combinación de ellas y el trasplante con células madre. Dado que afectan a las células y no generan una masa sólida, se considera un tumor circulante hematológico.

Linfoma: esta enfermedad se origina en los ganglios linfáticos, generalmente comienza con un fallo en los linfocitos, que son las células de la sangre encargadas de generar anticuerpos. Si bien es un tumor hema-

tológico, como la leucemia o el mieloma, se considera sólidos porque afecta zonas concretas del cuerpo. Se clasifica en dos subtipos: linfoma de Hodgkin y linfoma no- Hodgkin.

Tipos Mixtos: se utiliza este término para denominar aquellos tumores que poseen varios tipos de células tumorales. Algunos tumores dentro de esta clasificación son: el carcino-sarcoma, carcinoma adenoescamoso o el tumor mesodérmico mixto.

Como se ha mencionado, la enfermedad también puede clasificarse según la zona del cuerpo que afecta. En este punto, es preciso destacar que los cánceres que más inciden en mujeres son el de mama, pulmones y bronquios, colono-rectal, útero, tiroides y linfoma No- Hodking. En tanto que para los hombres, se tiene próstata, pulmones y bronquios, colono-rectal, vejiga, melanoma de piel y riñones, que para el año 2014 constituyen el 66 % (para cada sexo) del total de todos los nuevos casos de cáncer en EEUU (Siegel *et al.*, 2014). De acuerdo a estos autores, el cáncer de pulmón es el de mayor tasa de mortalidad en ambos sexos, seguido por los de próstata y colono-rectal en hombres; y mama y colono-rectal, en mujeres.

Según datos sumunistrados por la *International Agency for Research on Cancer* (Ferlay *et al.*, 2015), en el año 2012 se produjeron más de 8 millones de muertes por cáncer a nivel mundial y hubo 14 millones de nuevos casos que se proyectan aumenten a 22 millones en los siguientes veinte años. La Argentina, se encuentre dentro del rango de países con ocurrencia de cáncer media-alta dado que las estimaciones de la agencia indicaron una incidencia de 217 casos nuevos por año cada 100.000 habitantes en ambos sexos en el año 2012 para dicho país,comparado con la incidencia media a nivel mundial de 137.5 a 172.3 casos por cada 100000 habitantes. Predominan los cánceres de próstata (44 casos por cada 100000 habitantes) y pulmón (32.5 casos por

cada 100000 habitantes) en hombres; y mama en mujeres (con 71 casos por cada 100000 habitantes).

Debido a las causas expuestas, los estudios sobre el cáncer constituyen una de las áreas de investigación más importantes de los últimos tiempos y las revistas donde se trata este tema son las de mayor impacto. Por otra parte, debido al rápido avance de las tecnologías "ómicas", que miden la abundancia de ARNm (RNA mensajero), proteínas o metabolitos en un número pequeño de individuos (Goldstein y Guerra, 2010). Estos datos son generados por tecnologías de alto rendimiento, entre las que se encuentran espectrocopía de masas, chips de proteínas, RNA seq (secuenciamiento) y microarreglos de ADN.

Este trabajo, focaliza sobre datos de microarreglos, que son una colección de fragmentos de ADN de secuencia conocida. Luego, estos chips se hibridan con ADN complementario marcado, obtenido de una muestra biológica, para finalmente cuantificar la cantidad de ADN hibridado con cada fragmento del chip. La mayoría de los experimentos de microarreglos son motivados por el interés en algún sistema biológico o una enfermedad. En este aspecto, la investigación sobre el cáncer constituye una de las áreas más importantes, dado que los estudios podrían contribuir a identificar subtipos de cáncer y a mejorar la prognosis y diagnosis de los mismos (Parmigiani *et al.*, 2011). Por este motivo, el crecimiento de trabajos publicados sobre expresión genética en cáncer ha sido exponencial a lo largo de los últimos 35 años,ver Figura 1.

Entre los objetivos que persiguen estos estudios, se destacan los de encontrar genes responsables del desarrollo de distintos tipos de tumores, estudiar su biología y contribuir al descubrimiento de terapias alternativas (Moghaddas Gholami *et al.*, 2013). Particularmente, el programa de Desarrollo Te-

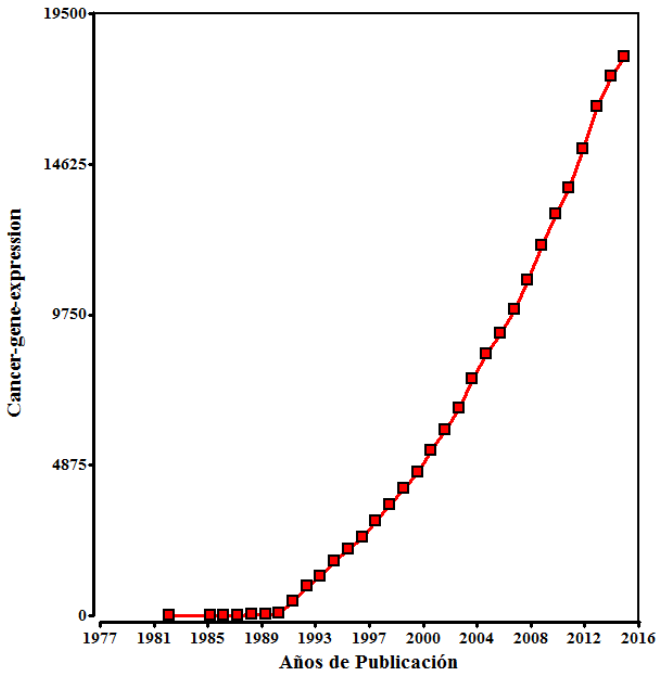


Figura 1: Evolución de la cantidad de trabajos publicados sobre: expresión genética en cáncer a lo largo de los últimos 36 años. Fuente: Reuters (2012).

rapéutico del Instituto Nacional del Cáncer de EEUU creó el panel de líneas celulares de NCI-60 para conducir un estudio sobre la respuesta a drogas de algunas de las clases de cáncer más frecuentes: colono-rectal, pulmón, mama, próstata, renal, de ovario, y los originados en el sistema nervioso central, como leucemias y melanomas (Shankavaram *et al.*, 2009). El panel contiene cientos de datos de ARNm, ADN, drogas y proteínas que fueron evaluados sobre líneas celulares de nueve sub-tipos de cáncer. Investigadores de todo el mundo dirigen sus esfuerzos a realizar un análisis integrado de éstos y de otros datos genéticos sobre la enfermedad.

Si bien, tradicionalmente el diagnóstico y pronóstico del cáncer se ha basado en un análisis morfológico complementado con el análisis de un único gen o proteína, el advenimiento de las tecnologías de microarreglos y la posibilidad

de evaluar la expresión de cientos o miles de genes o proteínas simultáneamente, generó un cambio de paradigma, con la capacidad de detectar múltiples bio-marcadores responsables del desarrollo de tumores (Ginsburg y McCarthy, 2001; Hood *et al.*, 2012). En consecuencia, el desafío de los estudios actuales es encontrar marcadores que se manifiesten simultánea e independientemente en distintas fuentes de datos y que permitan una clasificación más específica y generación de nuevos conocimientos sobre la enfermedad. Algunos investigadores sostienen que estas estrategias pueden aportar no sólo al descubrimiento de nuevos fármacos sino también al desarrollo de una medicina genómica personalizada (Joyce y Palsson, 2006).

El proceso de integración de datos en estudios genéticos de cáncer, posibilita identificar si una fuente de datos contiene más información y/o calidad, si existen bio-marcadores robustos o mostrar discrepancias en las distintas mediciones (Meng *et al.*, 2014). Generalmente, el grado de acuerdo entre estudios genéticos provenientes de distintas bases de datos es muy pobre, llegando a detectarse listas de genes estadísticamente significativos en distintas plataformas cuya intersección es mínima (Suárez-Fariñas *et al.*, 2005). Adicionalmente, el término integración también se utiliza en el sentido de analizar conjuntamente distintas fuentes de datos ómicos de un sistema particular (un organismo o una determinada enfermedad) que permite generar un conocimiento más completo del mismo (Zhang *et al.*, 2010).

La Figura 2, muestra la evolución temporal del número de investigaciones sobre integración de datos ómicos. Un análisis detallado de la figura, permite observar el creciente interés sobre el tema en los últimos años. Focalizando sobre integración de datos genéticos de cáncer (Figura 3) existen pocos trabajos publicados en la materia, aunque con una notable expansión en los últimos 4 años: con 2 publicaciones en 2012, 5 en 2013 y 2014 y 10 en 2015,

lo que sugiere la importancia que reviste en la actualidad el estudio sobre la temática.

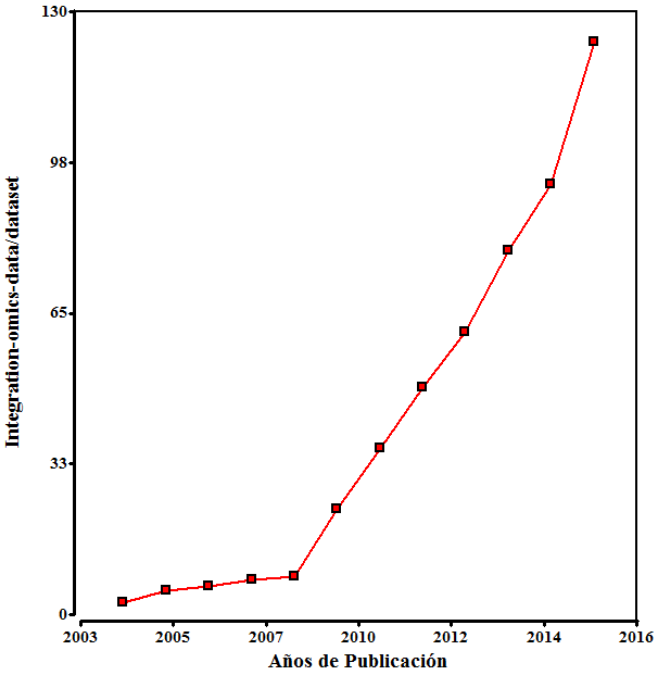


Figura 2: Evolución temporal de los trabajos publicados sobre integración de conjuntos de datos ómicos. Fuente: Reuters (2012)

1.2 Análisis de datos ómicos

Los ensayos de microarreglos generan grandes volúmenes de datos provenientes de distintas plataformas y laboratorios y una de las cuestiones más importantes es determinar si los experimentos son reproducibles (Goldstein y Guerra, 2010). Debido a la abundante cantidad de datos que estas tecnologías generan, el análisis presenta cierta complejidad.

En términos técnicos, los microarreglos contienen un conjunto de secuen-

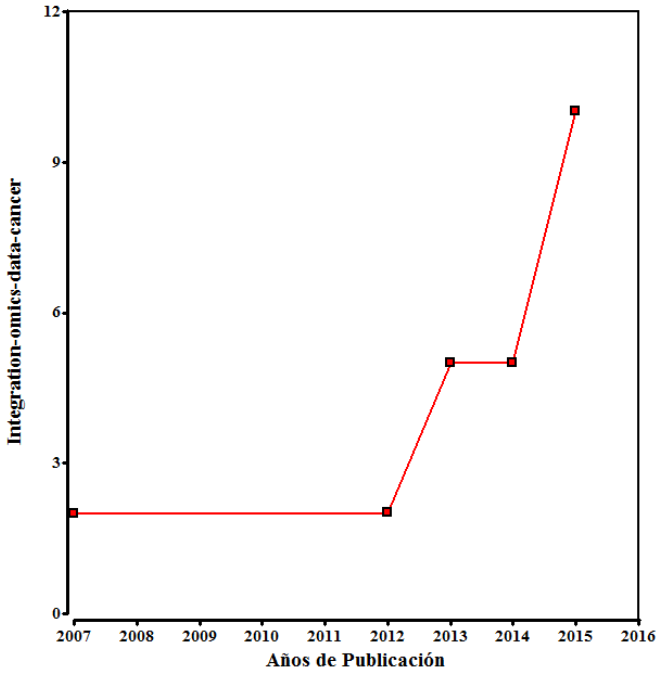


Figura 3: Evolución temporal de los trabajos publicados sobre integración de datos ómicos en cáncer. Fuente: Reuters (2012).

cias parciales de nucleóticos que permiten identificar inequívocamente un gen adherido a su superficie. Estos segmentos se conocen como *"probes"* o sondas (Wit y McClure, 2004). Para comprenderlos, es necesario conocer la estructura del ADN. Este se compone de dos cadenas de polímeros cuyas unidades básicas son los nucleótidos, que poseen una azúcar y una base nitrogenada. Hay cuatro tipos de bases nitrogenadas: adenina (A), citosina (C), guanina (G) y tiamina (T); cualquier ADN puede ser identificado por la secuencia lineal de éstas bases. La propiedad bio-molecular sobre la que se asientan las tecnologías de microarreglos es que el ADN está constituido por una doble cadena complementaria: cuando en una cadena hay una A en la otra hay una T y cuando hay una C, en la otra hay una G. Es decir, si una de las hebras de la cadena tiene la secuencia AATCGGT, entonces su cadena complementaria será TTAGCCA. La Figura 4 muestra la estructura

básica helicoidal del ADN.

Si bien hay distintas tecnologías de fabricación de chips, todas están basadas en el mismo principio. El chip es una estructura sólida, dónde se depositan miles de cadenas cortas de ADN diseñadas para identificar un gen y, posteriormente sobre ellas se hibridan las secuencias complementarias correspondientes, que se obtienen del ARNm de las muestras biológicas bajo análisis. Cuando un gen se exprese, lo hace transcribiendo su información desde el ADN que lo codifica a una secuencia complementaria de ARNm. Esta molécula viaja al citoplasma celular y allí participa como plantilla para la síntensis de proteínas. Luego la cuantificación de ARNm es una cuantificación de la actividad de un gen.

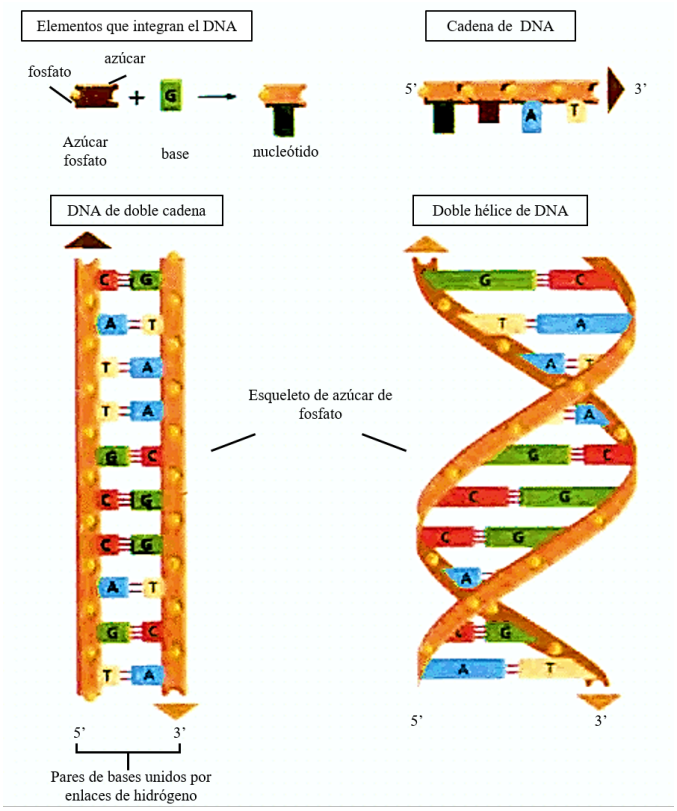


Figura 4: Estructura de doble hélice del ADN. Fuente: Alberts *et al.* (1996)

Las mediciones requieren una serie de pasos. En primer lugar, se marca con los fluorósforos Cy3 y Cy5 o sólo con uno de ellos, el ADNc (ADN complementario) que se genera con copia inversa del ARNm. La mayoría de los laboratorios usan marcado de fluorescencia con dos tintes (Cy3 y Cy5, que se activan diferencialmente a distintas longitudes de onda de un láser y que generan luminiscencia verde y roja respectivamente). De este modo, dos muestras son hibridizadas en los arreglos, una por cada tinte, lo que permite medirlas simultáneamente. En segundo lugar, se unen (mediante hibridación) las secuencias de la muestra con las secuencias complementarias que se encuentran inmovilizadas en el chip. La Figura 5 muestra el esquema básico para la obtención de los datos. En tercer lugar, los chips son leídos por un escáner que detecta la longitud de onda de los fluoróforos para la identificación de las secuencias. Finalmente, se generan los archivos de imágenes escaneadas que deben ser procesados con el fin de obtener una medida única de expresión para cada gen. Para experimentos de dos colores, generalmente se usa una expresión de intensidad relativa, en tanto que, para experimentos de un único color, se usa una expresión absoluta.

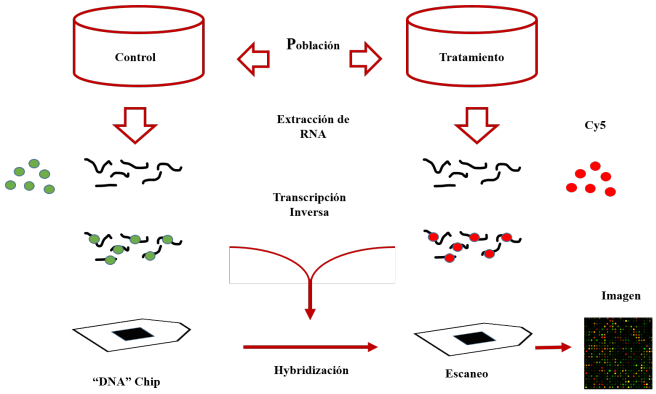


Figura 5: Esquema de un experimento de dos tintes.

Los datos ómicos son de naturaleza multivariada o multidimensional, con la característica de presentar siempre la medición de cientos o miles de varia-

bles sobre un pequeño número de individuos, que en el caso de los estudios sobre cáncer son muestras de células tumorales. Particularmente, en lo que respecta a los estudios sobre la enfermedad, la mayoría de los trabajos realizados implican la comparación de expresión genética de tejidos cancerígenos y normales con el fin de detectar genes que se expresen diferencialmente (Guo *et al.*, 2009; Ng *et al.*, 2009; Welsh *et al.*, 2001); otros comparan la expresión genética de tejidos correspondientes a distintos tipos de cáncer para establecer genes potencialmente ligados a cada uno de ellos y determinar subtipos (Lu *et al.*, 2005).

Respecto de las metodologías de análisis, la gran mayoría de las investigaciones publicadas sobre datos ómicos en general, y relacionados con el cáncer en particular, utilizan las mismas técnicas, entre las que se destacan: test t adaptado para estudios de microarreglos o modelos lineales mixtos cuando hay más de dos tratamientos (Cui y Churchill, 2003); algoritmos de clasificación (supervisada y no supervisada) tales como máquinas de soporte vectorial (Rakotomamonjy, 2003); redes bayesianas (Diaz-Uriarte, 2007); K vecinos más cercanos , conglomerados jerárquicos y no jerárquicos (Belacel *et al.*, 2006); modelos de regresión con selección de variables, usando mínimos cuadrados parciales (Nguyen y Rocke, 2002); Análisis de Componentes Principales, Correlaciones canónicas (Soneson *et al.*, 2010); entre otros.

Una problemática reciente, es la integración de la información provenientes de diversas fuentes de datos, dado que una enfermedad es un complejo de carácter multifactorial y para una mejor comprensión de los mecanismos que hay detrás de ella, es necesaria una aproximación que incluya distintas fuentes de información (Liu *et al.*, 2013b; Pastrello *et al.*, 2014).

Si bien se han utilizado algunas técnicas del análisis estadístico multiva-

riado, que permiten representar gráficamente individuos y variables (genes, proteínas) en un espacio de dimensión reducida facilitando la interpretación de la estructura de variabilidad global y la separación de la señal del ruido, éstas se restringen al análisis de dos tablas de datos, como máximo. En contraste, los métodos de integración de sub-espacios, que presentan una gran tradición en otras áreas y se utilizan para integración, combinación y visualización de datos, han tenido escaso desarrollo en este campo.

Las metodologías actualmente empleadas sólo permiten responder preguntas sobre agrupamiento de los diferentes tejidos tumorales basados en los perfiles de expresión, estableciendo relaciones entre observaciones o agrupamientos de genes (variables), sin la posibilidad de identificar relaciones individuos-variables, ni relaciones individuos- variables entre distintos estudios.

Por otra parte, muchos de los métodos utilizados requieren suposiciones sobre las matrices de datos que no siempre se cumplen y resultan inadecuados para trabajar con variables de distinta naturaleza (cuantitativa, mixta, ordinal), dificultando relacionar la información biológica con otra clase de información disponible: de carácter clínico, demográfico o cultural, que generalmente proviene de datos ordinales, nominales, binarios.

Los métodos más usados para la integración de la información proveniente de múltiples plataformas son: el método de Fisher o redes de co-expresión derivadas del uso del coeficiente de correlación. El primero realiza la integración desde una perspectiva del análisis de ontologías, a través del uso de pruebas estadísticas, que combinan múltiples probabilidades desde test independientes para generar un estadístico Chi-Cuadrado (Jia *et al.*, 2012); si bien el método es útil para obtener grupos de genes que comparten las mismas funciones biológicas y para determinar funciones enriquecidas o de-

preciadas por la presencia de una determinada enfermedad, no responde algunos interrogantes que podrían ser de interés de los investigadores,como por ejemplo, establecer relaciones directas entre genes y muestras biológicas; además, el procedimiento, no proporciona ninguna herramienta para visualizar la información, lo que dificulta la interpretación de los resultados. El coeficiente de correlación se utiliza para comparar el comportamiento de distintos conjuntos de datos o subconjuntos de ellos, por ejemplo: genes que se expresaron diferencialmente en cada plataforma, o genes que se expresaron diferencialmente con respuesta a determinados fármacos (Burkard, 2012). Resulta sensible al tamaño de la muestra y a la presencia de datos extremos. Adicionalmente, su uso requiere supuestos sobre distribuciones poblacionales y de que las configuraciones pertenecen a un mismo sistema de referencias (Demey, 2008).

Una aproximación mediante el uso de técnicas de integración de sub-espacios posee, entre otras, las siguientes ventajas:

1. reducir eficientemente la dimensión;
2. representar simultáneamente individuos y variables de distintos conjuntos de datos;
3. reducir los problemas de pequeño número de individuos, en comparación con el número de variables;
4. integración de datos en un espacio común e,
5. integración de datos de distinta naturaleza a partir del uso de métricas adecuadas, combinadas con otras técnicas de análisis.

A pesar de ello, pocos son los trabajos publicados sobre el tema que realizan el análisis desde esta perspectiva, entre los que se pueden mencionar, el

uso del análisis de co-inercia combinado con técnicas como componentes principales o correspondencias (Culhane *et al.*, 2003; Meng *et al.*, 2014).

A continuación, se discuten dos tópicos fundamentales en el análisis: los algoritmos utilizados en el pre-procesamiento de los datos de microarreglos y, la anotación de los mismos. En este sentido, Goldstein y Guerra (2010) señalan que sin un proceso de normalización adecuado, distintos arreglos no son comparables. Este es un problema a tener en cuenta cuando el objetivo central de un trabajo es integrar información desde múltiples fuentes de datos provenientes del mismo o de distintos fabricantes. En la actualidad hay diversos fabricantes que realizan estos experimentos, entre las más importantes se encuentran *Affymetrix*, *Illumina* y *Agilent*. Las plataformas proveen un aporte significativo para realizar comparaciones entre y dentro de los ensayos, siempre que cada conjunto de datos sea adecuadamente normalizado.

Para el caso de estudio de una enfermedad particular, por ejemplo el cáncer, las distintas plataformas miden expresión genética sobre un mismo conjunto de muestras, por lo que los datos disponibles son de carácter multi-vía (ver Figura 6), siendo los tejidos una configuración común a todas las tablas, en tanto que los genes pueden o no coincidir.

