Universitat Politècnica de València ETSIAMN Centro de Investigación Príncipe Felipe

TRABAJO FIN DE GRADO

GRADO EN BIOTECNOLOGÍA

ALTERACIONES EN LA COMUNICACIÓN ENTRE RUTAS DE SEÑALIZACIÓN: UN ESTUDIO TRANSVERSAL EN CÁNCER

PAU NEBOT FORCADA

CURSO 2015-2016

Tutores

José Carbonell Caballero, Departamento de Genómica de Sistemas, Centro de Investigación Príncipe Felipe (CIPF).

Joaquín Dopazo Blázquez, Departamento de Genómica de Sistemas, Centro de Investigación Príncipe Felipe (CIPF).

Jose Javier Forment Millet, Instituto de Biología Molecular y Celular de Plantas (IBMCP), UPV-CSIC.

Valencia, 19 de septiembre de 2016.

Título del TFG: Alteraciones en la comunicación entre rutas de señalización: un estudio transversal en cáncer.

Resumen (inglés):
Cancer is described as a set of diseases characterized by uncontrolled tissue development, where the affected cells evade the intrinsic control mechanisms against cell death or controlled proliferation. Those are multi-gene diseases where a series of mutations in key genes are responsible for the direction of the tumor development process.
In this context, systems biology provides the ideal framework for the study of cancer because it allows to describe biological function alterations as changes in the system, beyond the individual elements affected. Of particular relevance is the study of signaling pathways as they are involved in essential functions such as cell cycle.
This work focuses on the study of signaling pathways and their changes in a wide range of cancers collected through two international consortia (TCGA and ICGC).
The used methodology of analysis decomposes the set of pathways in signaling cascades responsible for carrying out specific functions. Of particular relevance in the work is the determination of communication changes between cascades with related functions as a result of the disease, as well as other specific parameters such as tumor subtype, stage or survival chance.

Resumen (español):
El cáncer se describe como un conjunto de enfermedades caracterizadas por un desarrollo tisular incontrolado, donde las células afectadas consiguen evadir los mecanismos intrínsecos de control frente a la muerte celular o la proliferación controlada. Se trata de enfermedades multi-génicas donde una serie de mutaciones en genes clave se encargan de dirigir el proceso de desarrollo tumoral.
En este contexto, la biología de sistemas aporta el marco idóneo para el estudio del cáncer ya que permite describir las alteraciones en la función biológica como cambios en el sistema, más allá de los elementos individuales afectados. De especial relevancia lo constituye el estudio de las rutas de señalización ya que se implican en funciones esenciales como el ciclo celular.
Este trabajo se centra en el estudio de las rutas de señalización y de sus alteraciones en un conjunto amplio de cánceres recopilados a través de dos consorcios internacionales (TCGA e ICGC).
La metodología de análisis empleada descompone el conjunto de rutas en cascadas de señalización encargadas de llevar a cabo funciones específicas. De especial relevancia dentro del trabajo es determinar qué cambios se producen en la comunicación entre cascadas con funciones relacionadas a consecuencia de la enfermedad, así como de otros parámetros más específicos como el subtipo tumoral, el estadio o la probabilidad de supervivencia.

Palabras clave (inglés): Computational biology, Systems biology, Signaling pathways, Crosstalk, Cancer.
Palabras clave (español): Biología computacional, Biología de sistemas, Rutas de señalización, Diafonía, Cáncer.

Autor del TFG: Alumno: D. Pau Nebot Forcada.
Localidad y fecha: Valencia, septiembre de 2016.
Tutor Académico: Prof. D. Jose Javier Forment Millet.
Cotutor: D. Joaquín Dopazo Blázquez.
Cotutor colaborador: D. José Carbonell Caballero.

Dedicatorias y/o agradecimientos:

Debo dar las gracias en primer lugar a Javier Forment, cuyas clases de Python me introdujeron en la bioinformática y su guía me permitió progresar en este campo.