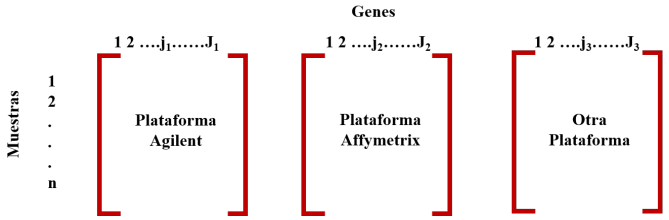


Figura 6: Esquema del conjunto de datos obtenidos a partir de varios experimentos de microarreglos. Se mide variabilidad genética sobre las mismas muestras.

Desde un punto de vista estadístico, las tablas de datos de expresión genética son matrices rectangulares X de orden $n \times p_k$, donde n son las filas que corresponden a los individuos (en el caso de cáncer pueden ser tejidos tumorales de distintas clases, por ejemplo) y, p_k las columnas que indican la expresión de un determinado gen o proteína en un individuo particular; x_{ij} indica la expresión del j-ésimo gen o proteína en el i-ésimo individuo.

Como es habitual medir expresión genética de un mismo conjunto de muestras en distintas plataformas, los datos pueden verse como un cubo (ver figura 7).

Combinar información de distintas fuentes genera la posibilidad de tener una visión más completa del sistema bajo estudio, sin embargo el proceso de integración requiere una inmensa atención sobre la representación de la información, anotación y el soporte (Pastrello *et al.*, 2014). Por esta razón, es imprescindible desarrollar una metodología de análisis que incluya: organización de los datos, control de calidad, normalización, precisión en la anotación y ontologías de genes que posibilite tener un vocabulario controlado.

Un paso de crucial importancia en el análisis consiste en obtener una correcta anotación de los genes. Los datos que aportan los fabricantes sólo tienen información de las sondas, con el problema adicional de que cada uno utiliza una nomenclatura específica. Un inconveniente adicional es que no existe una relación biunívoca de sondas a genes por lo que, una sonda podría reportar más de un gen y viceversa (Burguillo *et al.*, 2010).

Existen diversos organismos y consorcios que centran sus esfuerzos en mantener un vocabulario controlado que permite conservar funciones y anotaciones

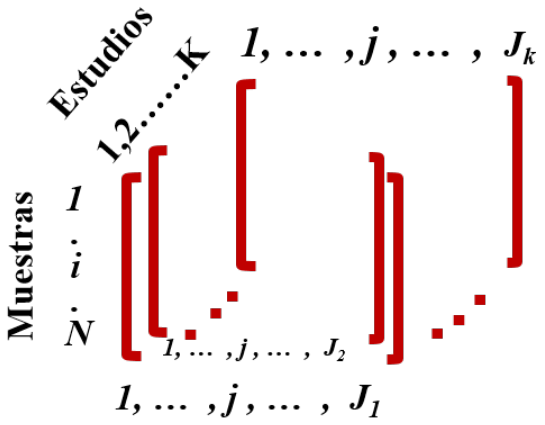


Figura 7: Esquema del conjunto de la matriz de datos generada a partir de varios experimentos de microarreglos realizados sobre las mismas muestras.

de genes específicos, entre los que se pueden nombrar: *LocusLink*, *Protein Information Resource* (**PIR**), *GeneCards*, *Proteome*, *Kyoto Encyclopedia of Genes and Genomes* (**KEGG**), **Ensembl** y **Swiss-Prot** (Dennis Jr *et al.*, 2003). Estas organizaciones proporcionan una excepcional cobertura y profundidad de datos funcionales disponibles para un determinado gen, pero no están diseñados para explorar eficazmente conocimientos biológicos asociado con cientos o miles de genes en paralelo. Por esta razón, Dennis Jr *et al.* (2003) crearon el portal **DAVID** (Base de datos para Anotación, Visualización y Descubrimiento Integrado) que proporciona un conjunto de herramientas de minería de datos que promueve el descubrimiento a través de la clasificación funcional y el acceso a fuentes de anotación biológica y que resuelve el problema de la disparidad de dimensiones (Dennis Jr *et al.*, 2003; Huang *et al.*, 2007). Además, este portal cuenta con una herramienta muy completa que permite distintas conversiones entre identificadores de genes.

Uno de los principales problemas de la anotación se debe al gran número

de genes con los que se trabaja, que imposibilita realizar un control manual de la nomenclatura. Zeeberg *et al.* (2004) advierten sobre la posibilidad de introducción de sesgo en los resultados de un experimento por el hecho de tener una anotación inadecuada. A modo de ejemplo, es posible mencionar que muchos de los identificadores universales de genes son transformados por el uso de planillas de cálculos (por ejemplo Excel) en un formato de fecha, lo que, por supuesto, resulta inoportuno para su tratamiento.

Particularmente, en caso que se desee comparar resultados desde distintos experimentos o integrar información, como en este trabajo, el proceso de conversión de los identificadores no sólo es imprescindible sino que, además, es tedioso y no trivial (Alibés *et al.*, 2007). Uno de los identicadores más utilizados es el *Universal Gene Name*: el Comité de Nomenclatura **HUGO** dependiente del Instituto de Bioinformática Europeo (**EMBL-EBI**) es el responsable de proveer una nomenclatura única para todos los genes del genoma humano (Gray *et al.*, 2014). Casi todos los símbolos y nombres de este repositorio son curados manualmente y se utilizan en la gran mayoría de las bases de datos que se centran en genes y proteínas humanos tales como Ensembl, UniProt, NCBI Gen, entre otros. Si bien los símbolos de genes oficiales actuales no son únicos entre las distintas especies (Kohl *et al.*, 2014), este comité es el encargado de proveer la anotación universal específicamente de la especie humana que se denomina **HGNC**, que provee un nombre corto para cada locus de gen humano conocido, así como también un nombre más largo que es descriptivo del mismo (Povey *et al.*, 2001).

Por estas razones, los creadores del repositorio CellMiner suministran bases de datos que son consistentes con las anotaciones de *Universal Gene Name* (alias, locación cromosómica, identificadores de secuencias de genes y proteínas y ubicación de las secuencias en el genoma) (Shankavaram *et al.*,

2009). Aunque los responsables del consorcio realizan un gran esfuerzo por normalizar la nomenclatura de los símbolos, estos son modificados constantemente, por un lado por los nuevos descubrimientos y, por otro, porque frecuentemente se determina que el primer atributo reconocido de un gen, que es la base de su nombre, no es finalmente un aspecto esencial del mismo o existe otra nomenclatura que resulta más racional desde una perspectiva biológica. Por lo que es recomendable actualizar las conversiones cuando se deseen hacer nuevos análisis.

A pesar de las dificultades mencionadas en el párrafo anterior, pero teniendo en cuenta que en el presente trabajo se utilizan datos correspondientes únicamente al genoma humano y, para generar consistencia con las anotaciones del repositorio donde se encuentran almacenados los datos (Reinhold *et al.*, 2012), se adopta como nomenclatura para la identificación de los genes el *Universal Gene Name*.

Por otro lado, debido al inconveniente mencionado de inexistencia de relación biunívoca entre sondas y genes, es posible que un mismo gen aparezca en varias ocasiones. Hay distintas posturas para dar solución a este problema y tener finalmente una única muestra de gen en cada tabla. En el marco de este trabajo, se usa la siguiente: se calcula la expresión media de cada gen y en caso de haber múltiples sondas para un mismo gen, se utiliza como representante la mayor de ellas.

1.2.1 Normalización de datos

El objetivo de la normalización de los datos es remover efectos de variaciones sistemáticas tales como, la diferencia en la preparación de muestras,

distintos niveles de intensidad de escaneo (Goldstein y Guerra, 2010). Este proceso permite reducir la variabilidad que no tiene origen biológico, y adicionalmentees utilizado en el sentido habitual del término, es decir para llevar todas las expresiones génicas a una escala comparable.

Wit y McClure (2004) indican que normalizar los arreglos de dos colores es más importante que los de un único color, dado que la metodología es, en gran medida, no estandarizada y existe correlación potencial entre los canales.

Una de las técnicas más utilizadas para normalizar es la Corrección de *Background*. Una vez que se obtienen los archivos de imagen escaneadas, se realiza el proceso de segmentación de las mismas, que divide cada rejilla en dos regiones: *foreground* (intensidad de la señal) y *background* (intensidad de fondo). Este último constituye un problema común en medición de señales ópticas (Wit y McClure, 2004). Se han propuesto varios métodos para contrarrestar este efecto, la mayoría de los cuáles se basan en un supuesto de aditividad:

$$S = B + T \tag{1}$$

donde S es la señal observada, B es la señal de fondo y T es la verdadera señal. No obstante, con los métodos tradicionales es imposible observar el *background* del *spot* y sólo se observa la medición de un *background* cercano al mismo (B^T). Los métodos más simples de corrección usan este valor para obtener una estimación de T:

$$\hat{T} = S - B^T \tag{2}$$

$\hat{T}$ (2), podría ser negativa, lo cual carece de interpretación en términos biológicos y no permite las posteriores transformaciones a escala logarítmica. Por lo tanto, en algunos casos, todas las intensidades que resultan menores o iguales a un número positivo fijo, son reemplazadas por este valor. Por ejemplo, en el paquete **limma** (Ritchie *et al.*, 2015) de **R** (R Core Team, 2015), la función que implementa corrección de *Background*, establece que todos los valores de intensidad observada menores o iguales a 0.5, se reemplazan por dicho valor.

Otro método de normalización se denomina **RMA** (*Robust Multiarray Average*), el cuál es utilizado en el marco de este trabajo. Fue sugerido por Irizarry (2003) para *chips Affymetrix* aunque su uso puede extenderse a datos provenientes de otras plataformas. Básicamente, el método asume distribución común de todas las intensidades de sonda y consta de tres pasos: (i) corrección de fondo, (ii) normalización de cuantiles y (iii) resumen del conjunto de sondas. La ventaja que posee respecto de otras técnicas es que no requiere el uso de un chip de referencia para la normalización.

El primer paso consiste en aplicar la corrección de *Background* que se describió en el párrafo anterior, usando sólo valores de intensidad positivos o a lo sumo, cero.

El segundo paso, radica en aplicar un algoritmo de normalización de cuantiles y asume que la gran mayoría de genes no se expresan diferencialmente y tienen, por lo tanto, un comportamiento basal. Para ello, las columnas de las matrices de datos son ordenadas en forma ascendente, obteniéndose una

matriz denominada X_{ord}, cuya i-ésima fila se denota como: $q_{i1}, q_{i2}, ..., q_{ip}$. El valor de normalización de esta matriz se define a través del producto interno de la ecuación 3:

$$\frac{\langle q_i, d\rangle}{\|d\|^2} \tag{3}$$

con $d = (1/\sqrt{P}, ..., 1/\sqrt{P})$ que es el vector director del plano donde todas las columnas son iguales, es decir, donde se cumplen las asunciones de igual distribución. Así, la fila i de la matriz viene dada por:

$$proj_{dq_i} = \frac{\langle q_{i,d}\rangle}{\|d\|^2} = \frac{1}{\sqrt{P}\sum_j q_{ij} d} = \Big(\frac{1}{P}\sum_j q_{ij}, ..., \frac{1}{P}\sum_j q_{ij}\Big) \tag{4}$$

Luego de este proceso se obtiene una matriz denominada $X^T ord$, que es devuelta a la ordenación original (matriz X_{norm}). Finalmente, todos los valores de la matriz X_{norm} son ajustados a escala logarítmica, obteniéndose una nueva matriz (Y).

El último paso, consiste en resumir el conjunto de sondas. Para la tecnología *Affymetrix*, un gen está representado por varias sondas cortas y en tal caso, este paso, resume las sondas en una única aplicando el concepto de afinidad. En otras tecnologías como *Agilent*, las sondas son únicas y no hay sumarización, excepto para aquellos genes que estan repetidos y cuyas sondas repetidas son idénticas; en este caso, la sumarización consiste simplemente en promediar o tomar la mediana (Irizarry, 2003).

1.3 Ontología de genes

Desde la filosofía, una ontología hace referencia a un sistema de conocimientos (Goldstein y Guerra, 2010). El término fue introducido por Aristóteles para describir el estudio del ser (Mayer *et al.*, 2014); en los últimos años, ha adquirido una creciente importancia en el contexto de la bioinfomática. Una ontología se refiere a la presencia de un vocabulario controlado con relaciones bien definidas que conectan los términos y que generalmente puede ser representada matemáticamente como una estructura de grafos (Goldstein y Guerra, 2010; Aibar *et al.*, 2015).

Dado que el origen evolutivo es común a distintas especies, aún en aquellas que tienen formas y comportamiento diferentes, es posible obtener un vocabulario único que caracterice los roles de los genes y su papel en los procesos vitales. Esto permite comparar datos provenientes de distintas fuentes y facilita la reproducción de los resultados del análisis.

En este sentido, después de la identificación de genes candidatos, es necesario evaluar cómo responde el sistema biológico en su conjunto sobre distintos procesos o funciones biológicas conocidas (por ejemplo, aquellos involucrados en el desarrollo de tumores). Para ello se consultan las ontologías , donde se almacena la información funcional disponible a nivel de genes. De acuerdo a Wei *et al.* (2015), combinar herramientas de minería de datos, aumenta las probabilidades de identificar procesos biológicos y genes candidatos que permitan tener un mejor conocimiento de determinadas enfermedades.

En esta tesis, se adopta el siguiente esquema de trabajo: una vez identificados genes potencialmente importantes, que se expresan diferencialmente en tejidos con cáncer, se realizan búsquedas que permitan identificar si exis-

ten relaciones probadas o tratadas experimentalmente en estudios anteriores entre los genes seleccionados y la presencia de la enfermedad. Para ello, se buscan oncogenes, genes supresores de tumores y genes mutados en distintas bases de datos como Intogen (Gundem *et al.*, 2010) o COSMIC (*Catalogue of Somatic Mutations in Cancer*) (Bamford *et al.*, 2004; Forbes *et al.*, 2010).

Un oncogen es un gen que participa en el crecimiento de las células normales pero su forma ha tenido una mutación. Los oncogenes pueden hacer crecer las células cancerosas y, por lo tanto estar sobre-regulados en diferentes tipos de cáncer (Thomas *et al.*, 2012; Ibrahim *et al.*, 2011). Las mutaciones de los genes que se convierten en oncogenes pueden ser heredadas o pueden resultar de la exposición a sustancias del ambiente que causan cáncer. Por otro lado, un gen supresor de tumores (o anti-oncogen) es un tipo de gen que elabora una proteína denominada supresora de tumores, la cual ayuda a controlar la proliferación celular e induce la apoptosis del tumor y por lo tanto, puede ayudar a controlar la enfermedad (Ibrahim *et al.*, 2011). Además, las mutaciones (cambios en el ADN) en estos genes pueden conducir al cáncer.

Adicionalmente, se usa el *Cancer Gene Index* (NCI, 2014b); que tiene como objetivo proveer una fuente de datos que consiste en genes asociados experimentalmente con cánceres humanos y/o con fármacos. Cuenta con datos sobre 6955 genes humanos, 12000 términos relacionados al cáncer y 2180 compuestos farmacológicos. Sus anotaciones fueron extraídas mediante el uso de tecnologías de minería de texto de aproximadamente 90 millones de sentencias en 20 millones de *abstracts* seleccionados desde MEDLINE (NCI, 2014a) para identificar presuntos genes asociados a enfermedades que han sido validados manualmente por expertos. La Figura 8 muestra un esquema del proceso.

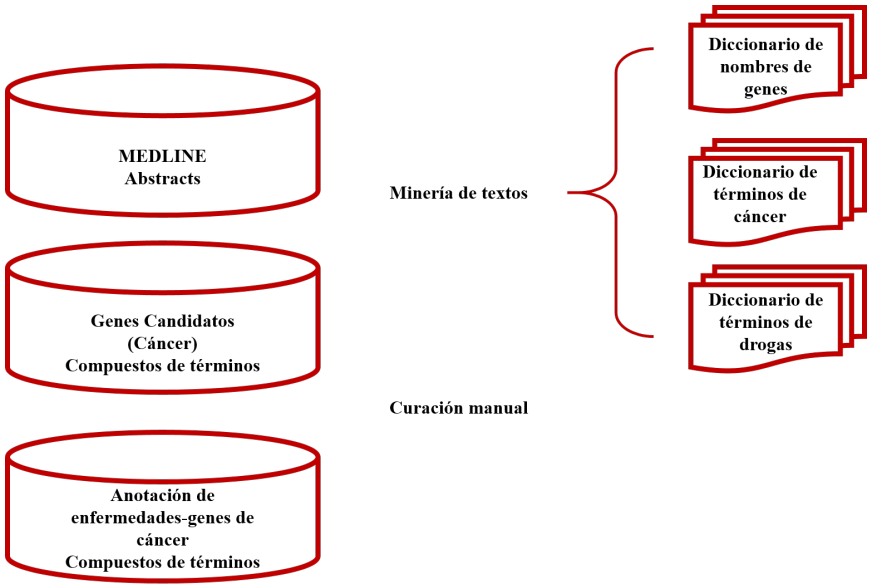


Figura 8: Esquema para obtener términos y genes asociados al cáncer.

Es posible contrastar la información sobre los genes seleccionados con la reportada en la literatura. La búsqueda en el *Cancer Gene Index* posee 13 categorías principales:

Behavior-Related-Disorder: un problema de comportamiento específico que se produce en los patrones persistentes y grupos característicos y que causa un deterioro clínicamente significativo.

Cancer-Related-Condition: se refiere a genes sobre los que fueron estudiados desórdenes asociados con un incremento del riesgo de padecer transformaciones malignas.

Disorder-by-Site: sitios donde se produjeron desórdenes.

Hamartoma: crecimiento excesivo y con patrones desorganizados de un tumor benigno de células maduras y tejidos normales.

Hyperplasia: incremento anormal en el número de células en un organismo o tejido con consecuente ampliación o engrosamiento.

Neoplasm: un crecimiento de tejido (benigno o maligno) que resulta de

una proliferación incontrolada de las células.

Non-Neoplastic-Disorder: un desorden que resulta de crecimiento anormal de tejidos resultando desde la proliferación descontrolada de las células.

Psychiatric-Disorder: desviación de la función normal del cerebro y que se traduce en un deterioro del normal funcionamiento cognitivo, emocional o conductual de un individuo y es causada por factores psicológicos o fisiológicos.

Radiation-Induced-Abnormalities: un trastorno no neoplásico o neoplásico que resulta de la exposición a la radiación.

Rare-Disorder: una enfermedad no maligna que afecta alrededor de 200000 personas en EEUU.

Polyp: una masa exofítica generalmente unida al tejido subyacente por una base amplia o un delgado tallo. Los pólipos pueden ser neoplásico o no neoplásicos.

Syndrome: no está definido en el diccionario.

Estas categorías contienen sub-categorías que están asociadas directamente a distintos tipos de cáncer y que permiten buscar genes asociados a la presencia de tumores. Además, el *Cancer Gene Index* permite buscar a través de la nomenclatura de cada tipo (por ejemplo: *breast*, *carcinoma*, *leukemia*, entre otros) y detectar los genes relacionados que fueron reportados en estudios previos. La última actualización es del año 2014.

Por otro lado, también se lleva a cabo un análisis de enriquecimiento funcional consultando las bases de datos de Gene Ontology (Ashburner *et al.*, 2000) y KEGG (Kanehisa y Goto, 2000) para establecer relaciones entre grupos de genes y términos enriquecidos.

El consorcio GO (Gene Ontology) se creó a finales del siglo XX con el objetivo de unificar una estrategia de recopilación y búsqueda de información genética (Ashburner *et al.*, 2000). Para ello, adopta tres vocabularios específicos que permiten reportar tres aspectos de la actividad genética (Goldstein y Guerra, 2010):

Procesos Biológicos (PB): esta ontología permite capturar el objetivo biológico al cuál determinado gen contribuye. Describe procesos de comunicación celular, estado biológico, entre otros. Solamente describe la variedad de procesos disponibles para una célula, sin ocuparse de las componentes bioquímicas necesarias para que éstos se lleven a cabo. Es la más desarrollada y utilizada en el área de la bioinformática.

Funciones Moleculares (FM): esta ontología representa el aspecto bioquímico de lo que produce un gen. Solamente describe lo que realiza, sin ocuparse del contexto de determinada reacción ni para qué sirve. Algunos ejemplos de ella son "quinasa", "enzima.º "transporte".

Componente Celular (CC): esta ontología permite describir la localización física del gen dentro de la célula. Es decir en qué región de ésta se establece: núcleo, membrana celular. También incluye términos que representan el complejo de multi-proteínas.

La **Enciclopedia de Genes y Genomas Kyoto (KEGG)**, fue creada en mayo del año 1995 y tiene como principal objetivo hacer corresponder información genómica con información funcional. En otras palabras, ligar un conjunto de genes con una red de moléculas interactuando a nivel celular, denominadas vías metabólicas (Kanehisa y Goto, 2000).

Tiene tres grandes bases de datos: una correspondiente a genes para distintas especies (29 en total), otra que vincula grupo de genes con vías metabólicas y una tercera, denominada ligandos, que tiene información sobre los compuestos químicos, encimas y reacciones enzimáticas.

El análisis de enriquecimiento funcional se realiza comparando la lista de genes candidatos con una lista de referencia (por ejemplo, la totalidad de genes considerados en el análisis) para encontrar términos biológicos enriquecidos. Esto significa que si, una determinada función tiene una probabilidad de ocurrencia de un 1 % en la lista completa y dicha probabilidad se incrementa a un 10 % en la lista de genes seleccionados, es necesario llevar a cabo análisis estadísticos para determinar si tales diferencias no son productos del azar. Entre los métodos de análisis más conocidos, se encuentran pruebas exactas de Fisher, pruebas Chi-Cuadrado, probabilidades binomiales y distribución Hipergeométrica. Para mayor detalle, consultar Huang *et al.* (2009) y Goldstein y Guerra (2010) .

1.4 Los métodos de k-tablas

El término k-tablas hace referencia a un conjunto de métodos exploratorios multivariados basados en el álgebra lineal y el espacio vectorial euclideano (Escoufier, 1973). Se desarrollaron para realizar análisis de datos multi- vía (Lavit *et al.*, 1994), y particularmente, permiten el tratamiento de datos cuando hay tres modos: individuos, variables y condiciones. Son generalizaciones del análisis de componentes principales (Abdi *et al.*, 2007, 2012) o el análisis de correlación canónica (Vivien y Sabatier, 2004), y para su aplicación, deben cumplirse algunas de las siguientes condiciones: tener varios conjuntos de variables, que fueron medidas en el mismo conjunto de indivi-

duos; poseer mediciones de las mismas variables sobre varios conjuntos de individuos o bien cuando un mismo conjunto de individuos y variables fue medido bajo distintos escenarios experimentales o a lo largo del tiempo. El esquema básico que representa a este tipo de datos es el mismo de la Figura 7. Entre estos métodos, es posible mencionar a los denominados STATIS y STATIS-Dual (des Plantes, 1976).

1.4.1 STATIS

El método STATIS se utiliza para realizar el análisis de un conjunto de k-tablas de datos cuantitativos que continen las mismas observaciones, es un acrónimo que significa en francés ***"Structuration des Tableaux 'aTrois Indices de la Statistique"*** y puede traducirse como: **"Estructuración de tablas estadísticas de tres vías"**. El objetivo es analizar la estructura de cada conjunto de datos individual para derivar pesos óptimos que permiten computar la mejor representación común de todas las tablas en un espacio euclideano. Se implementa en tres fases denominadas Interestructura, Compromiso e Intraestructura (des Plantes, 1976; Escoufier *et al.*, 1976; Lavit *et al.*, 1994).

1.4.1.1 Interestructura

El primer paso del análisis consiste en la evaluación de la similaridad entre todas las tablas a partir de la construcción de configuraciones de matrices. Sea $X_{[k]}$, la k-ésima matriz individual de dimensión $(I \times J_k)$, $M_{[k]}$, una matriz diagonal que contiene un conjunto de pesos de las variables para dicha matriz (que normalmente, son los mismos) y D una matriz diagonal con los pesos

de las observaciones a lo largo de todas las tablas. El análisis consiste en el estudio del triplete ($X_{[k]}$,$M_{[k]}$,D). Como se mencionó, en el STATIS, los elementos comunes a lo largo de todas las tablas son los individuos, por lo que, en primer lugar, se construyen k configuraciones de matrices simétricas $W_{[I\times I]} = X_{[k]}M_{[k]}X_{[k]}^T$.

La estructura de similaridad entre las matrices W se establece a partir de la definición del siguiente producto interno, que induce una norma:

$$\langle W_k|W_{k'}\rangle = tr(DW_kDW_{k'}) \tag{5}$$

este producto se denomina producto de Hilbert-Smith (**HS**). Geométricamente se interpreta como un producto escalar entre dos matrices semidefinidas positivas y es, por lo tanto, proporcional al coseno del ángulo entre las mismas. Adicionalmente, cuando estas matrices están normalizadas (es decir, la suma de sus elementos es igual a la unidad), define exactamente dicho coseno, que se conoce bajo el nombre de coeficiente de correlación vectorial **RV** y fue introducido por Escoufier (1973) para medir similaridad entre matrices cuadradas simétricas (Abdi *et al.*, 2012):

$$\rho_{k,k'} = \frac{\langle W_k|W_k\rangle_{HS}}{\langle W_k|W_k\rangle_{HS}\langle W_k'|W_{k'}\rangle_{HS}} \tag{6}$$

siendo la norma de cada una de las matrices:

$$\|W_k\|^2 = \langle W_k|W_k\rangle_{HS}$$

Debido a su definición, si el coeficiente es cercano a 1, las estructuras son congruentes. Es decir, no existen diferencias entre las estructuras factoriales

de las dos condiciones k y k'. En términos de STATIS, esto significa que los individuos se comportan de manera similar en ambas configuraciones.

Los productos escalares y las correlaciones se organizan en una matriz $S_{k\times k}$ semi-definida positiva, cuyos autovalores son reales y positivos o nulos y sus autovectores, ortogonales. Consecuentemente, por el teorema espectral, su auto- descomposición es:

$$S = Q\Delta Q^T \qquad con \qquad Q^T Q = 1 \tag{7}$$

y provee el Análisis de Componentes Principales de la estructura de similaridad entre las tablas que se representan como puntos bi o tri-dimensionales tomando las 2 o 3 primeras columnas de: $G = Q\Delta^{1/2}$. Por el teorema de Perron- Frobenius, todos los coeficientes del primer vector propio son del mismo signo y la proyección en las dos primeras dimensiones, tiene un aspecto como el que muestra la Figura 9.

En el ejemplo de la Figura 9, k es igual a 4 y las tres matrices representadas en el cuarto cuadrante poseen una estructura similar, pues los ángulos entre ellas son muy pequeños y sus cosenos son cercanos a 1. Además de la comparación entre estudios, el análisis provee los pesos óptimos que tienen cada una de las tablas sobre el compromiso y, que se calculan usando el valor correspondiente del primer autovector de la matriz S. Las tablas que más influencia poseen en la configuración son las que tienen el primer autovector más alto (en el ejemplo de la figura 9 son W_2,W_3 y W_4).

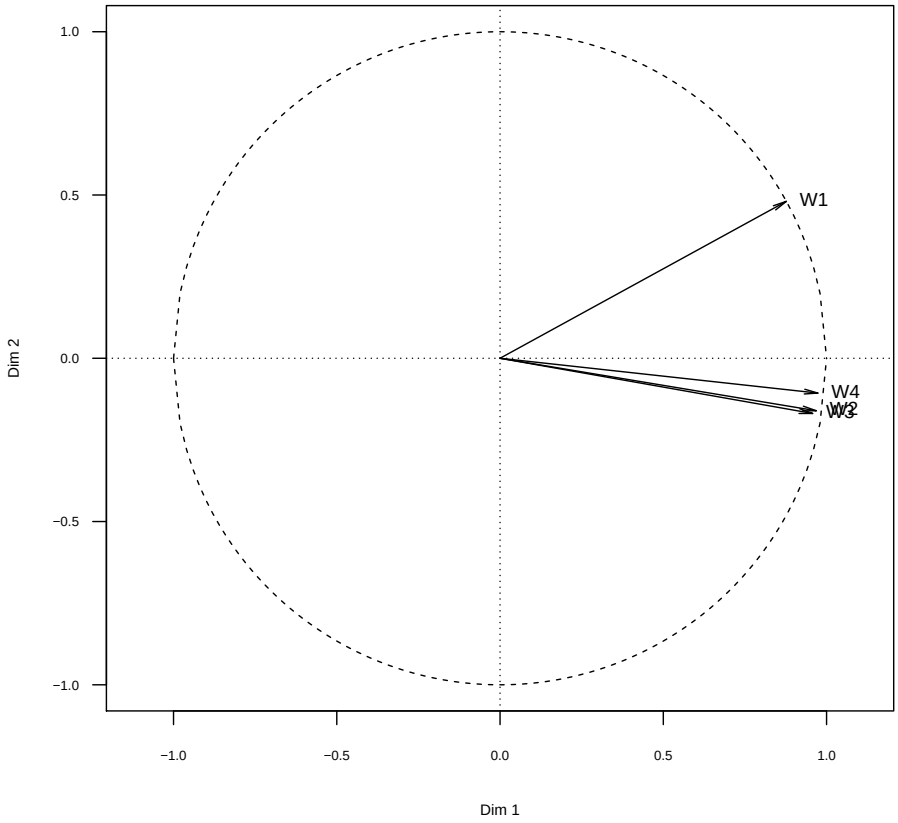


Figura 9: Proyección de la configuración euclídea de las tablas originales $W_k's$.

1.4.1.2 Compromiso

El objetivo de esta etapa es evaluar el comportamiento de los individuos a lo largo de todos las tablas. Para ello, se construye una matriz representante de las k configuraciones iniciales, denominada compromiso o consenso W que se define a partir de la siguiente suma ponderada (Lavit *et al.*, 1994):

$$W = \sum_{k=1}^{K} \alpha_k W_k \tag{8}$$

siendo:

$$\alpha_k = \frac{1}{\sqrt{\lambda_1^{(c)}}(\sum_{k=1}^{K} \sqrt{S_{kk}})} U_{1k}^{(S)} \tag{9}$$

$\lambda_1^{(c)}$ es el primer valor propio de la matriz W_k

$U_{1k}^{(S)}$ es la k-ésima componente del primer vector propio de la matriz S de correlaciones vectoriales.

El compromiso es la configuración más correlacionada (en el sentido del producto de **H-S**) con cada una de las configuraciones W_k. Para evitar la influencia de configuraciones que poseen norma elevada, en algunas ocasiones, se trabaja con las configuraciones normadas $W_k/|W_k|_{HS}$,obteniéndose:

$$W = \sum_{k=1}^{K} \alpha'_k \frac{W_k}{\|W_k\|_{HS}} \tag{10}$$

donde

$$\alpha'_k = \frac{1}{\sqrt{\lambda_1^{(c)}}} U_{1k}^{(S)} \tag{11}$$

siendo $\lambda_1^{(c)}$ el primer valor propio de la matriz W_k y $U_{1k}^{(S)}$ la k-ésima componente del primer vector propio de la matriz S de correlaciones vectoriales.

Dado que las matrices W_k son semidefinidas positivas y los coeficientes α_k (α_k') son positivos, la matriz W es semidefinida positiva y geométricamente se interpreta como la matriz de productos escalares que representa una configuración compromiso de los I individuos.

A partir de la configuración compromiso, es posible representar los I individuos en un espacio de dimensión dos o tres, utilizando la auto-descomposición de la matriz:

$$WD = \tilde{U}\tilde{\Delta}\tilde{U}^T \tag{12}$$

donde D es la matriz de pesos de los individuos. Para realizar dicha representación, se toman las 2 o 3 primeras columas de la matriz $F = \tilde{U}\tilde{\Delta}^{1/2}$, cuyas filas representan la imagen euclídea de cada uno de los I individuos y sus columnas son las componentes. La distancia entre dos puntos de la imagen euclídea del compromiso, representa la distancia media entre dichos individuos a lo largo de todas las tablas.

Esta metodología involucra un problema de optimización, que consiste en maximizar la norma cuadrática de W (Lavit *et al.*, 1994; Vivien y Sabatier, 2004; Abdi *et al.*, 2012):

$$v = \|W_k\|^2 = \langle W, W \rangle \qquad con \qquad a^T a = 1 \tag{13}$$

El criterio también puede expresarse a partir de su dual, que radica en minimizar la suma de las distancias entre el compromiso y cada una de las configuraciones de matrices W_k, como un problema de mínimos cuadrados en regresión.

$$\text{Đ} = \sum_{k=1}^{K} \|W_k - \alpha_k W\|^2 = \langle W, W \rangle \qquad con \qquad a^T a = 1 \tag{14}$$

1.4.1.3 Intraestructura

Dado que, a partir del compromiso es posible representar las posiciones "medias"de los individuos (ver ecuación 12), la configuración consenso (F) está dada por la ecuación 15.

$$\begin{aligned} WD &= \tilde{U}\tilde{\Delta}\tilde{U}^T \\ WD\tilde{U} &= \tilde{U}\tilde{\Delta}\tilde{U}^T\tilde{U} \\ WD\tilde{U}\tilde{\Delta}^{-1/2} &= \tilde{U}\tilde{\Delta}\tilde{\Delta}^{-1/2} \\ WD\tilde{U}\tilde{\Delta}^{-1/2} &= \tilde{U}\tilde{\Delta}^{1/2} \end{aligned} \tag{15}$$

ya que $\tilde{U}^T\tilde{U} = I$ por ser ortogonales.

A partir de 15, es posible construir matrices de proyección de los individuos de cada tabla sobre el compromiso:

$$F_k = W_k D\tilde{U}\tilde{\Delta}^{-1/2} \tag{16}$$

obteniendo una representación de las posiciones de los individuos dentro del espacio compromiso. Además, estas proyecciones permiten construir las trayectorias, que resultan de la unión de cada proyección individual a lo largo de las k- tablas. En el caso de tablas indexadas en el tiempo o el espacio, las trayectorias permiten visualizar la evoolución del fenómeno y se pueden nombrar dos clases de trayectorias: las envolventes y las excéntricas; las primeras indican que la diferencia entre el valor de la variable y la media es regular a lo largo de las condiciones, en tanto que las segundas, indican un cambio en la estructura del individuo a lo largo de las tablas.

Igualmente, también es posible construir la correlación entre las variables iniciales (J_k) de cada una de las tablas, con las componentes estándar que generan el espacio compromiso:

$$cor(J_k, \tilde{U}) = X_k^T D\tilde{U} \tag{17}$$

1.4.2 STATIS-Dual

El método STATIS- Dual (des Plantes, 1976; Lavit *et al.*, 1994) se utiliza para realizar el análisis de un conjunto de k-tablas de datos cuantitativos donde los elementos comunes en cada una de ellas son las variables. Se implementa en las mismas tres etapas que el método STATIS.

1.4.2.1 Interestructura

El primer paso consiste en la evaluación de la similaridad entre todas las tablas a partir de la construcción de configuraciones de matrices. Sea $X_{[k]}$ la k-ésima matriz individualde dimensiones $I_k \times J$, $D_{[k]}$, un conjunto de matrices diagonales que contienen los pesos de las observaciones para cada tabla y M una matriz diagonal con los pesos de las variables a lo largo de todas las tablas. El análisis consiste en el estudio del triplete ($X_{[k]}$,$D_{[k]}$,M) al partir del cuál se construyen k configuraciones de matrices $C_{[J\times J]} = X_{[k]}^T D_{[k]} X_{[k]}$.

La estructura de similaridad entre las matrices C se establece a partir de la definición del siguiente producto interno, que induce una norma y que en términos estadísticos indica la estructura de co-variación entre las variables:

$$\langle C_k | C_{k'} \rangle = tr(MC_kMC_{k'}) \tag{18}$$

que es el mismo producto de Hilbert-Smith (**HS**), definido en 5. Del mismo modo, el producto normalizado está dado por:

$$\rho_{k,k'} = \frac{\langle C_k | C_{k'} \rangle_{HS}}{\langle C_k | C_k \rangle_{HS} \langle C'_k | C_{k'} \rangle_{HS}} \tag{19}$$

cuya interpretación es la misma que la de la ecuación 6: si es cercano a 1, la estructura de co-variancián entre esos estudios es similar, siendo $\langle C_k | C_k \rangle_{HS} = \|C_k\|^2$.

Los productos escalares y las correlaciones se organizan en una matriz $R_{k\times k}$ semi-definida positiva, cuyos autovalores son reales y positivos o nulos y

sus autovectores, ortogonales. Consecuentemente, por el teorema espectral, su auto-descomposición es la misma que la referida en la ecuación 7 para el método STATIS que induce la misma representación geométrica de las tablas (ver figura 9).

1.4.2.2 Compromiso

El objetivo de esta etapa es evaluar el comportamiento de las variables de manera global (es decir, considerando cada uno de los k estudios). Para ello, se construye una matriz representante de las k configuraciones iniciales, denominada C, que se define a partir de la siguiente suma ponderada (Lavit *et al.*, 1994):

$$C = \sum_{k=1}^{K} \alpha_k C_k \tag{20}$$

α_k es el mismo que en la ecuación 9. Al igual que en STATIS, el compromiso es la configuración más correlacionada con cada una de las configuraciones C_k. Se define como una matriz de covarianzas promedio de las k tablas. También se puede trabajar con las configuraciones normadas para evitar influencia de configuraciones con normal elevada.

La configuración compromiso permite la representación de las J variables en un espacio de dimensión reducida, utilizando la auto-descomposición de la matriz:

$$CM = \tilde{U}\tilde{\Delta}\tilde{U}^T \tag{21}$$

para ello, se toman las primeras r columnas ($r < J$) de la matriz $V = \tilde{U}\tilde{\Delta}^{1/2}$, cuyas filas representan la imagen euclídea (consenso) de cada una de las variables.

1.4.2.3 Intraestructura

A partir de la configuración consenso de las variables ($V = \tilde{U}\tilde{\Delta}^{1/2}$), se pueden construir matrices de proyección para cada tabla, tal como lo indica la ecuación 22. Este análisis,permite representar todas las variables en un espacio común y da cuenta de los cambios y variabilidad de cada una a través de las distintas configuraciones.

$$V_k = C_k M \tilde{U}^T \tilde{\Delta}^{-1/2} \tag{22}$$

1.4.3 Análisis Parcial Triádico

El análisis Parcial Triádico (o X-STATIS) (Jaffrenou, 1978) es un caso particular de los denominados "Análisis de Tablas Múltiples"(Thioulouse y Chessel, 1987). Al igual que en los dos métodos descriptos anteriormente, el análisis se realiza en tres etapas: Inter-estructura, Compromiso e Intra-Estructura.

El método se utiliza cuando las k-tablas ($X[k]$) fueron medidas sobre las mismas observaciones y variables (Abdi *et al.*, 2012). La única diferencia que posee con los métodos anteriores es que en lugar de trabajar con operadores de matrices, utiliza directamente las matrices de datos. En este método, el compromiso es obtenido como una suma ponderada de las tablas originales

y tiene el mismo orden que ellas: $I \times J$.

El estudio de la Interestructura se basa en el concepto de "Vector de varianza" "vector de covarianza" (Escoufier, 1973). Se construye una matriz de producto escalar entre las distintas tablas como lo expresa la ecuación 23:

$$Covv(X_{[k]}, X_{[l]}) = trX_k^T D_I X_l Q_J \tag{23}$$

Siendo $X_{[k]}$, $X_{[l]}$, las k y l-ésimas tablas respectivamente. D_I la matriz de pesos de los individuos y Q_J una métrica en el espacio de las J variables. A partir de la ecuación 23, es posible obtener los coeficientes de correlación vectorial, que tienen las mismas interpretaciones que en el método STATIS o STATIS-Dual.

Como se mencionó, en este método, el compromiso es una combinación lineal de las tablas originales, pesadas por la componente del primer autovector resultante de la auto-descomposición de la matriz de correlaciones vectoriales (Thioulouse y Chessel, 1987; Thioulouse *et al.*, 2004; Thioulouse, 2011).

Finalmente, y al modo usual, la intraestructura se obtiene proyectando las filas y columnas de cada tabla de la serie en el compromiso, utilizando un análisis de componentes principales.

1.5 Medidas de calidad de representación

En los métodos multi-tabla, es posible incluir el cálculo de medidas de calidad de representación en un espacio de dimensión reducida (como el que se muestra en la figura 9), tanto para las k configuraciones iniciales , como para la proyección de los individuos y/o variables (según sea el método empleado) sobre la matriz compromiso.

1.5.1 Calidad de representación de las tablas

Para proyectar las tablas sobre un espacio de menor dimensión se realiza un Análisis de Componentes Principales, que resulta de la auto-descomposición de la matriz $S_{k\times k}$ de productos escalares (7), por lo que la calidad de representación de una tabla en las dos primeras dimensiones, viene dada por:

$$CRT = \frac{s_1^2 + s_2^2}{\sum_{k=1}^{K} s_k^2} \tag{24}$$

donde (s_1, s_2) son las coordenadas de la tabla en las dos primeras dimensiones y $\sum_{k=1}^{K} s_k^2$ representa las sumas de los cuadrados de todas las coordenadas del objeto. Es decir, la ecuación 24 indica la proporción de la información de dichas matrices recogidas por los dos primeros ejes. Si CRT es cercano a 1, significa que dicha tabla está muy bien representada en ese espacio.

1.5.2 Calidad de representación de los individuos

En el STATIS, los individuos son los elementos comunes a todas las tablas y se proyectan en el compromiso usando la auto-descomposición de la matriz

WD de la ecuación 12. Las coordenadas medias de los individuos a través de las k configuraciones, vienen dadas por $F = \tilde{U}\tilde{\Delta}^{1/2}$; así dos o tres primeras columnas de la matriz F constituyen las proyecciones de los objetos en el espacio compromiso de dimensión reducida. Del mismo modo que se explicó en la sección anterior, es posible calcular la calidad de representación del individuo i en el plano compromiso como:

$$CRT_i = \frac{f_{i1}^2 + f_{i2}^2}{\sum_{s=1}^{N} f_{is}^2} \tag{25}$$

donde (f_{i1}, f_{i2}) son las coordenadas del individuo en las 2 primeras dimensiones y $\sum_{s=1}^{N} f_{is}^2$ representa la suma de cuadrados de todas las coordenadas del objeto.

1.5.3 Calidad de representación de las variables

En el STATIS-Dual, los elementos comunes a todas las tablas son las variables y se proyectan en el compromiso usando la auto-descomposición de la matriz CM. Las coordenadas medias de las variables a través de las K configuraciones, son: $V = \tilde{U}\tilde{\Delta}^{1/2}$, cuyas dos primeras columnas constituyen las proyecciones de los variables en un espacio de dimensión 2. La calidad de representación de la j-ésima variable en el plano compromiso se calcula como:

$$CRT_j = \frac{v_{1j}^2 + v_{2j}^2}{\sum_{m=1}^{N} v_{mj}^2} \tag{26}$$

donde (v_{1j}, v_{2j}) son las coordenadas de las variables en las 2 primeras di-

mensiones y, $\sum_{m=1}^{N} v_{mj}^2$ representa la suma de cuadrados de todas sus coordenadas.

Se debe notar que el numerador de las ecuaciones 24, 25 y 26 siempre será menor o igual que el denominador. Es más, sólo se alcanzará la igualdad cuando K, P o N sean 2. Además, tanto el numerador como el denominador siempre serán números positivos, por ser sumas de cuadrados. Por lo tanto:

$$0 \leq CRT \leq 1 \tag{27}$$

multiplicando por 100, se obtienen los valores en porcentajes. Geométricamente, debe ser interpretada como el coseno cuadrado del ángulo entre el vector en el espacio completo y su proyección en el espacio de representación.

1.6 Variabilidad muestral

Siguiendo a Demey (2008) , los resultados de cualquier análisis de datos no están completos si no ofrecen información sobre la estabilidad de la solución, que permiten comprobar si la estructura detectada por el análisis no es producto del azar. Existen varias vías para cumplir con este propósito, incluyendo la introducción de pequeñas perturbaciones en los datos, las técnicas de remuestreo o la aplicación de permutaciones.

Al igual que para otras técnicas de ordenación, el estudio de sensibilidad de las soluciones en los métodos de k-tablas prácticamente no ha recibido atención. Por lo tanto, como parte de este trabajo, se estudia la estabilidad muestral de las proyecciones medias de los individuos y las variables sobre las matrices compromiso de los métodos STATIS o STATIS Dual.

Formalmente, se miden las alteraciones o perturbaciones que producen modificaciones sobre la matriz W (C) del compromiso (o sobre alguna de sus transformaciones). El procedimiento consiste en eliminar un elemento, cambiarlo, o simular errores en los datos para comprobar o verificar la estabilidad respecto a la configuración inicial.

En su trabajo de tesis doctoral, Demey (2008), introduce el uso de técnicas de remuestreo para estudiar la estabilidad de los resultados generados por el Análisis de Coordenadas Principales. Es posible extender la aplicación de estas ideas sobre otros métodos de ordenamiento y particularmente, en la matriz compromiso generada por los métodos abordados en el presente trabajo.

Los métodos de remuestreo que más se han utilizado para el análisis de estabilidad de matrices de datos que involucran la auto-descomposición son el *jackknife* (Tukey, 1958) y el *bootstrap* (Efron y Tibshirani, 1994; Lebart, 2007), que permiten la construcción de regiones de confianza para los individuos o variables proyectados sobre las dos o tres primeras dimensiones del compromiso sin el conocimiento previo de su distribución. Además, pueden usarse para generar intervalos de confianza para el porcentaje de inercia retenido por las distintas dimensiones.

Con el fin de construir las regiones de confianza para los individuos proyectados sobre la matriz compromiso, en esta tesis se utiliza el *bootstrap*. Esta técnica tiene la desventaja de que cada nueva matriz puede generar planos principales con direcciones diferentes, por lo que para corregir esto, es necesario generar transformaciones sobre las B configuraciones obtenidas.

De acuerdo a Lebart (2007) existen tres tipos de transformaciones. La pri-

mera, consiste en corregir las reflexiones sobre los ejes haciendo cambios de signos (si son necesarios) en las nuevas coordenadas. La segunda transformación, asigna secuencialmente los ejes originales a los derivados del remuestreo donde tengan correlación máxima y la última transformación se denomina Procrustes y consiste en superponer, tanto como sea posible, la configuración original y las generadas por el muestreo a través de una traslación, rotación y re-escalamiento de los ejes para generar consenso entre éstos y la proyección original. Como lo mostró Demey (2008), la transformación Procrustes es la que genera mayor estabilidad y por lo tanto, es la que se utiliza para la generación del consenso de las perturbaciones en este trabajo.

Las B configuraciones para la obtención de la variabilidad muestral, se generan a partir de WD. El algoritmo empleado, utiliza la matriz de residuos, tal como se detalla a continuación:

La auto-descomposición de WD es:

$$\hat{WD} = \tilde{U}_q \Delta_q \tilde{U}_q^T \tag{28}$$

así, tomando $q = 2$ o $q = 3$, se obtiene una configuración en dimensión reducida de dicha matriz. De este modo, WD, puede aproximarse como $WD = \hat{WD} + \varepsilon$, siendo ε una matriz de residuales con las mismas propiedades que WD y $\hat{WD}$ (la estimación de rango q $(q < r)$ de WD), es decir que ε es una matriz simétrica.