En segundo lugar agradezco a Joaquín Dopazo y José Carbonell por permitirme pasar un año en el departamento de Genómica de Sistemas del CIPF y por su apoyo en la realización de este trabajo.

En tercer lugar agradezco a todos los compañeros del laboratorio la excelente acogida e integración en el grupo. Hicisteis que cada día de trabajo fuera para mí una experiencia muy llevadera, divertida e instructiva. En especial debo mencionar a Carol, Edgar y Yunlong por su incondicional ayuda y amistad.

Por último, debo dar las gracias a mis padres y a mi novia por su apoyo y sobre todo por su paciencia.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La complejidad de los sistemas biológicos requiere una correcta gestión de la información del entorno, desde la generación del mensaje hasta su consecuencia en el organismo vivo (Mc Mahon et al., 2014). Ésta información es transmitida por diferentes cascadas de señalización, que a su vez están compuestas por diferentes caminos o subrutas (Hou et al., 2013).
La señalización ocurre mediante varios mecanismos como interacciones proteína-proteína, modificaciones post-translacionales a proteínas, actividad de enzimas o la organización y/o localización de los componentes en el interior de la célula (Landry et al., 2015).

A medida que avanzan los estudios y aumenta el conocimiento que tiene la comunidad científica sobre este campo se desvela la complejidad del control al que se encuentran sometidos los seres vivos a todos los niveles. El efecto de una señal es modulado de manera compleja por otras moléculas presentes por lo que se ve afectado por el contexto espacial y temporal en el que ocurre (Mc Mahon et al., 2014).

Por lo tanto, los diferentes caminos de transducción de señal transmitirán la información de receptores presentes tanto dentro como fuera de la célula al núcleo, donde se iniciará la respuesta celular adecuada, principalmente por factores de transcripción (Fossett, 2013).
Gracias a esta transmisión de estímulos, las células reciben información del exterior, reaccionan ante cambios y se adaptan para mantener la homeostasis. Lógicamente, desde un punto de vista evolutivo, se favorece la optimización de esta transmisión de señales (Mc Mahon et al., 2014). Se cree que esta información, en concreto para algunas señales de carácter esencial, es transmitida por un número pequeño de caminos conservados entre diferentes sistemas celulares. Además, la especificidad de la señal se ve reforzada por interacciones entre diferentes cascadas de señalización (Fossett, 2013).

La capacidad de entender los procesos moleculares de transmisión de información está relacionada directamente con la capacidad de cuantificar los componentes responsables de estos (Mc Mahon et al., 2014). Así mismo, poco a poco se está cambiando el modo de clasificar las diferentes rutas de señalización, pasando de emplear receptores de membrana como identificativo de la ruta, (por ejemplo en el caso de la proteína G), a nombrar la ruta en sí (Hou et al., 2013).

Considerar las redes de rutas de señalización en lugar de obviarlas, en estudios de respuesta a enfermedades o fármacos, puede facilitar la tarea de encontrar nuevas dianas de origen genómico o transcriptómico. Dicho sea de paso, la comprensión de esta red no es tarea fácil dada la variedad de maneras con las que se pueden transmitir estas señales y la complejidad de sus interacciones (Landry et al., 2015). Como se concretará en los apartados siguientes del trabajo, esta comunicación o interacción entre caminos de rutas de señalización, (frecuentemente nombrada usando el término inglés "crosstalk"), en muchos casos no ha sido comprendida o estudiada en su totalidad. Cabe mencionar que esto probablemente es consecuencia del enfoque reduccionista empleado en el campo a lo largo de la historia.

En cuanto a qué método (ómico) elegir para el estudio de la señalización, ninguno es adecuado en solitario por la multitud de formas en las que la señal puede ser codificada, sino que lo ideal es combinar varios de ellos en modelizaciones, cosa que en un futuro puede estar a la altura de

la naturaleza dinámica, reprogramable y plástica que caracteriza las cascadas de señalización (Landry et al., 2015).