Remuestreando B veces en los $n(n-1)/2$ elementos fuera de la diagonal de la matriz ε, se generan B réplicas de la matriz WD: $WD_i^* = \hat{WD} + \varepsilon_i^*, i = 1..B$. Usando la auto-descomposición de las matrices WD_i^* se generan B nuevas matrices F^* $(F^* = \tilde{U}^* \tilde{\Delta}^{*1/2})$ que pueden ser comparadas con la configuración original (F) y crean la variabilidad muestral. La Figura 10,

ilustra esta metodología.

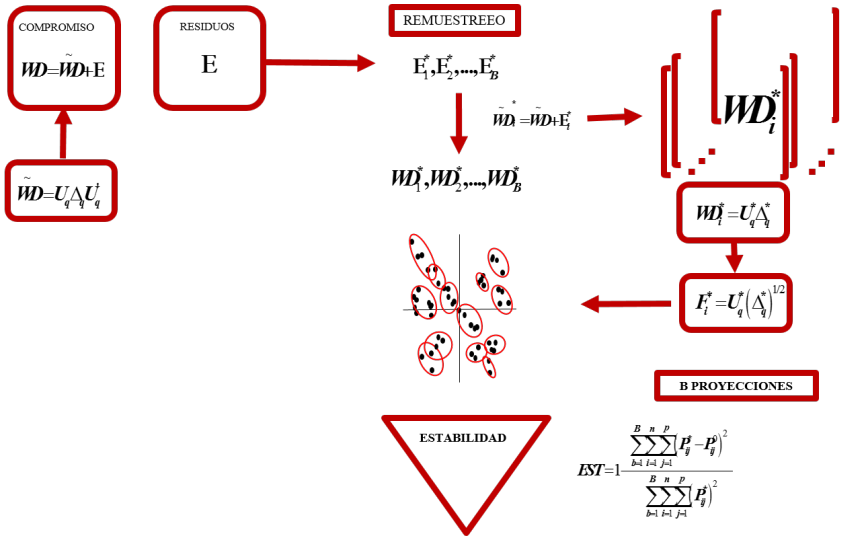


Figura 10: Esquema de remuestreo sobre la matriz compromiso.

Las réplicas bootstrap de la matriz original podrían no cumplir las condiciones necesarias para ser una matriz de productos escalares, es decir: que los elementos sobre la diagonal sean positivos y que la matriz sea semi-definida positiva.

La primer condición está garantizada porque no se realizan perturbaciones sobre la diagonal. Respecto de la segunda condición, si bien no está garantizada; en la práctica no debería representar un problema dado que se utilizan sólo los primeros autovalores de la matriz (Demey *et al.*, 2008).

La estabilidad del método empleado puede ser evaluada a través de la siguiente expresión:

$$EST = 1 - \frac{\sum_{b=1}^{B}\sum_{i=1}^{n}\sum_{j=1}^{p}(F_{ij}^{*} - F_{ij})^2}{\sum_{b=1}^{B}\sum_{i=1}^{n}\sum_{j=1}^{p}(F_{ij}^{*})^2} \tag{29}$$

donde F^*_{ij} es la representación para cada punto en las configuraciones de los diferentes remuestreos, en tanto que F es la configuración inicial. EST puede ser interpretada como la proporción de variabilidad asociada a la capacidad del método de análisis para reproducir la configuración original. Nótese que si cada F^* se aproxima a F, las sumas del numerador tienden a cero y por lo tanto, la estabilidad tiende a 1. En otras palabras, la estabilidad es máxima si las perturbaciones ejercidas sobre los datos o sobre los residuales reproducen la configuración original. En tanto que, la estabilidad tiende a cero (valor teórico más pequeño) siempre que el cociente participante de la ecuación 29 tienda a 1; esto ocurre cuando el numerador aumenta debido a que la perturbación produce una variabilidad del mismo orden que la de los datos mismos.

Tal como está descrito, el método puede proporcionar variabilidad espúrea debida al procedimiento: es decir, a las rotaciones ortogonales aleatorias y no a la perturbación de los datos. Esto se debe al hecho de que los vectores propios son únicos salvo el signo y en algunas réplicas podrían invertirse produciendo una variabilidad no real (Efron, 1987). Más aun, es posible que en las réplicas se intercambien los ejes factoriales o incluso se rote la solución, proporcionando, de nuevo, una variabilidad debida a las propiedades matemáticas específicas de las matrices y no a la perturbación estadística.

Si bien aquí se ejemplificó sobre el método STATIS, es posible realizar el mismo algoritmo que el indicado en la Figura 10 para el STATIS Dual (usando la matriz CM). Además de la evaluación de estabilidad de cada proyección, la construcción de elipses de confianza (o elipsoides si se usan 3 dimensiones en el análisis), permite evaluar cuál/cuáles objetos o variables no se diferencian entre sí teniendo en cuenta su comportamiento general a lo largo de todas las tablas involucradas en el análisis.

1.7 Representación Biplot para medir interacción entre individuos y variables

Los métodos presentados sólo posibilitan proyectar en un gráfico individuos o variables y no individuos y variables (o en términos de la problemática sobre análisis funcional del genoma en tejidos tumorales: genes o tejidos tumorales y no genes y tejidos). En muchas ocasiones, los investigadores no sólo desean saber las relaciones que se establecen entre observaciones, sino también entre éstas y variables involucradas o entre las propias variables. Por ejemplo, en el caso de estudios genéticos donde las k tablas son distintas plataformas de microarreglos, los individuos son tejidos tumorales y las variables genes; además del interés en la proyección de los tejidos tumorales y sus relaciones, incluyendo el estudio de su variabilidad interna, también resulta importante determinar cuáles son los genes asociados a dichos tejidos o incluso genes responsables de las similitudes o de las diferencias entre los mismos. Por lo tanto, se realiza una descripción y desarrollo teórico de los métodos Biplot y cómo pueden aplicarse a este tipo de estudios.

1.7.1 Formulación

La definición clásica del método Biplot es que posibilita la aproximación gráfica de una matriz de datos multivariantes -matriz de datos X de orden $(I \times J)$ y de rango r- usando marcadores filas y columnas para estudiar las relaciones entre individuos y variables, a partir de su descomposición en valores singulares:

$$X = UDV^T \tag{30}$$

De acuerdo a Gabriel (1971), esta matriz puede aproximarse por una de rango q (con $q < r$) tal que 30 se vea como:

$$X \approx= X_q = U_q D_q V_q^T \tag{31}$$

de aquí, puede escribirse como:

$$X \approx= X_q = U_q D_q^s D_q^{1-s} V_q^T \tag{32}$$

U y V los vectores singulares por izquierda (XX^T) y por derecha (X^TX), respectivamente y $s \in R$ y $0 \leq s \leq 1$. Las coordenadas filas y columnas son: $A = U_q D_q^s$ y $B^T = D_q^{1-s} V_q^T$. Estas estimaciones dan lugar, de acuerdo a los diferentes valores elegidos para s, a distintos métodos Biplot. Los más conocidos son el *Row Metric Preserving* (**RMP**) que se obtiene con $s = 1$, el *Column Metric Preserving* (**CMP**), cuando $s = 0$.

En el caso particular de los métodos STATIS y STATIS-Dual, las matrices compromiso que derivan son, como lo indican las fórmulas 10 y 21, cuadradas y sólo representan una configuración: o de individuos o de variables, por lo que, desde estas matrices, es imposible desarrollar un Biplot clásico (ver 32). Sin embargo, la matriz X también puede aproximarse a través de un modelo bilineal general del tipo multiplicativo:

$$X \approx= X_q = AB^T + \varepsilon \tag{33}$$

La ecuación 33 puede entenderse como una regresión multivariante de X sobre las coordenadas de los individuos A, cuando éstas están fijadas, o como una regresión multivariante de X^T sobre las coordenadas de las variables B, si son éstas quienes están fijas.

Es decir, si las filas están fijas, la estimación de B es:

$$B^T = (A^T A)^{-1} A^T X \tag{34}$$

del mismo modo, si las columnas están fijas, A se obtiene por:

$$A^T = (B^T B)^{-1} B^T X^T \tag{35}$$

Estas ecuaciones se aplican alternadamente, se normaliza A (por ejemplo cada dos iteraciones) y el algoritmo converge a la descomposición en valores singulares de la matriz X (Vicente-Villardón *et al.*, 2006). Además, si se ortogonaliza la matriz, se puede asegurar una solución única.

El modelo de la ecuación 34 permite generar la aproximación Biplot a partir de un modelo lineal clásico, es decir $E(X) = AB^T$ si la distribución de las variables es normal o bien como un modelo generalizado ($E(X) = g(\mu) = AB^T$) si la distribución de las variables pertenece a la familia exponencial y sus valores esperados se encuentran relacionados con predictores lineales (η) a través de alguna función de enlace (por ejemplo podría ser *logit*, *probit*, *complemento log-log*, entre otras) (Demey *et al.*, 2008; Cárdenas *et al.*, 2003). Es evidente que los modelos clásicos son un caso particular de los modelos generalizados cuya función de enlace es la identidad:$(g(\mu) = \mu)$.

Las ideas de la ecuación 33 pueden utilizarse para aproximar cualquier matriz de datos X. Así, si X representa una de las k-matrices originales en el método STATIS (de orden $n \times j_k$) y A es la matriz de proyección de los individuos, derivada por ejemplo de la auto-descomposición de la matriz compromiso; es decir, es la matriz $F = \tilde{U}\tilde{\Delta}^{1/2}$ que representa las coordenadas medias de los individuos en dicho espacio, es posible plantear un modelo lineal que consista en realizar regresiones simples de cada columna de X (x_{jk}) sobre F (de orden $n \times 2$ o $n \times 3$) tal como lo indica la ecuación 36.

$$x_{jk} = FB^T + \varepsilon \tag{36}$$

De la ecuación 36, se puede estimar $\hat{B}$ como:

$$\hat{B} = (F^T F)^{-1} F^T x_{jk} \tag{37}$$

que representa la coordenada de la j_k variable dentro de la estructura compromiso donde previamente fueron proyectadas las coordenadas de los individuos. Vicente-Villardón *et al.* (2006) describen la geometría del Biplot ajustado a través de modelos de regresión lineal. Si el ajuste Biplot es bidimensional, la matriz F de la ecuación 37 tiene dos columnas y geométricamente se representa en un plano. Llamando L a este espacio, el objetivo es encontrar los marcadores columnas (b_{jk}) cuya proyección prediga los valores de la variable j_k de la mejor manera posible. En este trabajo, los autores añaden una tercera dimensión a la j_k-ésima variable y ajustan el plano de regresión habitual (H); demuestran que la predicción de un valor fijo está dada por la intersección entre el plano normal al tercer eje para el valor particular de x_{jk} y el plano de regresión, obteniéndose rectas paralelas en el

plano H. Al eje de referencia, que permite predecir los valores y representa la dirección de H normal a todas esas rectas paralelas en el plano de regresión se le denomina ξ_{jk}, en tanto que los puntos en L que predicen diversos valores de la variable también están en líneas rectas paralelas; la proyección de ξ_{jk} sobre L es normal a todas las líneas y se denomina eje Biplot b_j. De este modo, para encontrar un marcador en los ejes Biplot que prediga un valor fijo (μ) de la variable observada, se debe resolver el siguiente sistema de ecuaciones:

$$y = \frac{b_{j2k}}{b_{j1k}}x \quad y \quad \mu = b_{j0k} + b_{j1k}x + b_{j2k}y \tag{38}$$

cuyas soluciones son:

$$x = \mu\frac{b_{j1k}}{b_{j1k}^2 + b_{j2k}^2} \quad y = \mu\frac{b_{j2k}}{b_{j1k}^2 + b_{j2k}^2} \tag{39}$$

o de manera general:

$$(x, y) = \mu\frac{b_{jk}}{b^T j_k b j_k} \tag{40}$$

es decir que el marcador que predice un valor fijo de la j_k-ésima variable, viene dado por la razón entre las coordenadas y su longitud ajustada. Además, los coeficientes de determinación de cada variable, permiten medir su calidad de representación y se interpreta como en los Biplots clásicos.

Esta aproximación Biplot sobre el compromiso en el método STATIS, permite proyectar las variables de las distintas tablas de datos y determinar relaciones entre observaciones y variables y, variables entre sí. Adicionalmente, el método provee una vía para seleccionar variables usando las medidas de

bondad de ajuste de cada uno de los modelos lineales estimados.

Esto muestra que, en cualquier método de reducción de la dimensión, es posible representar variables e individuos en un mismo plano utilizando las proyecciones generadas y haciendo regresiones entre las variables originales y los ejes retenidos.

1.8 Redes y Grafos

Los métodos Biplot permiten la selección de genes candidatos y si bien las ontologías son muy importantes para determinar su función específica y los términos biológicos enriquecidos, en la gran mayoría de los casos proveen largas listas de genes que son difíciles de analizar e interpretar. Razón por la cuál, en el marco de este trabajo, los resultados se analizan usando la herramienta **FGNet** que permite obtener asociación funcional entre genes e identificar genes multifuncionales (Aibar *et al.*, 2015).

Por lo tanto, se definirá brevemente el modo de construcción de los grafos o redes utilizados (*"network"*). Básicamente, una red es un conjunto de vértices y aristas que se puede definir matemáticamente como $G(V, E)$ y que representa una relación binaria de E (el conjunto de aristas) sobre V (el conjunto de vértices). El número de vértices indica el orden del grafo y generalmente se denota n, en tanto que el número de aristas es m (Grimaldi, 1998).

Para construir cualquier grafo, son necesarias dos matrices que se describen en los párrafos siguientes.

En primer lugar, se requiere una matriz de incidencia M de orden $n \times s$; que en este caso hacen referencia al número de genes enriquecidos en la lista seleccionada (n) y el número de grupos de genes (s), ambos determinados por el análisis de ontologías. Un elemento m_{ns} es igual a 1 si el n-ésimo gen está presente en el grupo s y es 0, en caso contrario.

La matriz M, puede transformarse en una matriz de adyacencia A (de orden $n \times n$); donde cada elemento a_{ij} está dado por la ecuación 41:

$$a_{ij} = \sum_{s} (m_{i,s} \times m_{j,s})(1 - \delta_{i,j}) \tag{41}$$

con $\delta_{i,j} = 1$ si $i = j$ y 0 en caso contrario. Es decir, la ecuación 41, indica la cantidad de grupos que comparten dos pares de genes. Esto evidencia la imposibilidad de que la matriz de adyacencia tenga algún elemento negativo. Más aun, esta matriz tiene traza igual a cero y por lo tanto, la suma de todos sus autovalores es nula.

Finalmente, M es utilizada para generar los pesos de las aristas que conectan a cada par de vértices (es decir, $a_{ij} \neq 0$). Con esta información, se realiza el grafo no dirigido que relaciona grupos de genes de acuerdo al análisis ontológico.

2. Ilustración de la metodología propuesta: integración de datos ómicos provenientes de líneas celulares de cáncer del panel NCI60

2.1 Conjunto de Datos NCI60

A modo de ilustración de la metodología de análisis propuesta, se aborda un estudio del panel de datos de cáncer NCI-60, tomados desde el repositorio CellMiner (Shankavaram *et al.*, 2009; Reinhold *et al.*, 2012).

En el año 1985, la junta de asesores científicos del Instituto Nacional del Cáncer de los Estados Unidos decidió rediseñar su programa de detección

de drogas (Burkard, 2012). El nuevo programa fue desarrollado para evaluar la sensibilidad a sustancias químicas de 60 líneas celulares humanas de distintos tejidos tumorales. Los tejidos evaluados corresponden a (las abreviaturas hacen referencia al nombre en inglés): melanomas (ME), leucemias (LE), mama (BR), renal (RE), ovario (OV), sistema nervioso central (SNC) pulmón (LC), próstata (PR) y colon (CO) (Shankavaram *et al.*, 2009; Reinhold *et al.*, 2012). Estas líneas han sido estudiadas intensamente a partir del año 1992 (Liu *et al.*, 2013a) y procesadas con tecnologías de alto rendimiento en años recientes, donde fueron utilizadas múltiples plataformas para caracterizarlas entre las que se incluyen arreglos de genómica comparativa, análisis de mutación de ADN, huella de ADN, microarreglos o análisis de proteínas (Reinhold et al, 2012). Entre los variados objetivos que tienen los investigadores del área, se encuentran los de integrar la información proveniente de los estudios nombrados como también integrar información biológica con farmacológica. En este sentido, el sitio web CellMiner brinda algunas herramientas de análisis para investigadores sin experiencia en informática y, además, organiza y almacena tanto datos crudos (raw data) como normalizados de microarreglos de ADN, RNA, proteínas y drogas de las líneas celulares mencionadas (Shankavaram *et al.*, 2009).

En este trabajo, se utilizan los datos crudos de transcriptoma de las líneas celulares de NCI-60 disponibles en CellMiner provenientes de cuatro plataformas de análisis: *Affymetrix* HG-U133 plus 2.0 (47000 transcripciones aproximadamente), HG-U133 (44000 sondas, 2 conjuntos de chips (A-E)), HG-U95 (65000 sondas, 5 conjuntos de chips (A-E)) y *Agilent* GE 4x44K (Alrededor de 44000 sondas para 41000 genes, aproximadamente). En total, se eligen 58 líneas celulares debido a que en las tablas originales hay dos líneas ausentes: $RE.UO_{31}$ en la plataforma *Affymetrix* $HG-U133$ plus 2,0 y $LCNCI_{H23}$ $HG-U133$.

Los cálculos y representaciones gráficas se realizan con la librería **kimod** (A k-tables approach to integrate multiple Omics-Data)(Zingaretti *et al.*, 2015).

2.2 Procesamiento de Datos

Los datos se pre-procesan según lo establecido en la sección 1.2. La Figura 11 muestra el esquema de trabajo utilizado. Como se indicó en la sección 1.2, para los genes repetidos en cada tabla, se elije el de mayor expresión media. Para realizar las correcciones de *Backround*, se utilizan los paquetes **Limma** (Ritchie *et al.*, 2015) y **Affy** (Gautier *et al.*, 2004) de **R** (R Core Team, 2015), combinados con códigos desarrollados en el contexto de este trabajo.

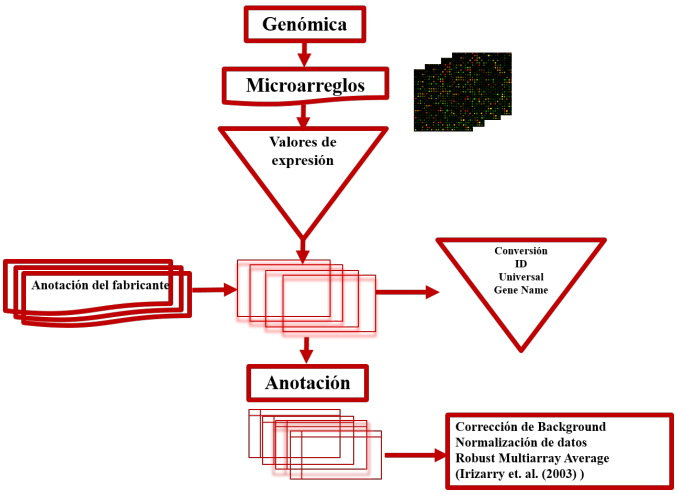


Figura 11: Flujo de trabajo pre-procesamiento de datos.

La Tabla 1 muestras las dimensiones de las tablas generadas por las 4 plataforma incluidas en este estudio. La dimension fila, corresponde a los genes y la dimension columna a los tejidos tumorales.

Cuadro 1: Dimensiones de las tablas de datos

Tabla	Dimensión
HG-U133	18133x58
HG-U133 plus 2.0	21469x58
HG-U95	17719x58
Agilent	16726x58

Los genes se ordenan en filas y los tejidos tumorales en columnas. En total, hay 14931 genes comunes a todos los estudios (ver Figura 12).

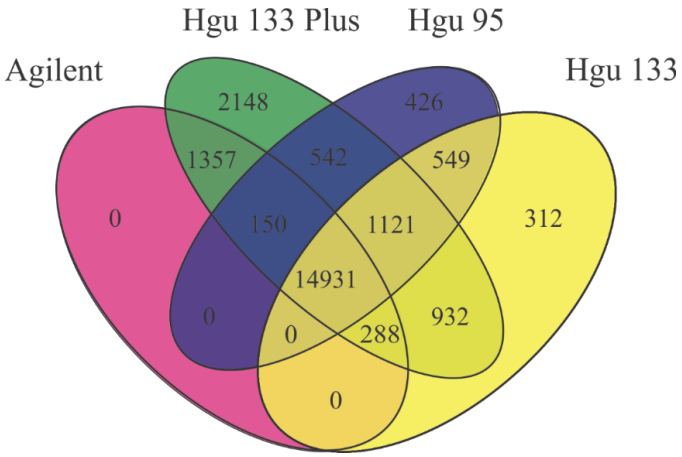


Figura 12: Diagrama de Venn que muestra las intersecciones de los genes entre las distintas tablas.

Además, se realiza otro proceso de normalización, que tiene en cuenta la contribución relativa de las filas, columnas e individuos a la variación total de cada tabla (Meng y Gholami, 2014). Llamando Y a cada matriz de expresión, con elementos: $Y = [y_{ij} 1 \leq i \leq n, 1 \leq j \leq J_k]$. Sean $m_{i.}$ y $m_{.j_k}$ las sumas de los elementos de cada fila y columna de Y, respectivamente y sea $m_{..}$ la suma de todos sus elementos, la contribución relativa de cada fila a la variación total, se denota $r_i = \frac{m_{i.}}{m_{..}}$,en tanto que la de cada columna: $c_{jk} = \frac{m_{.j_k}}{m_{..}}$. Por otra parte, la contribución de cada individuo, se computa como: $p_{ij_k} = \frac{m_{ij_k}}{m_{..}}$. De aquí, se derivan las matrices estandarizadas (X) para realizar los análisis, cuyos elementos se definen:

$$x_{ij_k} = \frac{p_{ij_k}}{r_i} - c_{j_k} \tag{42}$$

una vez realizados estos pre-procesamientos, se realiza un análisis STATIS sobre el conjunto de tablas.

Finalmente, se utiliza el Biplot lineal para seleccionar genes mejor representados y potencialmente responsables de las diferencias entre los tipos de cáncer y se realizan redes de co-expresión y un análisis ontológico-funcional.

2.2.1 Análisis de ontologías: integración de datos en una red

Luego de seleccionar los genes comunes y mejor representados en todas las tablas, se realiza un análisis ontológico, se agrupan los genes en meta-grupos y se construye un grafo de relaciones entre ellos, tal como se explicó en la sección **Redes y Grafos**, usando el algoritmo *Organic Layout*, que pertenece al grupo de algoritmos "dirigido por fuerzas"(yWorks GMBH, 2013). Este grupo de algoritmos modelan al grafo como un sistema físico y tienen el objetivo de lograr un equilibrio en el sistema de fuerzas que interactúan entre los vértices de modo tal que alcancen una posición final en que la suma de las fuerzas totales del sistema sea nula. El algoritmo que se utiliza en este trabajo, tiene la particularidad que permite la identificación de grupos de vértices fuertemente conectados (Tuikkala *et al.*, 2012). Finalmente, se utiliza el flujo de trabajo de la Figura 8 para obtener información provista por la literatura relacionada con el cáncer y se realiza un análisis ontológico del listado de genes . Para realizar los grafos y búsqueda de enfermedades fueron utilizados **Cytoscape** (Shannon *et al.*, 2003), su plug-in **Reactome** (Croft *et al.*, 2014) y el paquete **FGNet** (Aibar *et al.*, 2015).

2.3 Resultados

2.3.1 Interestructura

En la Figura 13, se observan las imágenes euclídeas de los estudios: las dos primeras dimensiones explican más del 90 % de la variabilidad total.