Los estudios realizados en este trabajo se realizaron a escala celular. Las mediciones tomadas, aunque directamente pudieran ser correlacionadas en busca de relaciones lineares, pueden resultar más informativas al combinarlas con estudios de transmisión de la señal o modelizaciones, permitiendo la reconstrucción y comprensión de toda la red (Mc Mahon et al., 2014).

La modelización de una cascada de señalización generalmente da como resultado datos de interacción entre componentes en forma de red. Estos datos contienen esencialmente elementos, (principalmente proteínas), constituyendo los nodos de la red y sus interacciones, estas últimas representadas como aristas o "edges". Además, estas interacciones no tienen que ser físicas o directas, sino que también admiten otras de naturaleza indirecta (Landry et al., 2015).

A continuación se presentarán brevemente algunas de las estrategias seguidas en modelización: suelen estar basadas en datos, en agrupaciones (en inglés "clustering"), regresiones multivariable y otras técnicas de reducción de la dimensionalidad (Landry et al., 2015).

Entrando en detalle, las estrategias basadas en datos combinadas con información ya disponible sobre las interacciones que se estén estudiando pueden ayudar a entender la red de señalización, las influencias, conexiones entre los componentes y en general el flujo de información. Estos análisis asumen que las relaciones entre proteínas presentan patrones estadísticamente observables en los datos (Pe'er y Hacohen, 2011).

Como se concreta en los siguientes apartados, el software utilizado en este trabajo (HiPathia) utiliza una estrategia de este tipo para transformar los datos iniciales de expresión en valores de activación de caminos.

Por otro lado, entre las estrategias basadas en agrupaciones se encuentran el agrupamiento jerárquico, K-medias, K-vecinos..., y se emplean para identificar subgrupos de muestras con algún tipo de relación, por ejemplo individuos del mismo subtipo.
Por último, las regresiones y otras técnicas de reducción de la dimensionalidad incluyen el análisis de componentes principales (PCA), regresión de mínimos cuadrados parciales (PLSR) y análisis discriminante lineal (LDA) entre otros.

Tras presentar las cascadas de señalización y la modelización, se ampliará en los siguientes párrafos el concepto de comunicación entre rutas de señalización y se comentarán distintos abordajes seguidos por varios grupos de investigación para obtener resultados coherentes y fiables.

Se han planteado diferentes definiciones para la interacción entre cascadas de señalización (o también conocido por el término inglés "crosstalk") en función de lo que se entiende por "interacción" en cada estudio. Por ejemplo, mientras que unos grupos consideran que hay comunicación entre rutas de señalización si comparten simplemente un nodo (proteína), otros exigen que esta interacción ocurra con una mayor frecuencia que la esperada aleatoriamente (Tegge et al., 2016). Con el tiempo han aparecido grupos de investigación con interpretaciones del concepto de "interacción" de mayor complejidad, al mismo tiempo que se avanza en la comprensión de éstas y se proponen modelos más aproximados a la realidad.

A continuación se presentarán métodos de estudio de las interacciones entre cascadas de señalización empleados en trabajos recientes:

- Relaciona dos rutas de señalización si entre ellas hay más conexiones de las esperadas en una red aleatoria (Li et al., 2008).
- Observa las proteínas de dos rutas de señalización y relaciona ambas rutas si éstas presentan un número mayor de conexiones entre sí que el esperado al azar (Dotan-Cohen et al., 2009).
- Empleo de las correlaciones de coexpresión entre los subsets de genes de una cascada de señalización para inferir las redes de comunicación entre caminos (Wang et al., 2013).
- Considera que dos rutas de señalización tienen interacción si comparten proteínas (Donato et al., 2013).
- Conexión de una pareja de rutas si tienen componentes en común (aparte de proteínas, por ejemplo genes) (Han et al., 2015). Emplean la base de datos KEGG.
- Asocia dos cascadas de señalización si una estimulación de los receptores de una de ellas causan una respuesta en la otra, (normalmente llevada a cabo por factores de transcripción). Este acercamiento considera los caminos más cortos ya que supone que así la señal viaja rápido, y la posibilidad de que esta comunicación ocurra de manera asimétrica (Tegge et al., 2016).