A continuación, se muestra el código para obtener los resultados:

```
library("kimod")
library("xtable")
options(width = 98)
setwd("C:/Users/Laura/Desktop/TesisSweavePrueba/Analisis/Datos")
load("Datos.RData")
Z1<-DiStatis(Datos,Center = FALSE, Scale = FALSE,
             CorrelVector = TRUE, Frec = TRUE, Traj = FALSE)

Tissues<-c(rep("Mama",5),rep("SNC",6),rep("Colon",7),
           rep("Leucemia",6),rep("Melanoma",10),
           rep("Pulmón",8),rep("Ovario",7),
           rep("Próstata",2),rep("Renal",7))
Colors<-c(rep("limegreen",5),
          rep("palevioletred4",6),rep("navyblue",7),
          rep("yellow",6),rep("sienna1",10),
          rep("red",8),rep("steelblue",7),
          rep("maroon3",2),rep("midnightblue",7))
```

2.3.2 Configuración Compromiso

La Figura 14 muestra las proyecciones de las observaciones comunes (líneas) en el compromiso, en tanto que la Figura 15, destaca cada uno de los tejidos. Los tejidos de cáncer de mama, pulmón y ovario son los que presentan mayor variabilidad. En tanto que el cáncer de Sistema Nervioso Central (SNC) es el menos variable. Algunos tejidos de cáncer de mama se agrupan con los tejidos de cáncer de colon, en tanto que otros se asemejan a los tejidos de cáncer de Sistema Nervioso Central. Algunos perfiles de cáncer de pulmón se asemejan a los de cáncer de ovario y otros a los de cáncer renal. Además, los tejidos que más se diferencian son los de SNC, colon, leucemia y melanoma.

Aunque no se muestran las figuras, se observaron las proyecciones entre todas las dimensiones desde la primera hasta la octava, dado que al haber nueve clases, podrían ser necesarias hasta 8 dimensiones para la separación. En este sentido, es preciso aclarar que en la dimensión 5, se diferencian los tejidos de cáncer renal y de pulmón. En la dimensión 6, el cáncer renal se separa de los tejidos de cáncer de ovario y pulmón y en la dimensión 8, se separan los tejidos de ovario y pulmón entre sí. En tanto que los tejidos de cáncer de mama, son altamente variables en todas las dimensiones estudiadas.

La Figura 16 muestra el análisis de estabilidad de los resultados. Allí, se observan las elipses de confianza realizadas sobre las proyecciones en el compromiso. Los tipos de cáncer que más variabilidad interna presentaron fueron leucemia y melanoma. En tanto que, los tejidos de cáncer de colon presentaron una menor variabilidad interna y están bien separados (a excepción de uno) del resto de los tejidos. El tejido de cáncer de mama muestra la misma tendencia que en la Figura 14. Uno de ellos, muestra un perfil similar a uno de los tejidos de cáncer de colon y a los de leucemia (aunque es mucho menos

variable), en tanto que otros presentan patrones similares a los tejidos de Sistema Nervioso Central y, en menor medida, a los melanomas. Además, es importante notar que el porcentaje de inercia explicado por cada eje fue calculado usando *bootstrap*.

La Figura 17 muestra las diferencias *bootstrap* en el primer eje. Los tejidos de cáncer de mama se separan, quedando dos de ellos ("BR.MCF7", "BR.T47D") agrupados junto a los de colon. Además, en el primer eje, éstos se agrupan con la mayoría de los tejidos de melanoma y ovario. Por otra parte, los tejidos de cáncer renal, pulmón, sistema nervioso central, dos de mama y la mayoría de leucemia, tienen marcadores positivos en esta dimensión y conforman otro grupo. Por otro lado, de acuerdo a la Figura 18, la segunda dimensión separa los tejidos de melanoma y pulmón, del resto de los tejidos. Adicionalmente, es importante notar que la variabilidad es mayor en la segunda dimensión que en la primera (lo cuál puede observarse a través de las formas de las elipses en la Figura 16). Las razones *bootstrap* son computadas por medio del cociente entre la media y el desvío estándar bootstrap, respectivamente.

2.3.3 Proyección de genes usando Biplot

En la Figura 19, se muestran las proyecciones de todos los genes sobre el espacio compromiso.

2.3.4 Selección de genes comunes a todas las tablas usando Biplot

Se eligieron 0.5 % de genes con mayor R^2 ajustado (y por lo tanto, mejor representados) usando modelos lineales tal como se explicó en la sección 1.7. En total, 156 resultaron comunes a todas las matrices (ver Cuadro 2).

ADAM9	CLCF1	FLNB	LEPREL1	PKIG	SH2D4A
AMIGO2	CLDN3	FSTL3	LGALS1	PKP3	SLC24A5
AMOTL2	COL6A1	FZD2	LHFP	PLAU	SOX10
ANKRD44	CRIM1	GJA1	LOXL2	PLCXD2	SPARC
ANXA2P1	CTGF	GLIPR1	LRP12	PLK2	SPINT2
AP1S2	CYR61	GPNMB	MAP1B	PLS1	SRGAP2
AREG	DAAM2	GPR176	MAP3K14	PTPN14	SRGAP2P1
ARNTL2	DDR2	GRB7	MARVELD2	PTPN6	ST3GAL6
ASAM	DFNA5	GYPC	MLANA	PTRF	ST6GAL1
AXL	DKK3	HEG1	MST1R	PVR	ST8SIA1
BDNF	DST	HRH1	MXRA7	PYGO1	SYDE1
C17ORF91	DZIP1L	IL16	MYB	RAI14	SYN3
C19ORF21	EGFR	IL6ST	MYLK	RASSF7	TGFB2
C1ORF172	EPB41L4B	ITGA3	MYO5C	RBM9	TJP1
C6ORF218	EPCAM	ITGB1	MYOF	RBMS3	TMEM45A
C9ORF167	EPHA2	JUB	NAV3	RECK	TNFRSF12A
CALD1	EPS8L2	JUP	NRG1	RIN2	TNS4
CALU	FAM126A	KAT2B	NTN4	RNF11	TPM1
CAPN3	FAM127A	KIAA0802	NUAK1	RRAS	TPST1
CARD10	FAM20C	KIRREL	OSMR	RUSC2	TRPV2
CAV1	FAM78A	KRT19	OSTM1	S100A2	TWSG1
CAV3	FCGR2A	KRT8	P2RX7	S100B	TYR
CCDC88C	FCRLA	KRT80	PDGFC	SCHIP1	VCL
CD151	FERMT2	LAMC1	PDLIM7	SERPINE1	VEGFC
CFL2	FGD3	LAPTM4A	PEA15	SGCD	VIM
CHMP4C	FLNA	LARP6	PHLDB2	SH2D3A	WIPF1

Cuadro 2: Lista de genes seleccionados comunes a todas las tablas.

En la Figura 20 se muestran los vectores que indican la dirección de sus proyecciones medias sobre el compromiso. Para facilitar la visualización de los resultados, los genes seleccionados se agruparon en 3 conjuntos con el

criterio de agrupamiento de Ward.

Adicionalmente, el paquete **kimod** implementa una función que permite establecer relaciones entre los genes seleccionados y las muestras (tejidos tumorales en este caso), usando la proyección ortogonal de cada uno de los tejidos sobre las direcciones de proyección de los genes. Los genes marcados en negro, se caracterizan por estar sobre-expresados en los tejidos de Sistema Nervioso Central (y los dos de mama que están junto a ellos) y renal y por tener una baja expresión en los tejidos de cáncer de colon, leucemia y melanoma.

Los genes marcados en color rojo, se caracterizan por tener una alta expresión en los tejidos de melanoma, leucemia y de Sistema Nervioso Central.

Finalmente, los genes marcados en verde, se caracterizan por tener alta expresión en el cáncer de colon, ovario, pulmón y sub- expresión en Sistema Nervioso Central y melanoma. En el Cuadro 3, se puede encontrar un listado de los mismos. Cabe aclarar que los genes comunes son un subconjunto de los genes de esta tabla, dado que aquí se muestran los seleccionados comunes y no comunes. Esto es importante, porque por ejemplo, en el grupo 1, se encuentra el gen BCAR3, que en la literatura está asociado a la presencia de cáncer de mama y que no es común a todas las tablas.

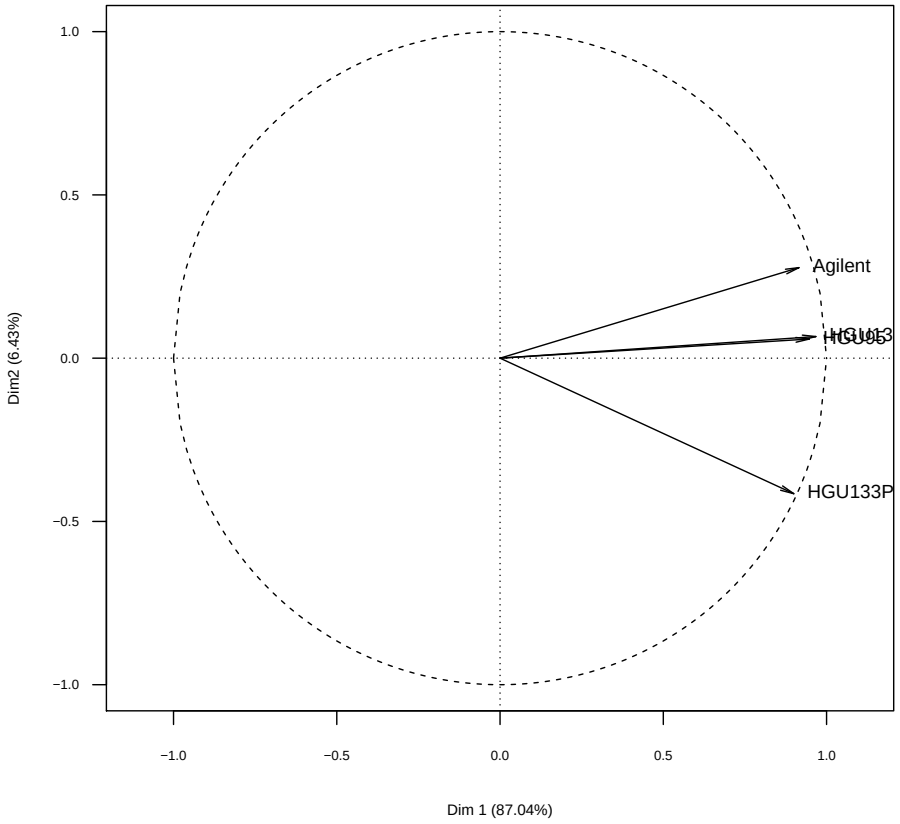


Figura 13: Proyección de la configuración euclídea de las tablas. Las plataformas *Affymetrix* HGU133 y HGU95 tienen la misma estructura y son similares a la plataforma *Agilent*.

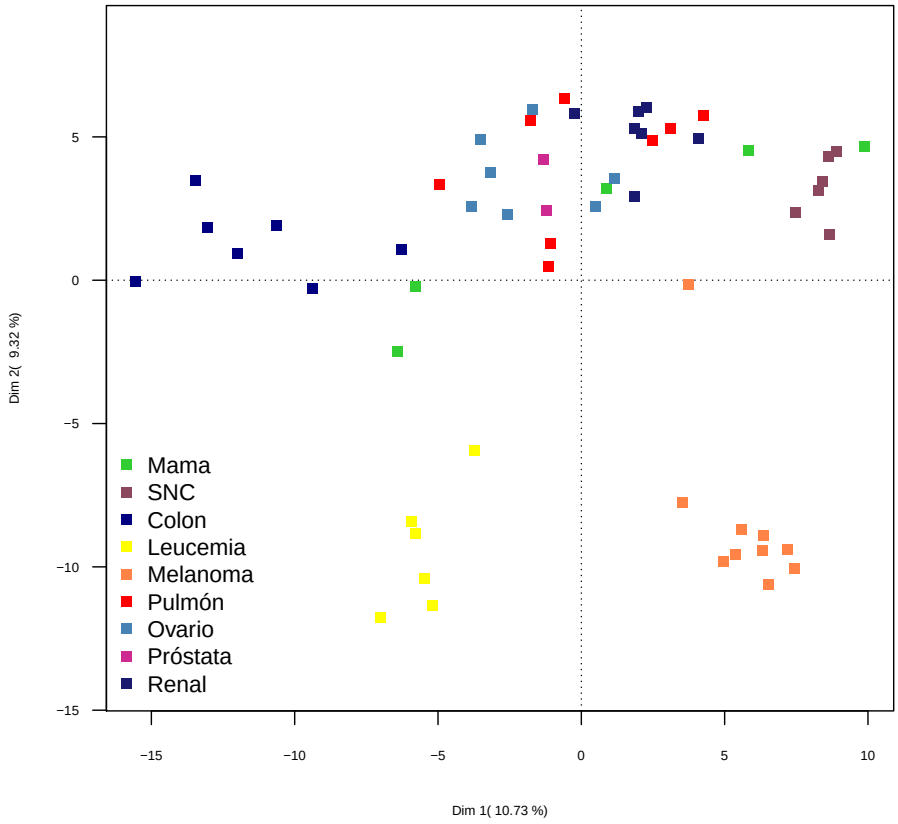


Figura 14: Proyección de las observaciones medias (líneas celulares) en el compromiso.

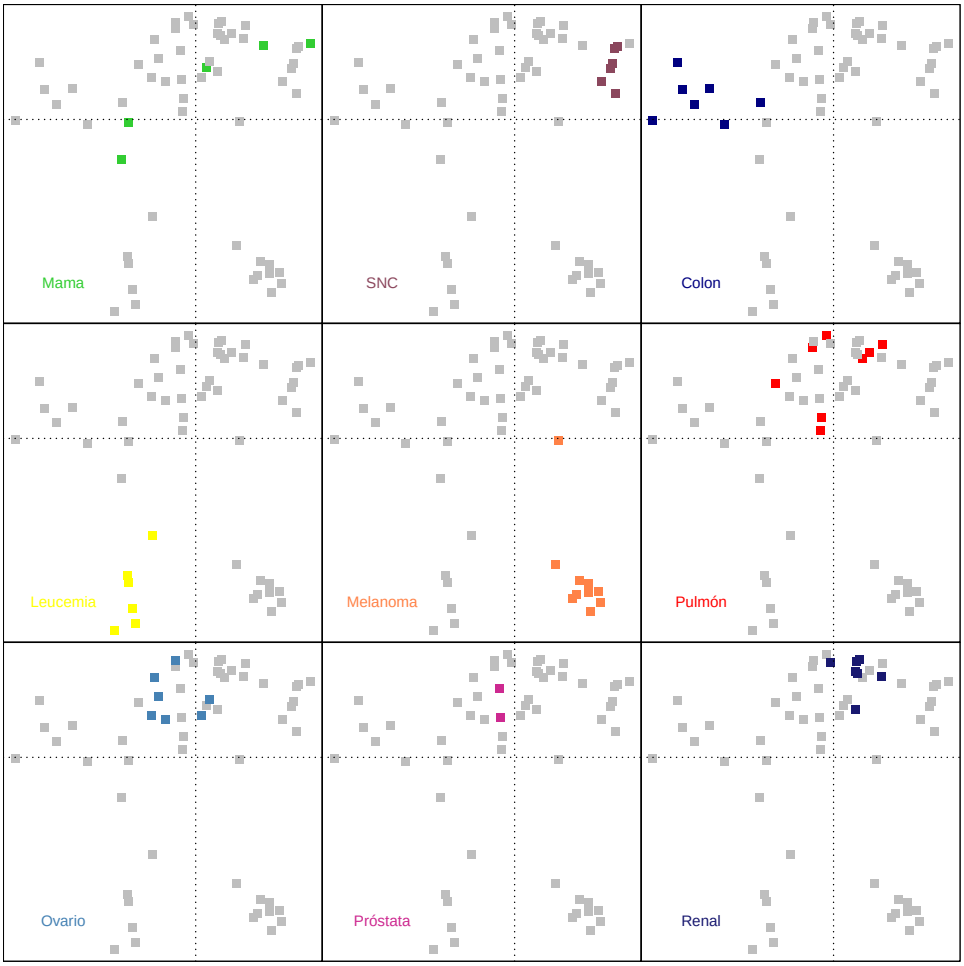


Figura 15: Proyección de las observaciones medias (líneas celulares) en el compromiso. En cada uno de los gráficos, se destaca la proyección de un tejido.

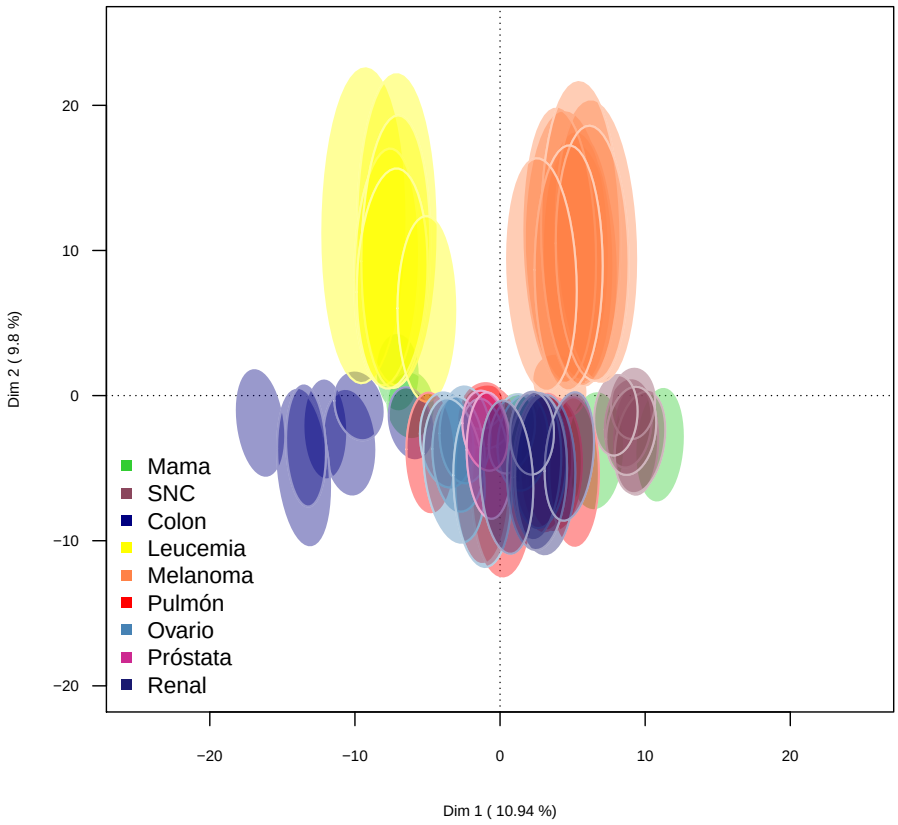


Figura 16: Elipses de confianza para las líneas celulares.

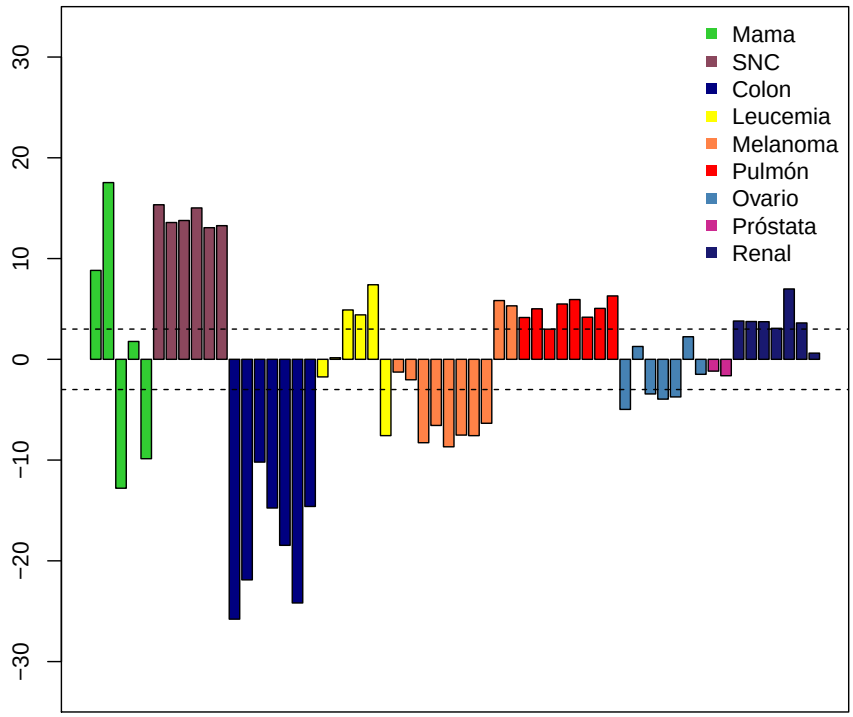


Figura 17: Razones *bootstrap* para la primer dimensión del compromiso.

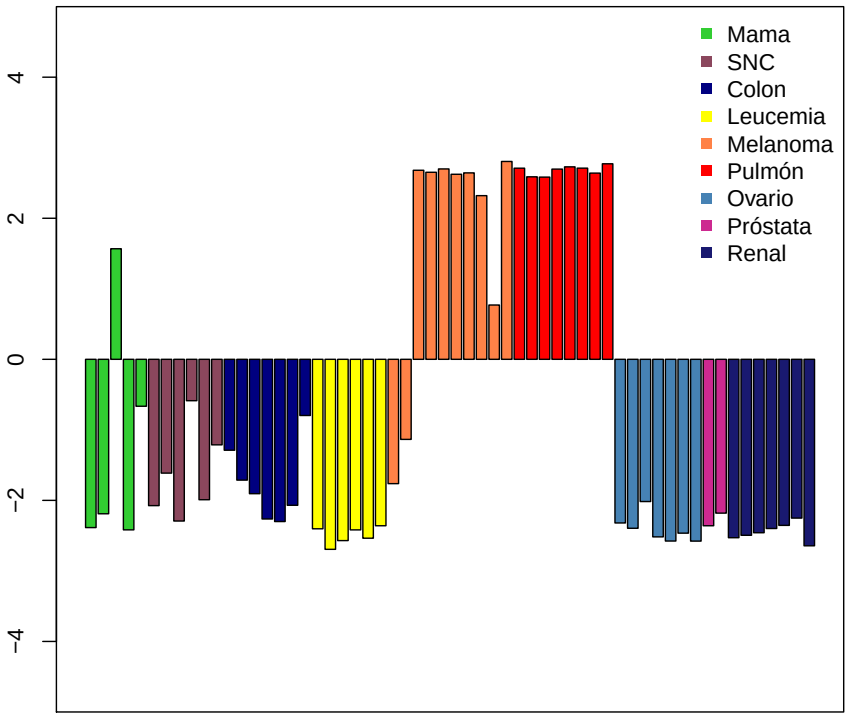


Figura 18: Razones *bootstrap* para la segunda dimensión del compromiso.

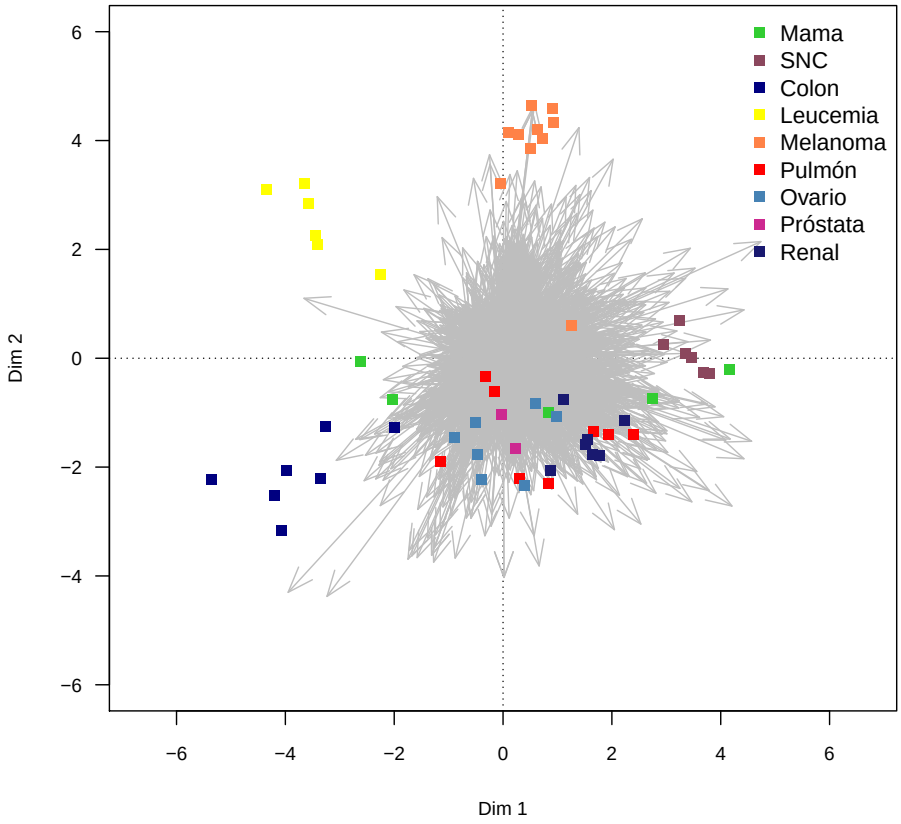


Figura 19: Proyección de los genes de todas las tablas sobre el compromiso usando Biplot.

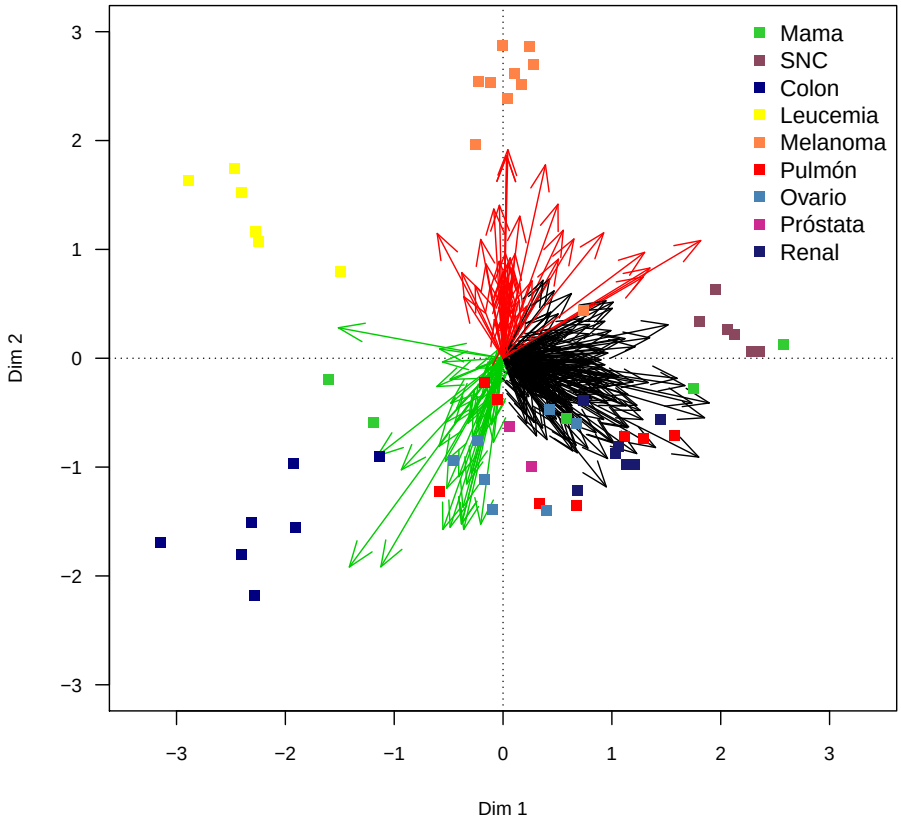


Figura 20: Proyección de los genes de todas las tablas sobre el compromiso usando Biplot.

Grupo	Genes						Sobre-Exp	Sub-Exp
Negro	ABL2	CFL2	FLNA	LGALS1	PDGFC	RUSC2	Sistema Nervioso Central, renal	colon, leucemia, melanoma
	ADAM9	CHST3	FLNB	LHFP	PDLIM7	S100A2		
	AFAP1L1	CKAP4	FOSL2	LIFR	PEA15	SCHIP1		
	AMIGO2	CLCF1	FSTL3	LOC652725	PHLDB2	SERPINE1		
	AMOTL2	COL12A1	FZD2	LOC653653	PKIG	SH2D4A		
	ANLN	COL6A1	GJA1	LOX	PLAU	SHISA4		
	ANXA2P1	CRIM1	GLIPR1	LOXL2	PLCXD2	SHOX2		
	AP1S2	CTGF	GPR176	LRP12	PLK2	SIPA1L3		
	ARNTL2	CYR61	GPR39	MAP1B	PPP1R13L	SPG20		
	ASAM	DCLK2	GPX8	MAP3K14	PPP2R3A	SRGAP2		
	AXL	DDR2	HEG1	MGAT5B	PTPN14	SRGAP2P1		
	BCAR3	DFNA5	HRH1	MSRB3	PTRF	SYDE1		
	BDNF	DKK3	IL6ST	MXRA7	PVR	TGFB2		
	BNC2	DST	ITGA3	MYLK	PYGO1	THBS1		
	C17ORF91	DZIP1L	ITGB1	MYOF	RAI14	TIMP2		
	CALD1	EGFR	JUB	NAV3	RBM9	TJP1		
	CALU	EPHA2	KIAA0802	NNMT	RBMS3	TMEM45A		
	CAMSAP1L1	EVC	KIRREL	NRG1	RECK	TNFRSF12A		
	CARD10	FAM126A	KRT80	NTN4	RHOQ	TPM1		
	CAV1	FAM127A	LAMC1	NUAK1	RHOQP2	TPST1		
	CAV3	FAM127B	LAPTM4A	OSMR	RIN2	TWSG1		
	CCDC80	FAM20C	LARP6	OSTM1	RNF11	VCL		
	CD151	FERMT2	LEPREL1	PAWR	RRAS	VEGFC		
Rojo	ANKRD44	GAS7	P2RX7	ST3GAL6			melanoma, leucemia,Sistema Nervioso Central	colon, pulmón, ovario, renal, próstata.
	AP1S2	GPNMB	RENBP	ST6GAL1				
	C6ORF218	GYPC	ROPN1	ST8SIA1				
	CAPN3	HMCN1	S100B	SYN3				
	CNRIP1	IGSF11	SGCD	TRIM63				
	DAAM2	IL16	SLC1A4	TRPV2				
	DOCK10	IL1RAP	SLC24A5	TYR				
	EDNRB	KAT2B	SNAI2	TYRL				
	FAM78A	LCP2	SOX10	VIM				
	FCGR2A	LRRC33	SOX6	WIPF1				
	FCRLA	MLANA	SPARC	ZEB2				
Verde	AREG	EPB41L4B	KRT8P12	PLS1			colon, ovario, pulmón	Sistema Nervioso Central, melanoma
	AREGB	EPCAM	KRT8P9	PSD4				
	ARHGAP27	EPS8L2	LLGL2	PTPN6				
	C19ORF21	FGD3	LOC146880	RASSF7				
	C19ORF33	GRB7	LOC149501	SH2D3A				
	C1ORF172	JUP	MARVELD2	SLC27A2				
	C9ORF167	KIAA1543	MST1R	SPINT2				
	CCDC88C	KLF5	MYB	TNS4				
	CHMP4C	KRT15	MYO5C					
	CLDN3	KRT19	NBEAL2					
	ELF3	KRT8	PKP3					

Cuadro 3: Relaciones entre genes seleccionados y tejidos tumorales.

Usando los genes comunes a todas las tablas, se resumió la información en una única matriz compromiso, derivada de aplicar un análisis X-STATIS sobre las matrices de dimensión 156×58. A partir de dicha matriz, se realizó un mapa de calor ("heatmap") que permite observar si la estructura de los datos se mantiene con la cantidad reducida de genes seleccionados, lo que constituye una forma de determinar si la selección ha sido adecuada. Para medir la distancia entre individuos se computa la distancia euclídea, en tanto que para hacer el análisis de conglomerado se usa el método de agrupamiento de Ward.En la Figura 21 se observan los resultados. Se distingue 3 grupos de genes (en columnas) y 6 grupos de tejidos tumorales. La estructura de los datos se mantiene.

En la Figura 22, se puede observar la proyección de los tejidos tumorales sobre el compromiso usando sólo los genes seleccionados. Aquí también se puede apreciar que la estructura de los datos se mantiene, es decir la posición relativa de la mayoría de los tejidos es la misma que cuando se usa toda la información, siendo los tejidos de melanoma, leucemia, colon, sistema nervioso central los más fáciles de diferenciar. Asimismo, los tejidos de cáncer de mama siguen mostrando una alta variabilidad interna, con perfiles muy diferenciados entre sí. Esto muestra que el método es muy eficaz para la selección de información.

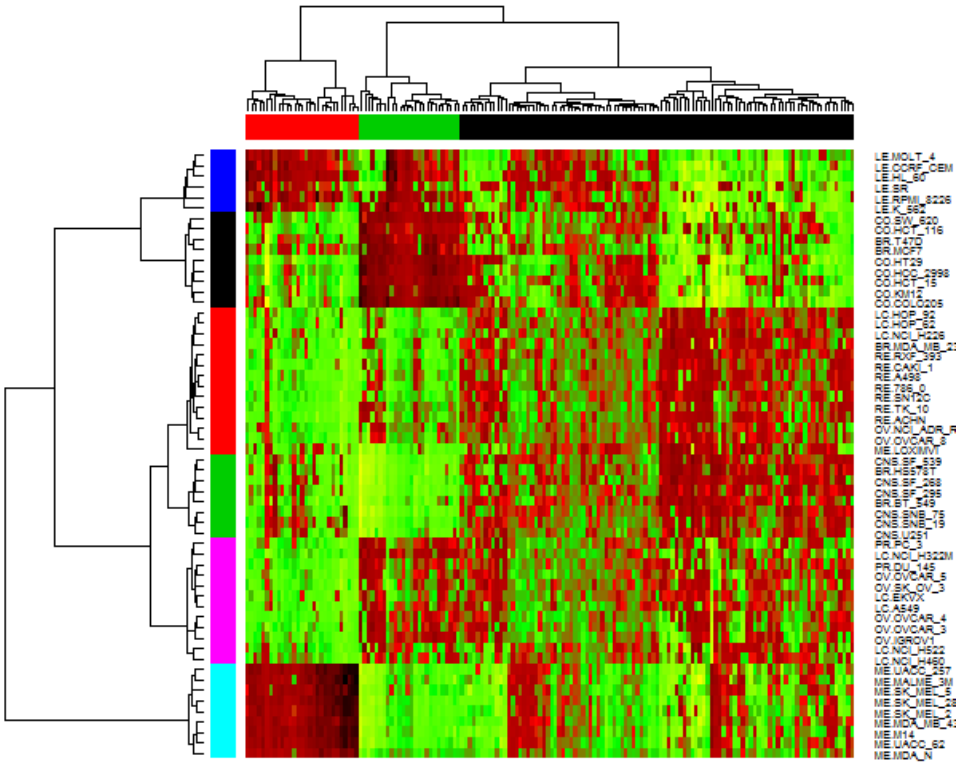


Figura 21: Mapa de calor de los genes seleccionados y tejidos tumorales.

2.4 Análisis de ontologías: red de relaciones de genes seleccionados y genes activados ante la presencia de distintos tejidos

Una vía para validar los resultados obtenidos por el análisis es contrastarlos con la información existente en la literatura. La búsqueda de estos genes usando el esquema de la Figura 8 y un procedimiento manual, indica que 98 de ellos tienen relaciones probadas con cáncer, es decir aproximadamente el

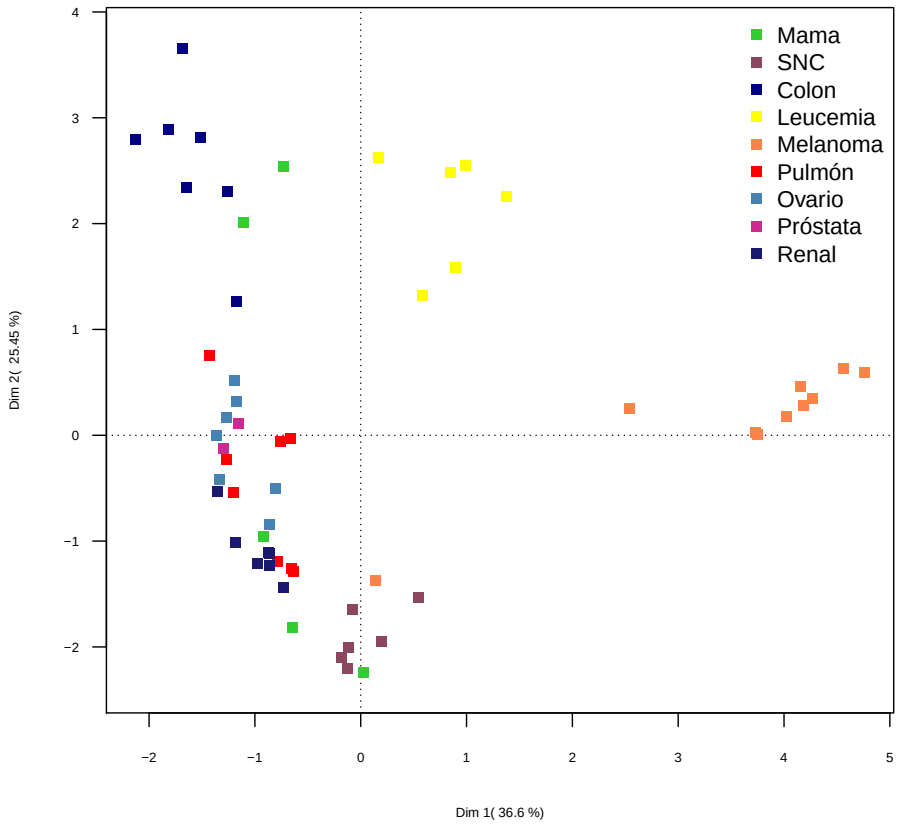


Figura 22: Proyección de los tejidos tumorales en el compromiso, usando sólo los genes seleccionados.

63 %. Este número se contrasta seleccionando 10000 listas al azar desde el conjunto de genes comunes. La proporción de términos anotados al azar en el *Cancer Gene Index* es de 0.29, por lo que se puede concluir que existen evidencias suficientes para suponer que la lista seleccionada no es producto del azar.

Por otra parte, usando la herramienta **FGNet** (Aibar *et al.*, 2015), se construyó un grafo dirigido por fuerzas de los genes seleccionados y los

meta-grupos obtenidos que los asocian a determinados procesos o funciones biológicas, tal como se indicó en la sección **Redes y Grafos**.

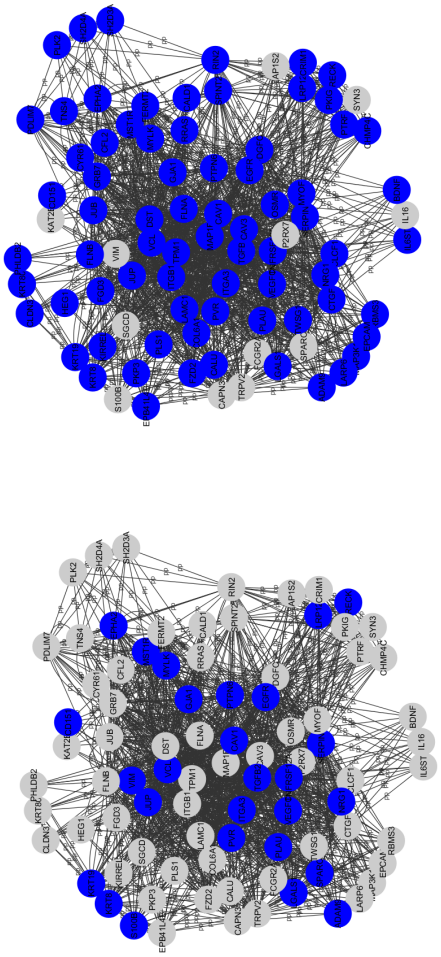
Se buscan términos específicos establecidos en la literatura y referentes a los tipos de cáncer de la base de datos NCI-60. Se compararon estas relaciones con las obtenidas a partir del análisis detallado en la sección anterior. Se observa que la mayoría de los genes encontrados se relacionan con más de un tipo de cáncer. La interacción de los genes, permite identificar genes que no han sido estudiados y tienen relación directa con genes que se expresan diferencialmente o mutan ante la presencia de alguna de las clases de cáncer.

En las Figuras 23 a 31, se muestran las comparaciones entre los genes reportados en la literatura y los obtenidos por este análisis relacionados a la presencia de los diferentes tejidos tumorales.

(a) Genes relacionados con Carcinoma de mama reportados por la literatura.

(b) Genes relacionados con Carcinoma de mama obtenidos por la metodología del trabajo.

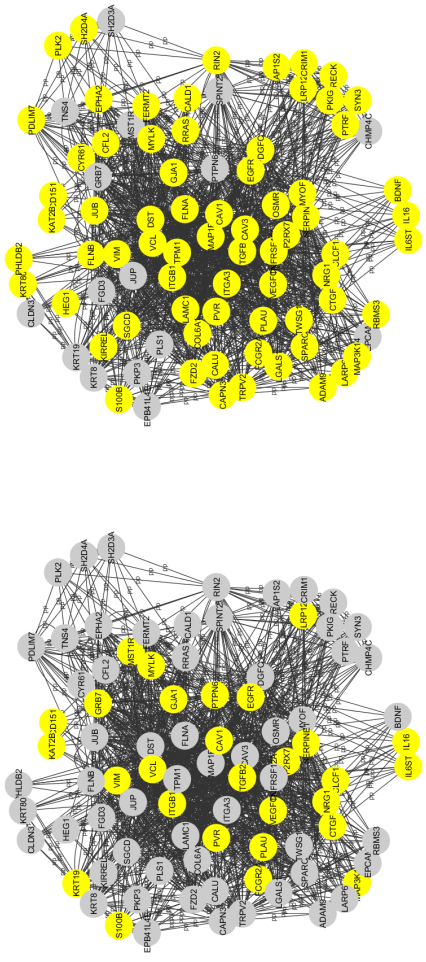
Figura 23: Genes relacionados a la presencia de carcinoma de mama.



(a) Genes relacionados con carcinoma de mama reportados por la literatura.

(b) Genes relacionados con carcinoma de mama obtenidos por la metodología del trabajo.

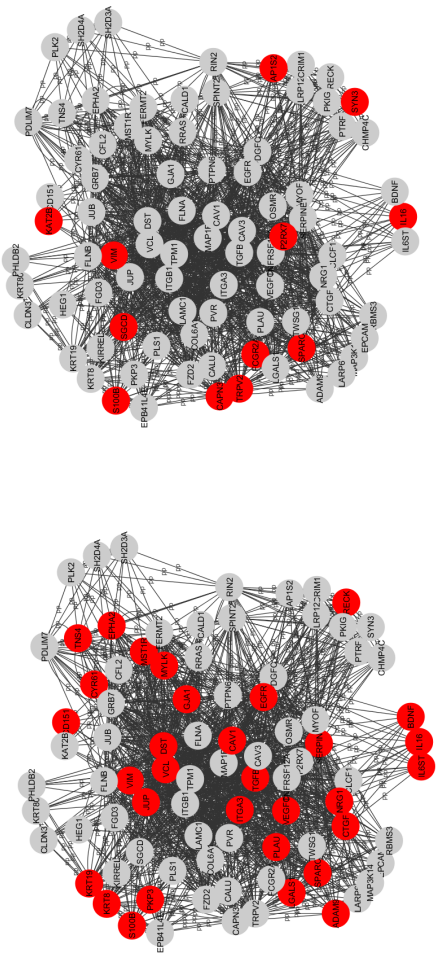
Figura 24: Genes relacionados a la presencia de carcinoma de colon.



(a) Genes relacionados con leucemia reportados por la literatura.

(b) Genes relacionados con leucemia obtenidos por la metodología del trabajo.

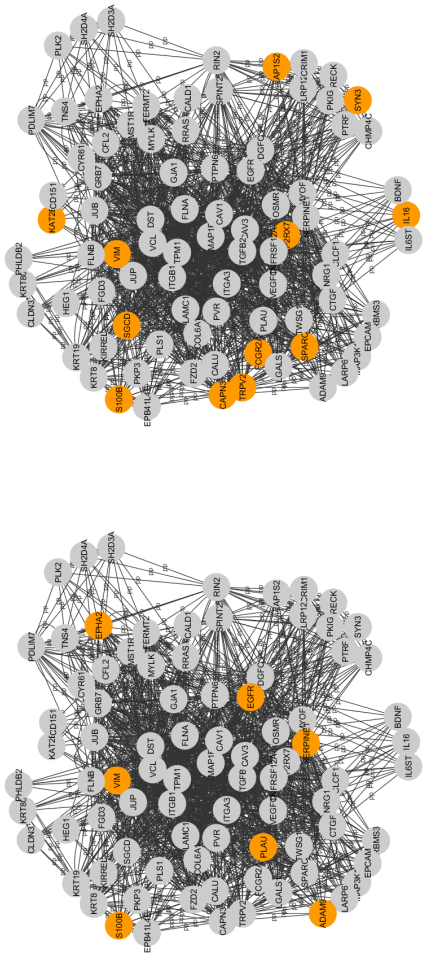
Figura 25: Genes relacionados a la presencia de leucemia.



(a) Genes relacionados con carcinoma de pulmón reportados por la literatura.

(b) Genes relacionados con carcinoma de pulmón obtenidos por la metodología del trabajo.

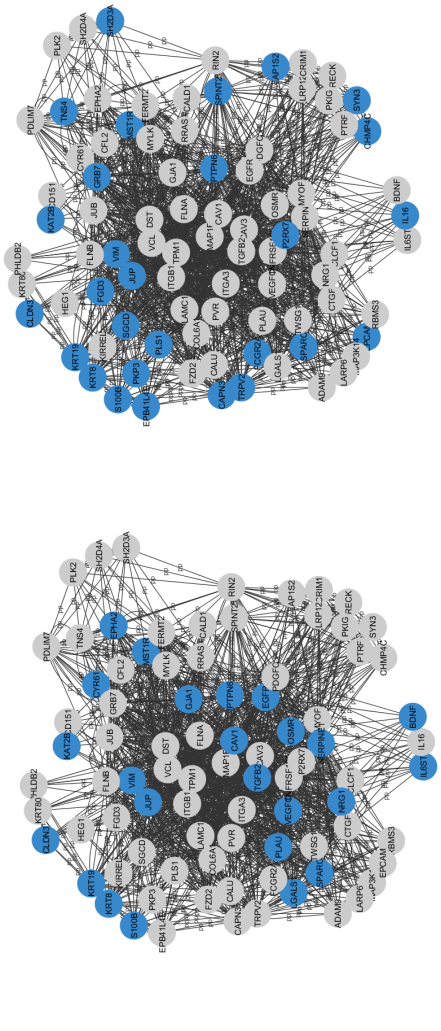
Figura 26: Genes relacionados a la presencia de carcinoma de pulmón.



(a) Genes relacionados con melanoma reportados por la literatura.

(b) Genes relacionados con melanoma obtenidos por la metodología del trabajo.

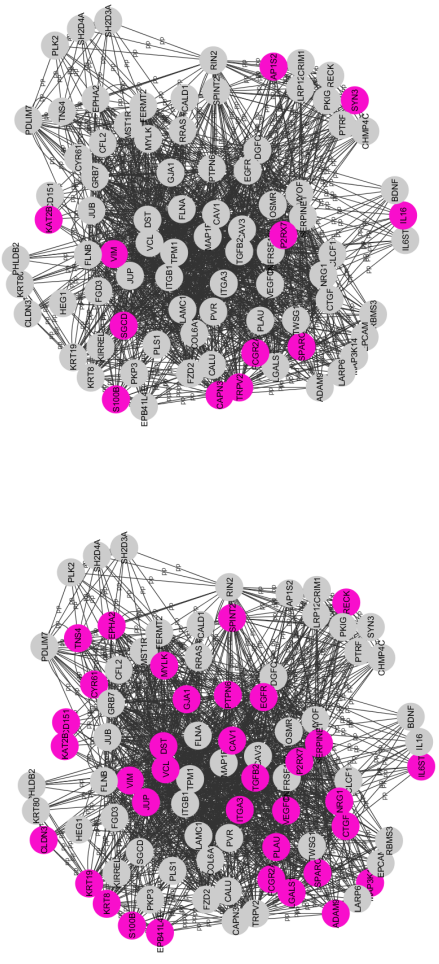
Figura 27: Genes relacionados a la presencia de melanoma.



(a) Genes relacionados con carcinoma de ovario reportados por la literatura.

(b) Genes relacionados con carcinoma de ovario obtenidos por la metodología del trabajo.

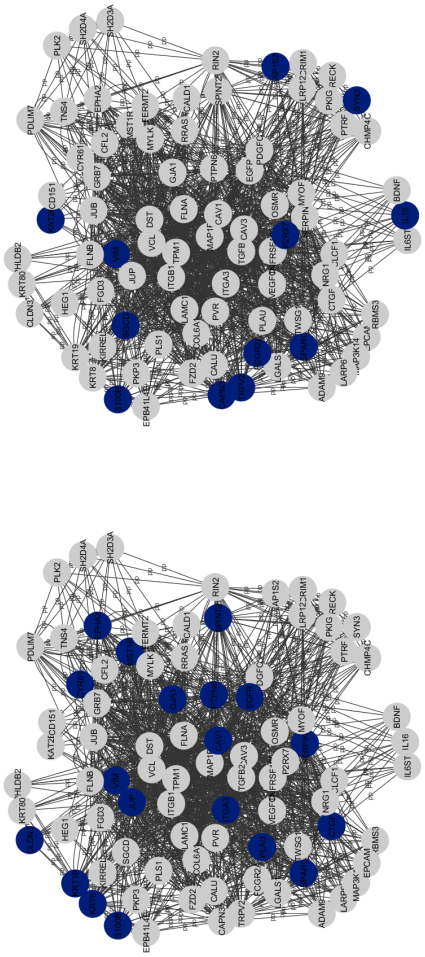
Figura 28: Genes relacionados a la presencia de carcinoma de ovarios.



(a) Genes relacionados con cáncer de próstata reportados por la literatura.

(b) Genes relacionados con cáncer de próstata obtenidos por la metodología del trabajo.

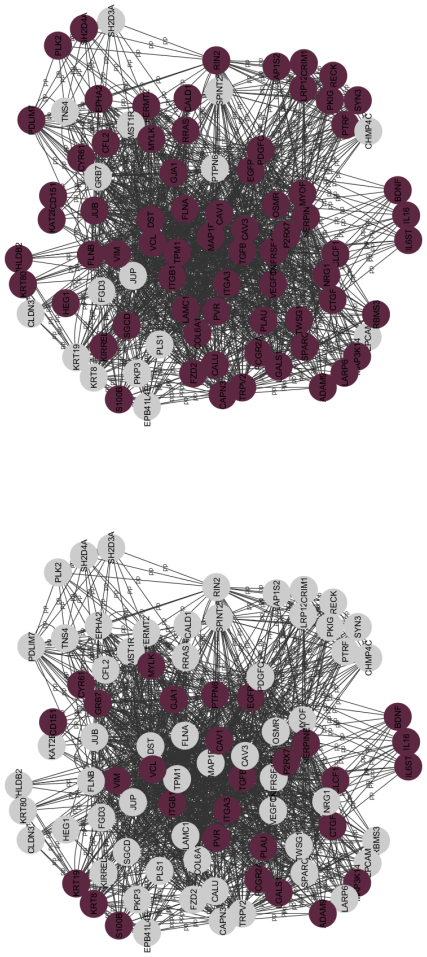
Figura 29: Genes relacionados a la presencia de cáncer de próstata.



(a) Genes relacionados con cáncer renal reportados por la literatura.

(b) Genes relacionados con cáncer renal obtenidos por la metodología del trabajo.

Figura 30: Genes relacionados a la presencia de cáncer renal.



(a) Genes relacionados con cáncer de SNC reportados por la literatura.

(b) Genes relacionados con cáncer de SNC obtenidos por la metodología del trabajo.

Figura 31: Genes relacionados a la presencia de cáncer de SNC.

La gran mayoría de los genes están asociados a alguno de los 13 términos nombrados en la sección 1.3, que se refieren a la presencia de enfermedades, muchos de ellos ligados de manera directa a la presencia de tumores, tales como: *"Cancer-Related-Conditions"*, *"Hamartoma"*, *"Hyperplasia"*, *"Neoplasm"*, *"Non-Neoplastic-Disorder"*. Estas relaciones se pueden ver en el Anexo.

Asimismo, en los gráficos se puede observar que hay un alto consenso entre los genes reportados anteriormente y los genes seleccionados en este estudio.

Es importante destacar, que en la literatura muchos de los genes reportados, se encuentran, de acuerdo al análisis realizado en este trabajo, agrupados a otros que no han sido reportados y que se relacionan a la presencia de los mismos tejidos. Por ejemplo, en trabajos anteriores, el cáncer de mama se encontró asociación con los genes AP1S2, LRP12 y RECK y en este trabajo, se determinó que los demás genes en el mismo grupo, también están asociados a la presencia de esta enfermedad. Lo mismo ocurre con el cáncer de colon, leucemia, próstata y pulmón.

En ambas redes, la distribución de los genes asociados a melanoma (tanto reportados en trabajos anteriores, como determinados en este análsis), muestran la alta variabilidad interna que presenta este tipo de cáncer,cuyos genes asociados pertenecen a diversos grupos.

Adicionalmente, se muestran los resultados de **FGNet** (Aibar *et al.*, 2015). En total, hay 18 grupos, 3 de los cuáles son excluidos por tener un coeficiente de silueta negativo (Rousseeuw, 1987), por lo que se analizan términos en 15 grupos. Los genes en color blanco en la Figura 32, pertenecen a más de un grupo simultáneamente (es decir, la mayoría de los genes).

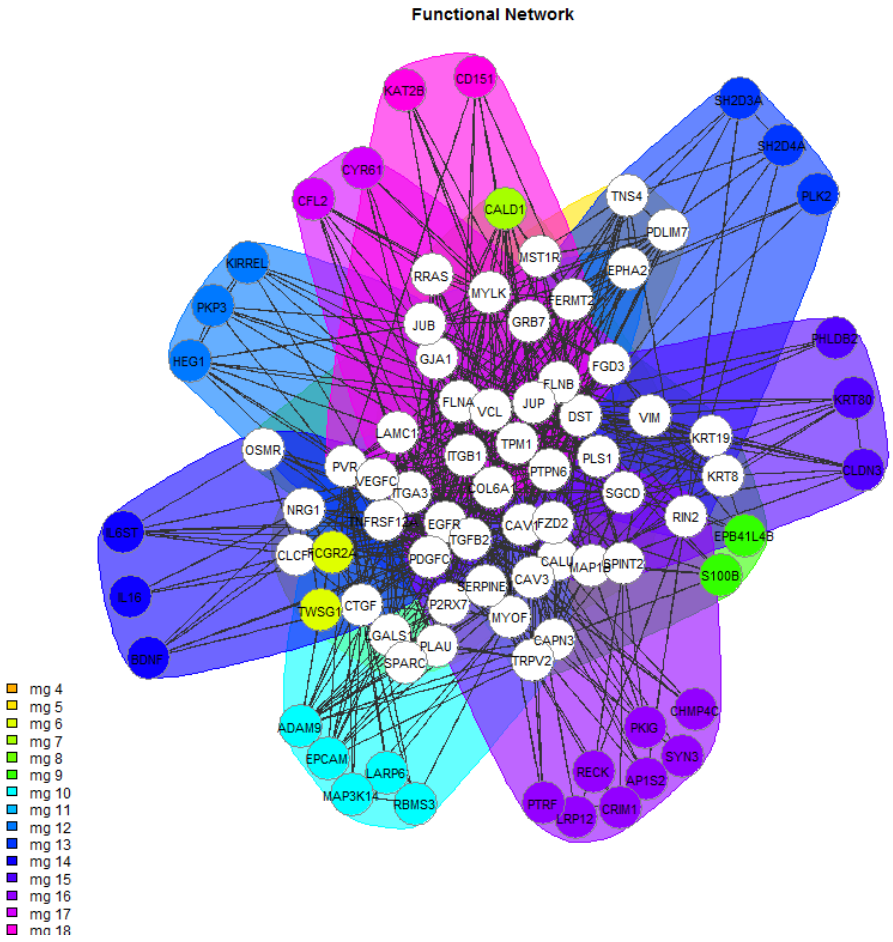


Figura 32: Red funcional: lista de genes seleccionados.

Los grupos más cercanos entre sí son el 5 y el 11, seguidos por el grupo 13. En el anexo, se puede ver la matriz de distancias entre ellos. Los 3 conglomerados nombrados están relacionados con el proceso *"Focal Adhesion"*, el cuál tiene un destacado rol en cáncer porque es un mediador, tanto de la proliferación y migración celular, como de la sobrevida de la célula y el desarrollo de una malignidad, está asociado con frecuencia a perturbaciones en estos procesos (McLean *et al.*, 2005). El listado completo de genes en estos grupos, se puede encontrar en la información complementaria.

Otro proceso biológico presente en uno de los grupos es *"Cell juntion assembly"*, estos son sitios de adhesión intracelular que mantienen la integridad de los tejidos epileliales, regulan la comunicación entre células y juegan un importante rol en la implementación, transformación e invasión tumoral (Knights *et al.*, 2012).

Otro término enriquecido es: *"Cytokine-cytokine receptor interaction"*, procesos que también están involucrados en el crecimiento, diferenciación, crecimiento celular y en la angiogénesis y que, por lo tanto, tiene un rol en la presencia de distintos tipos de cáncer (Culig, 2011).

También está enriquecido el término *"leishmaniasis"*, que es una enfermedad infecciosa; está demostrado que puede, de manera directa o indirecta, relacionarse a la presencia y proliferación de desórdenes malignos, especialmente de la piel y membranas mucosas (Kopterides *et al.*, 2007).

El término *"Axonogénesis"* está enriquecido en esta lista de genes. En la literatura, está reportada una relación entre este proceso biológico y el comportamiento agresivo del cáncer de próstata (Olar *et al.*, 2014; Ayala *et al.*, 2008).

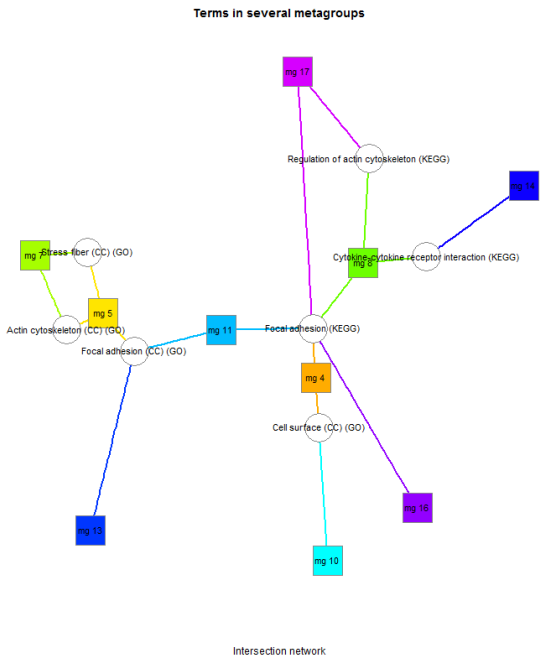


Figura 33: Términos enriquecidos presentes en varios grupos.

Otro término enriquecido, es el proceso celular conocido como *"Endocytosis"*, cuya desregulación probablemente está asociada a contribuir en una proliferación sostenida de las célular, una mayor invasividad y evitación de apoptosis (Mellman y Yarden, 2013), todos procesos involucrados con el cáncer. La Figura 33, muestra las relaciones entre términos que están en varios grupos.

Es decir, en los grupos derivados del análisis de enriquecimiento funcional, se encuentran enriquecidos términos relacionados a la presencia del cáncer, lo que también indica y complementa la información obtenida anteriormente de que esta lista no es azarosa y contribuye a validar este método como una alternativa novedosa para integrar información de distintas fuentes de datos y seleccionar genes.

Por otra parte, es importante destacar que, posiblemente los 58 genes de la lista que no están presentes en el *"Cancer Gene Index"*, estén muy probablemente asociados a alguno de los nueve tipos de cáncer aquí tratados. Para determinar la asociación, sería necesario llevar adelante estudios confirmatorios.

2.5 Discusión

La combinación de los métodos STATIS, representación Biplot, minería de textos y redes de co-expresión, favorecen la integración de sub-espacios y el estudio simultáneo de las relaciones entre individuos, variables y variables e individuos en el espacio consenso. Asimismo, posibilitan la selección de genes candidatos y la búsqueda de ontologías y términos enriquecidos, como también de las relaciones previamente establecidas entre los genes y las distintas enfermedades, incrementando, de este modo, la calidad y cantidad de información otorgada, respecto de la que otorgan los métodos comúnmente empleados.

A continuación, se presenta un cuadro que compara los métodos STATIS y STATIS DUAL clásicos y el enfoque aquí desarrollado.

Cuadro 4: Comparación entre los enfoques clásico y propuesto.

Criterios de comparación	**STATIS/STATIS DUAL (enfoque clásico)**	**Combinación de métodos (enfoque propuesto)**
1. Estudio de las relaciones entre individuos.	✓	✓
2. Estudio de la calidad de representación de los individuos.		✓
3. Estudio de la variabilidad muestral.		✓
4. Estudio de las relaciones entre las variables.	✓	✓
5. Estudio de las relaciones entre individuos y variables.		✓
6. Selección de variables.		✓
7. Estudio de las relaciones entre las variables seleccionadas y los individuos.		✓
8. Redes de co-expresión entre las variables seleccionadas.		✓

Conclusiones

El trabajo realiza una contribución a la integración de múltiples conjuntos de datos ómicos. Desarrolla una propuesta metodológica que considera las propiedades de las distintas clases de datos e introduce mejoras que permiten cuantificar la sensibilidad de los distintos métodos de k-tablas, estudiando su variabilidad a partir del uso de técnicas de re-muestreo y medidas de calidad de representación de individuos y grupos.

Además, se implementan métodos Biplot que permiten responder preguntas respecto de la relación entre tejidos tumorales, genes, tejidos y genes, así como también, tejidos, genes y términos implicados en el desarrollo de una enfermedad (en este caso, el cáncer) y métodos de minería de texto para buscar información disponible en la literatura respecto de los genes seleccionados por el método.

Bajo estas condiciones se concluye:

1. Actualmente, se dispone de un gran caudal de información de datos ómicos y su integración es muy importante para tener un conocimiento más completo del sistema bajo estudio; no obstante, en la literatura se observa un atavismo a las mismas técnicas de análisis,

que no permiten responder muchos de los interrogantes de interés de los investigadores. Por lo tanto, en el marco de este trabajo se abordó el problema usando distintas metodologías de k-tablas, cuya utilización en el área era escasa hasta el momento.

2. Se introdujo el uso de métodos de remuestreo sobre los individuos y/o variables proyectados en el compromiso. De esta manera, es posible el estudio de la variabilidad muestral en el plano bi o tri-dimensional y realizar un análisis de la estabilidad de los resultados.

3. Se introdujo el cálculo de medidas de calidad de representación sobre el compromiso, de individuos y variables.

4. Se introdujo el uso del método Biplot que permite la proyección de todos los genes estudiados sobre el espacio compromiso, logrando así una representación conjunta del individuo consenso (tejidos tumorales) y las variables (genes).

5. La aproximación Biplot, permite seleccionar genes responsables de la estructura común de los individuos proyectados. Además, en casos donde existe estructura de grupos, es posible identificar genes responsables o que se sobre-expresan en cada uno de ellos.

6. La representación gráfica, producto de la combinación de las metodologías STATIS y los métodos Biplot, favorece el estudio simultáneo de las relaciones entre tejidos tumorales, entre genes, entre tejidos y genes y entre tejidos genes y términos anotados sobre la enfermedad.

7. El uso de los métodos Biplot constituye una novedosa herramienta para la selección de genes candidatos, potencialmente ligados a la presencia y desarrollo de la enfermedad.

8. El análisis de los grafos, que se obtienen a partir del análisis de enriquecimiento funcional de los genes seleccionados y la búsqueda de términos en la literatura, comparados con las asociaciones entre los

genes y tejidos derivadas de los métodos Biplot, permiten conjeturar sobre la interacción de grupos de genes en la presencia de distintos tipos de tejidos, como también muestran las similaridades existentes entre algunas clases de cáncer.

9. Finalmente, esta metodología ofrece una comprensión holística de la estructura de datos y facilita las interpretaciones de los resultados. En este sentido, se recomienda su utilización en el abordaje de problemas que requieren integración de datos provenientes de múltiples fuentes, donde existe alguna configuración en común (individuos y/o variables).

Bibliografía

Abdi, H., Valentin, D., Chollet, S., y Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference*, 18(4):627–640.

Abdi, H., Williams, L. J., Valentin, D., y Bennani-Dosse, M. (2012). STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):124–167.

Acharjee, A. (2013). *Systems biology and statistical data integration of˜ omics data sets*. Wageningen University.

Aibar, S., Fontanillo, C., Droste, C., y De Las Rivas, J. (2015). Functional gene networks: R/bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics*, 31(10):1686–1688.

Alberts, B. B., Lewis, D., Raff, J., Roberts, M., Watson, K., Alberts, J. D. B., *et al.* (1996). *Biología molecular de la célula.* Number 576.3 BIO.

Alibés, A., Yankilevich, P., Cañada, A., y Díaz-Uriarte, R. (2007). IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC bioinformatics*, 8(1):9.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry,

J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Ayala, G. E., Dai, H., Powell, M., Li, R., Ding, Y., Wheeler, T. M., Shine, D., Kadmon, D., Thompson, T., Miles, B. J., *et al.* (2008). Cancer-related axonogenesis and neurogenesis in prostate cancer. *Clinical Cancer Research*, 14(23):7593–7603.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M., *et al.* (2004). The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer*, 91(2):355–358.

Belacel, N., Wang, Q., y Cuperlovic-Culf, M. (2006). Clustering methods for microarray gene expression data. *Omics : a journal of integrative biology*, 10(4):507–531.

Burguillo, F. J., Martin, J., Barrera, I., y Bardsley, W. G. (2010). Meta-analysis of microarray data: The case of imatinib resistance in chronic myelogenous leukemia. *Computational biology and chemistry*, 34(3):184–92.

Burkard, M. E. (2012). Integrating the nci-60 data with omics for drug discovery. *Drug Development Research*, 73(7):420–429.

Cárdenas, O., Noguera, J. L., y Vicente-villardón, J. (2003). María . P Galindo. IX(2):257–276.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., y D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic acids research*, 42(Database issue):D472–7.

Cui, X. y Churchill, G. a. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome biology*, 4(4):210.

Culhane, A. C., Perrière, G., y Higgins, D. G. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC bioinformatics*, 4(1):1.

Culig, Z. (2011). Cytokine disbalance in common human cancers. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1813(2):308 – 314.

Demey, J. R. (2008). *DIVERSIDAD GENETICA EN BANCOS DE GERMOPLASMA : UN ENFOQUE BIPLOT*. PhD thesis, Universidad de Salamanca.

Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P., y Zambrano, a. Y. (2008). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics (Oxford, England)*, 24(24):2832–8.

Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., Lempicki, R. A., *et al.* (2003). David: database for annotation, visualization, and integrated discovery. *Genome biol*, 4(5):P3.

des Plantes, H. L. (1976). *Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe*. PhD thesis, Université des sciences et techniques du Languedoc.

Di Rienzo, J., Casanoves, F., Balzarini, M., Gonzalez, L., Tablada, M., y Robledo, C. (2011). Infostat.

Diaz-Uriarte, R. (2007). GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC bioinformatics*, 8(1):328.

Efron, B. (1987). *The jackknife, the bootstrap, and other resampling plans*. CBMS-NSF Regional conference series in applied mathematics 38. Society for Industrial and Applied Mathematics.

Efron, B. y Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, pages 751–760.

Escoufier, Y., Cazes, P., *et al.* (1976). Opérateurs et analyse des tableaux à plus de deux dimensions. *Cahiers du bureau universitaire de recherche opérationnelle*, 25:61–89.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., y Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5):E359–E386.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., *et al.* (2010). Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, page gkq929.

Gabriel, K. R. (1971). The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*, 58(3):453.

Gautier, L., Cope, L., Bolstad, B. M., y Irizarry, R. A. (2004). affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315.

Ginsburg, G. S. y McCarthy, J. J. (2001). Personalized medicine: revolutionizing drug discovery and patient care. *TRENDS in Biotechnology*, 19(12):491–496.

Goldstein, D. y Guerra, R. (2010). A brief introduction to meta-analysis, genetics and genomics. *Meta-Analysis and Combining Information in Genetics and Genomics*, 1:3–20.

Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., y Bruford, E. A. (2014).

Genenames. org: the hgnc resources in 2015. *Nucleic acids research*, page gku1071.

Grimaldi, R. P. (1998). *Matemáticas discreta y combinatoria: introducción y aplicaciones*. Pearson Educación.

Gundem, G., Perez-Llamas, C., Jene-Sanz, A., Kedzierska, A., Islam, A., Deu-Pons, J., Furney, S. J., y Lopez-Bigas, N. (2010). Intogen: integration and data mining of multidimensional oncogenomic data. *Nature methods*, 7(2):92–93.

Guo, J., Miao, Y., Xiao, B., Huan, R., Jiang, Z., Meng, D., y Wang, Y. (2009). Differential expression of microrna species in human gastric cancer versus non-tumorous tissues. *Journal of gastroenterology and hepatology*, 24(4):652–657.

Hood, L., Balling, R., y Auffray, C. (2012). Revolutionizing medicine in the 21st century through systems approaches. *Biotechnology journal*, 7(8):992–1001.

Huang, D. W., Sherman, B. T., y Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., y Lempicki, R. a. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, 35(Web Server issue):W169–75.

Ibrahim, A. F., Weirauch, U., Thomas, M., Grünweller, A., Hartmann, R. K., y Aigner, A. (2011). Microrna replacement therapy for mir-145 and mir-33a is efficacious in a model of colon carcinoma. *Cancer research*, 71(15):5214–5224.

Irizarry, R. a. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):15e–15.

Jaffrenou, P.-A. (1978). *Sur l'analyse des familles finies de variables vectorielles: bases algébriques et application à la description statistique*. PhD thesis.

Jia, P., Liu, Y., y Zhao, Z. (2012). Integrative pathway analysis of genome-wide association studies and gene expression data in prostate cancer. *BMC systems biology*, 6 Suppl 3(Suppl 3):S13.

Joyce, A. R. y Palsson, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nature reviews. Molecular cell biology*, 7(3):198–210.

Kanehisa, M. y Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Knights, A. J., Funnell, A. P., Crossley, M., y Pearson, R. C. (2012). Holding tight: cell junctions and cancer spread. *Trends in cancer research*, 8:61.

Kohl, M., Megger, D. A., Trippler, M., Meckel, H., Ahrens, M., Bracht, T., Weber, F., Hoffmann, A.-C., Baba, H. A., Sitek, B., *et al.* (2014). A practical data processing workflow for multi-omics projects. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1844(1):52–62.

Kopterides, P., Mourtzoukou, E. G., Skopelitis, E., Tsavaris, N., y Falagas, M. E. (2007). Aspects of the association between leishmaniasis and malignant disorders. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 101(12):1181–1189.

Lavit, C., Escoufier, Y., Sabatier, R., y Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis*, 18(1):97–119.

Lebart, L. (2007). Which bootstrap for principal axes methods? In *Selected Contributions in Data Analysis and Classification*, pages 581–588. Springer.

Liu, J., Huang, J., y Ma, S. (2013a). Integrative analysis of multiple cancer genomic datasets under the heterogeneity model. *Statistics in medicine*, 32(20):3509–21.

Liu, Y., Devescovi, V., Chen, S., y Nardini, C. (2013b). Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC systems biology*, 7:14.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., *et al.* (2005). Microrna expression profiles classify human cancers. *nature*, 435(7043):834–838.

Mayer, G., Jones, A. R., Binz, P.-A., Deutsch, E. W., Orchard, S., Montecchi-Palazzi, L., Vizcaíno, J. A., Hermjakob, H., Oveillero, D., Julian, R., Stephan, C., Meyer, H. E., y Eisenacher, M. (2014). Controlled vocabularies and ontologies in proteomics: overview, principles and practice. *Biochimica et biophysica acta*, 1844(1 Pt A):98–107.

McLean, G. W., Carragher, N. O., Avizienyte, E., Evans, J., Brunton, V. G., y Frame, M. C. (2005). The role of focal-adhesion kinase in cancera new therapeutic opportunity. *Nature Reviews Cancer*, 5(7):505–515.

Mellman, I. y Yarden, Y. (2013). Endocytosis and cancer. *Cold Spring Harbor perspectives in biology*, 5(12):a016949.

Meng, C. y Gholami, A. M. (2014). Multiple Co-inertia Analysis of Multiple OMICS Data using omicade4. pages 1–6.

Meng, C., Kuster, B., Culhane, A. C., y Moghaddas Gholami, A. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC bioinformatics*, 15:162.

Moghaddas Gholami, A., Hahne, H., Wu, Z., Auer, F. J., Meng, C., Wilhelm, M., y Kuster, B. (2013). Global proteome analysis of the NCI-60 cell line panel. *Cell reports*, 4(3):609–20.

NCI (2014a). Creation of the cancer gene index. cancer gene index end user documentation. [Web]; Recuperado el 06 de Febrero de 2015 de https://wiki.nci.nih.gov/display/cageneindex/Creation+of+the+Cancer+Gene+Index.

NCI (2014b). Icr-cancer gene index. cancer gene index end user documentation. [Web]; Recuperado el 07 de Febrero de 2015 de https://wiki.nci.nih.gov/display/cageneindex/Cancer+Gene+Index+End+User+Documentation.

NCI (2014c). Seer training modules, cancer as a disease. u. s. national institutes of health, national cancer institute. [Web]: Recuperado el 06 de Febrero de 2015 de http://training.seer.cancer.gov/.

Ng, E. K., Chong, W. W., Lam, E. K., Shin, V. Y., Yu, J., Poon, T. C., Ng, S. S., Sung, J. J., *et al.* (2009). Differential expression of microrna s in plasma of colorectal cancer patients: a potential marker for colorectal cancer screening. *Gut*.

Nguyen, D. V. y Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50.

Olar, A., He, D., Florentin, D., Ding, Y., y Ayala, G. (2014). Biologic correlates and significance of axonogenesis in prostate cancer. *Human pathology*, 45(7):1358–1364.

OMS (2015). Cáncer. [Web]; Recuperado el 10 de Marzo de 2015 de http://www.who.int/mediacentre/factsheets/fs297/es/.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., y Gabrielson, E. (2011). A statistical framework for expression-based molecular classification in cancer. (2002):717–736.

Pastrello, C., Pasini, E., Kotlyar, M., Otasek, D., Wong, S., Sangrar, W., Rahmati, S., y Jurisica, I. (2014). Integration, visualization and analysis

of human interactome. *Biochemical and biophysical research communications*, 445(4):757–73.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., y Wain, H. (2001). The HUGO Gene Nomenclature Committee (HGNC). *Human genetics*, 109(6):678–80.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rakotomamonjy, A. (2003). Variable Selection Using SVM-based Criteria. *Journal ofMachine Learning Research*, 3:1357–1370.

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., Doroshow, J., y Pommier, Y. (2012). CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer research*, 72(14):3499–511.

Reuters, T. (2012). Web of science.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., y Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Shankavaram, U. T., Varma, S., Kane, D., Sunshine, M., Chary, K. K., Reinhold, W. C., Pommier, Y., y Weinstein, J. N. (2009). CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC genomics*, 10:277.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., y Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

Siegel, R., Ma, J., Zou, Z., y Jemal, A. (2014). Cancer Statistics , 2014. 64(1):9–29.

Soneson, C., Lilljebjorn, H., Fioretos, T., y Fontes, M. (2010). Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics*, 11:191.

Suárez-Fariñas, M., Noggle, S., Heke, M., Hemmati-Brivanlou, A., y Magnasco, M. O. (2005). Comparing independent microarray studies: the case of human embryonic stem cells. *BMC genomics*, 6:99.

Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. *The Annals of Applied Statistics*, 5(4):2300–2325.

Thioulouse, J. y Chessel, D. (1987). Les analyses multitableaux en écologie factorielle 1. - De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. 8(4):463–480.

Thioulouse, J., Simier, M., y Chessel, D. (2004). Simultaneous analysis of a sequence of paired ecological tables. *Ecology*, 85(1):272–283.

Thomas, M., Lange-Grünweller, K., Weirauch, U., Gutsch, D., Aigner, A., Grünweller, A., y Hartmann, R. (2012). The proto-oncogene pim-1 is a target of mir-33a. *Oncogene*, 31(7):918–928.

Tuikkala, J., Vähämaa, H., Salmela, P., Nevalainen, O. S., y Aittokallio, T. (2012). A multilevel layout algorithm for visualizing physical and genetic interaction networks, with emphasis on their modular organization. *BioData mining*, 5(1):1.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. In *Annals of Mathematical statistics*, volume 29, pages 614–614. INST MATHEMATICAL STATISTICS IMS BUSINESS OFFICE-SUITE 7, 3401 INVESTMENT BLVD, HAYWARD, CA 94545.

Vicente-Villardón, J. L., Galindo Villardón, M. P., Blázquez Zaballos, A.,

Greenacre, M., y Blasisus, J. (2006). Logistic biplots. *Multiple correspondence analysis and related methods. London: Chapman & Hall*, pages 503–521.

Vivien, M. y Sabatier, R. (2004). A generalization of STATIS-ACT strategy: DO-ACT for two multiblocks tables. *Computational Statistics & Data Analysis*, 46(1):155–171.

Wei, S., Wang, Y., Xu, H., y Kuang, Y. (2015). Screening of potential biomarkers for chemoresistant ovarian carcinoma with mirna expression profiling data by bioinformatics approach. *Oncology letters*, 10(4):2427–2431.

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., y Hampton, G. M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer research*, 61(16):5974–5978.

Wit, E. y McClure, J. (2004). *Front matter*. Wiley Online Library.

yWorks GMBH (2013). Automatic graph layout. [Web]; Recuperado el 03 de Marzo de 2015 de http://www.who.int/mediacentre/factsheets/fs297/es/.

Zeeberg, B. R., Riss, J., Kane, D. W., Bussey, K. J., Uchio, E., Linehan, W. M., Barrett, J. C., y Weinstein, J. N. (2004). Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC bioinformatics*, 5:80.

Zhang, W., Li, F., y Nie, L. (2010). Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology (Reading, England)*, 156(Pt 2):287–301.

Zingaretti, M. L., Demey-Zambrano, J. A., Vicente-Villardon, J. L., y Demey, J. R. (2015). *kimod: A k-tables approach to integrate multiple Omics-Data*. R package version 0.99.5.

Anexo

Metagrupo	Silhouette	P-value	Genes
Metagroup 4	0.11	1.30E-17	
Arrhythmogenic right ventricular cardiomyopathy (ARVC)			
Cell surface (CC) *			CAV1, CAV3, COL6A1, EGFR, FLNA, GJA1,
Dilated cardiomyopathy			GRB7, ITGA3, ITGB1, JUP, LAMC1, PLAU,
ECM-receptor interaction			PTPN6, SGCD, TGFB2, TPM1, VCL, VEGFC
Focal adhesion *			
Hypertrophic cardiomyopathy (HCM)			
Leukocyte migration (BP)			
Metagroup 5	0.17	1.80E-15	
Actin cytoskeleton (CC) *			CAV1, DST, EPHA2, FERMT2, FLNA,
Cell cortex (CC)			FLNB, GRB7, ITGB1, JUB, MST1R,
Focal adhesion (CC) *			MYLK, PDLIM7, TNS4, TPM1, VCL
Stress fiber (CC) *			
Metagroup 6	0.05	4.50E-14	
Embryonic development (BP)			
Extracellular matrix (CC)			CALU, CAV1, CLCF1, COL6A1, CTGF, DSTSPARC,
Leishmaniasis			FCGR2A, FLNA, FZD2, ITGB1, LAMC1,
Platelet alpha granule lumen (CC)			LGALS1, MAP1B, NRG1, P2RX7, PTPN6, SERPINE1,
Platelet degranulation (BP)			TGFB2, TPM1, TWSG1, VCL, VEGFC
Receptor binding (MF)			
Response to wounding (BP)			
Wound healing (BP)			
Metagroup 7	0.28	6.90E-14	
Actin cytoskeleton (CC) *			CALD1, DST, FERMT2, FLNA, GJA1, JUP,
Cellular component movement (BP)			MST1R, MYLK, MYOF, RRAS, SERPINE1,
Muscle contraction (BP)			SPINT2, TNFRSF12A, TPM1, VCL, VIM
Negative regulation of cell migration (BP)			
Stress fiber (CC) *			
Metagroup 8	0.18	8.40E-14	
Actin filament binding (MF)			COL6A1, EGFR, FLNA, FLNB, ITGA3,
Cytokine-cytokine receptor interaction *			ITGB1, JUB, LAMC1, OSMR, PDGFC, PLS1,
Focal adhesion *			TNFRSF12A, VCL, VEGFC
Regulation of actin cytoskeleton *			
Metagroup 9	0	1.30E-13	
Axon (CC)			
Axonogenesis (BP)			
Calcium-binding EF-hand			CALU, CAPN3, CAV3, COL6A1, DST, EPB41L4B,
Cytoskeleton organization (BP)			FGD3, FLNB, FZD2, ITGB1, JUP, KRT19,
Muscle organ development (BP)			KRT8, MAP1B, PLS1, S100B, SGCD,
Sarcolemma (CC)			TGFB2, TPM1, TRPV2, VIM
Sarcomere organization (BP)			
Structural constituent of cytoskeleton (MF)			
Z disc (CC)			
Metagroup 10	:0.12	3.60E-12	
Cell surface (CC) *			ADAM9, CAPN3, CAV1, CAV3, CTGF,
Intracellular membrane-bounded organelle (CC)			EPCAM, ITGA3, ITGB1, LARP6, LGALS1,
Response to calcium ion (BP)			MAP3K14, MYOF, P2RX7, PLAU, PVR,
Response to glucocorticoid stimulus (BP)			RBMS3, SERPINE1, SPARC, TNFRSF12A, TRPV2
Metagroup 11	0.44	5.80E-10	
Bacterial invasion of epithelial cells			CAV1, CAV3, EPHA2, FERMT2,
Focal adhesion *			FLNB, GRB7, ITGB1, JUB, VCL
Focal adhesion (CC) *			
Metagroup 12	0.2	6.60E-09	
Cell-cell junction (CC)			HEG1, ITGB1, JUB, JUP, KIRREL,
Cell-cell junction organization (BP)			LAMC1, P2RX7, PKP3, PVR, TGFB2, VCL
Cell migration (BP)			
Metagroup 13	0.31	6.90E-09	
Cell projection (CC)			DST, EPHA2, FERMT2, GRB7,
Focal adhesion (CC) *			PDLIM7, PLK2, PTPN6, RIN2,
SH2 motif			SH2D3A, SH2D4A, TNS4
Metagroup 14	0.37	7.20E-09	
Cytokine-cytokine receptor interaction *			BDNF, CLCF1, CTGF, EGFR,
Cytokine activity (MF)			IL16, IL6ST, NRG1, OSMR, PDGFC,
Growth factor activity (MF)			TGFB2, TNFRSF12A, VEGFC
Metagroup 15	0.37	2.70E-08	
Intermediate filament (CC)			CAV1, CLDN3, DST, JUP,
Intermediate filament cytoskeleton (CC)			KRT19, KRT8, KRT80, MAP1B,
Keratin, type I			PHLDB2, VCL, VIM
Structural molecule activity (MF)			
Metagroup 16	0.13	3.10E-08	
Caveola (CC)			
Cytoplasmic vesicle (CC)			AP1S2, CAV1, CAV3, CHMP4C,
Endocytosis			CRIM1, EGFR, LRP12, MAP1B, MYOF,
Endocytosis (BP)			P2RX7, PDGFC, PKIG, PTPN6, PTRF,
Focal adhesion *			RECK, RIN2, SERPINE1, SPINT2, SYN3, TGFB2
Negative regulation of MAPKKK cascade (BP)			
Serine-type endopeptidase inhibitor activity (MF)			
Soluble fraction (CC)			
Metagroup 17	0.36	8.70E-08	
Focal adhesion *			CFL2, CYR61, EGFR, FGD3, GRB7,
Positive regulation of cell migration (BP)			ITGA3, ITGB1, MYLK, PDGFC, RRAS, VCL
Regulation of actin cytoskeleton *			
Metagroup 18	0.44	5.20E-07	CD151, FERMT2, FLNA, GRB7,
Cell junction assembly (BP)			ITGB1, JUP, KAT2B, PTPN6, PVR

Cuadro 5: Funciones y genes dentro de los grupos.

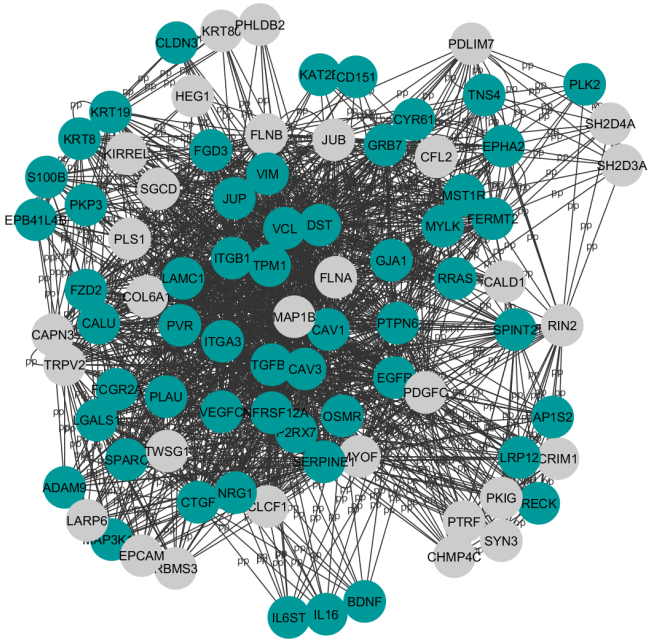


Figura 34: Lista de genes relacionados con carcinoma por la literatura.

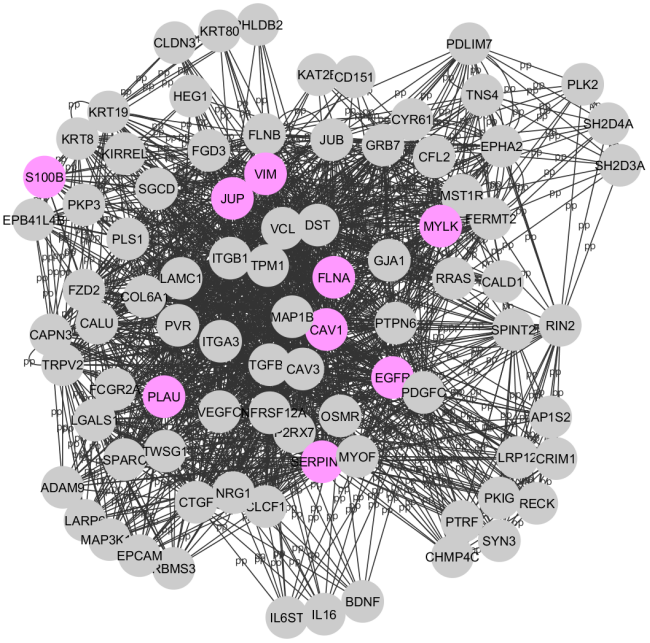


Figura 35: Lista de genes relacionados con Genetical Disorden en la literatura.

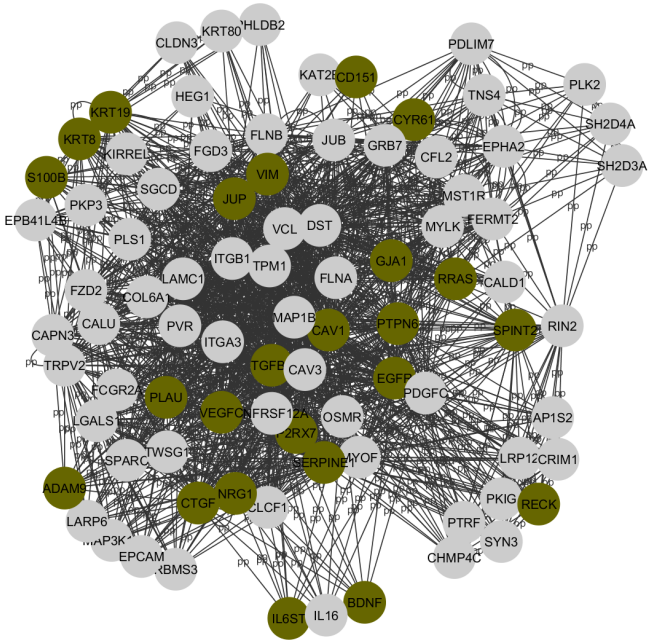


Figura 36: Lista de genes relacionados con hiperplasia por la literatura.

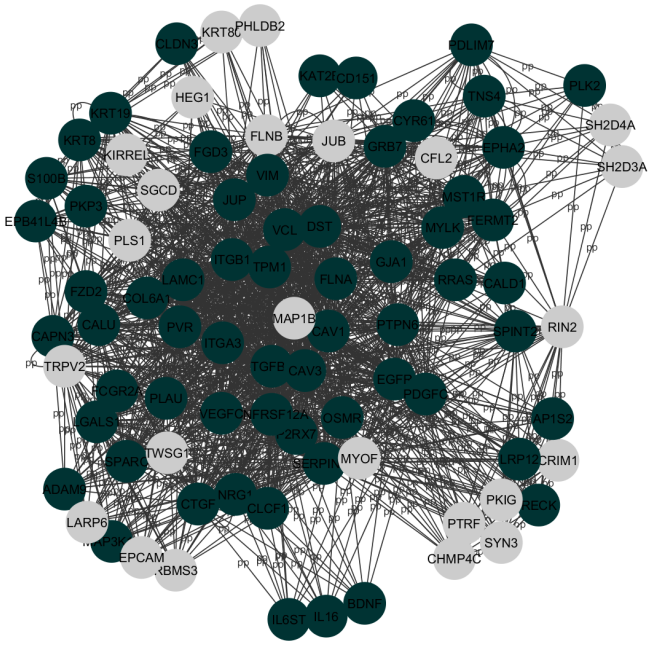


Figura 37: Lista de genes relacionados con neoplasma por la literatura.

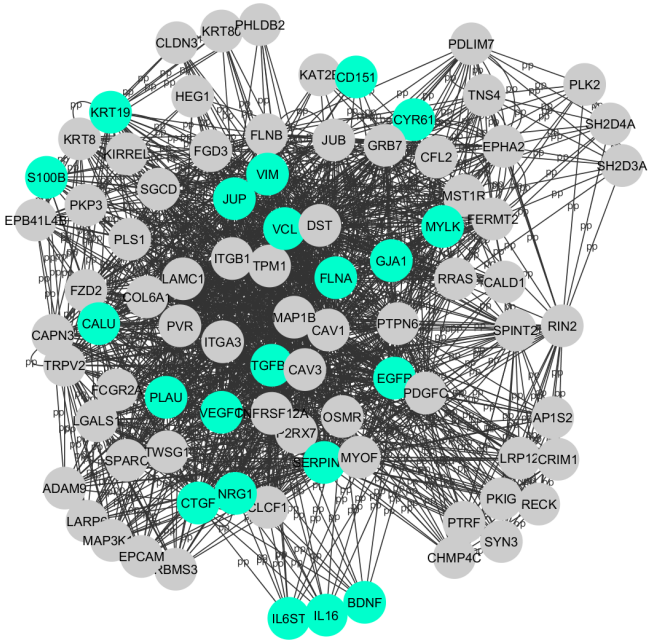


Figura 38: Lista de genes relacionados con NonNeoplastic por la literatura.

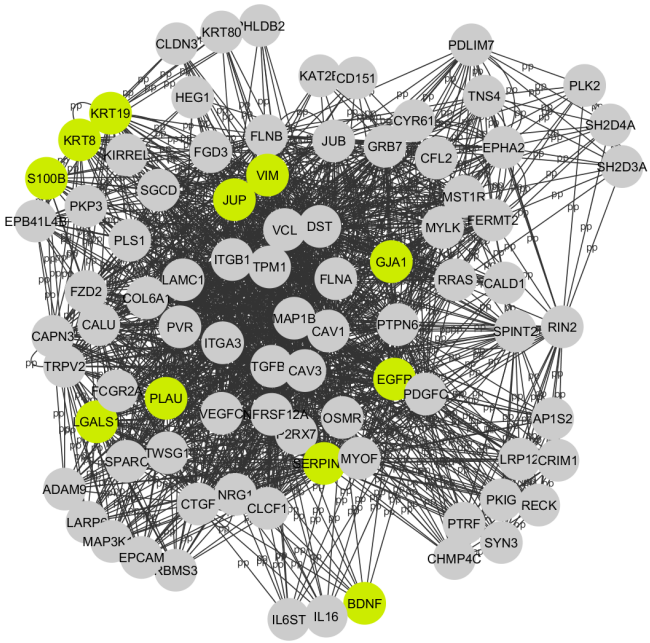


Figura 39: Lista de genes relacionados con polipos por la literatura.

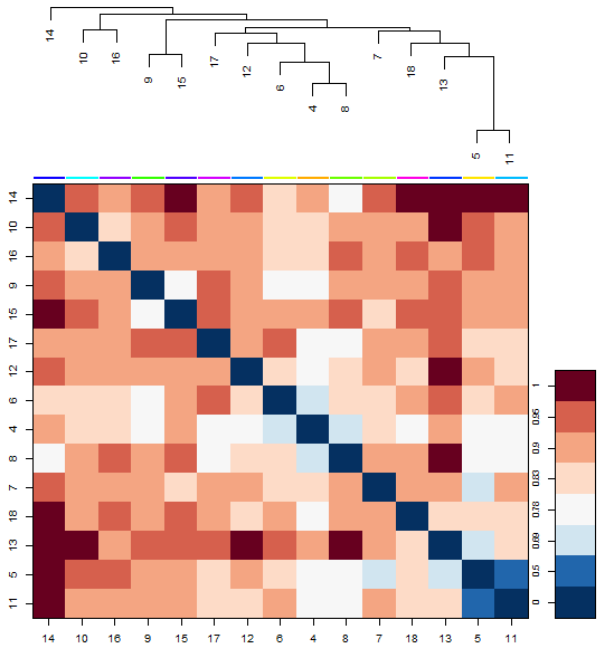


Figura 40: Distancia entre los grupos derivados del análisis de ontologías.