Otra manera de clasificar el tipo de estudio es mediante las categorías de "crosstalk" o interacción estática o dinámica. Las estrategias que emplean simplemente los datos de cascadas de señalización, (nodos e interacciones entre elementos de una misma cascada, por ejemplo con la información de la base de datos KEGG), son las conocidas como "crosstalk estático" ya que los resultados que encuentren son independientes de las fluctuaciones presentes al incluir datos reales. Por el contrario, el "crosstalk dinámico" emplea datos reales generados normalmente por tecnologías ómicas, opcionalmente combinados con otros datos estáticos para dar medidas de comunicación entre caminos específica del contexto de los datos.

Tras introducir estos conceptos, se presenta la definición de interacción entre rutas de señalización empleada en este trabajo y en el grupo receptor:
Se dice que ocurre una interacción entre dos cascadas de señalización A y B si una señal que entra en la red por un "nodo de inicio" de A se transmite y causa un efecto en un "nodo final o efector" en B. Se entiende por "nodo de inicio" cualquier proteína o componente celular que pueda iniciar una cadena de comunicación que se transmitirá a los siguientes nodos de la red (normalmente un receptor de membrana), y por "nodo final" el elemento directamente responsable de llevar a cabo una respuesta celular y que recibe la señal en último lugar (proteína efectora).

Habiendo comentado la señalización y el concepto de interacción entre rutas, queda introducir la diana de este trabajo: el cáncer.

El diverso conjunto de enfermedades comúnmente llamadas cáncer se originan cuando una población celular sufre modificaciones y evoluciona progresivamente a un estado neoplásico (Hanahan y Weinberg, 2011).
El conocimiento sobre cáncer al inicio del siglo XXI ya reconocía los cambios en el genoma como elemento importante en la evolución de la enfermedad. Se conocía que ocurren mutaciones productoras de oncogenes con una ganancia de función por un lado, y por otro se pierde la función de genes supresores de tumores (Bishop y Weinberg, 1996). La complejidad de los tumores es mayor que la de un conjunto de células cancerígenas proliferando, se trata de tejidos compuestos por muchos tipos celulares diferentes con interacciones entre sí y con las células normales de los alrededores. Estas últimas actúan como matriz o estroma y son responsables de algunas de las capacidades del tumor (Hanahan y Weinberg, 2011).

En un intento de facilitar la comprensión de la biología del cáncer, Hanahan y Weinberg (2000) definieron seis capacidades distintivas y complementarias que permiten el crecimiento y dispersión metastática de los tumores. (Esta definición también se conoce como "Hallmarks of Cancer"). Éstas son:

- Resistir la muerte celular
- Mantener la señalización proliferativa
- Evadir los supresores del crecimiento
- Activar los mecanismos de invasión y metástasis
- Alcanzar la inmortalidad replicativa
- Inducción de la angiogénesis

Sin embargo, los descubrimientos de la siguiente década aunque respetan las capacidades distintivas anteriores presentan algunas nuevas, así como características que predisponen o permiten a células normales el paso a malignas.

Las capacidades distintivas (o "hallmarks") incluidos recientemente son la reprogramación del metabolismo energético (para que sea capaz de lidiar con la actividad exacerbada de las células cancerígenas), y la evasión de la destrucción por parte del sistema inmune.
Por otro lado, y relacionado con todo lo anterior, se pueden definir dos características que son causa y a la vez consecuencia de la transición celular a un estado neoplásico: la inestabilidad genómica y las mutaciones asociadas a ésta, y la inflamación promotora de tumores.
Estas dos últimas características se favorecen entre sí, y a su vez generan el contexto para que aparezcan el resto de capacidades distintivas del cáncer (Hanahan y Weinberg, 2011).

Como diana de este trabajo se decidió estudiar células de distintos tipos de cáncer porque la naturaleza neoplásica lleva asociados importantes cambios en la señalización y gestión metabólica, cosa que favorece la existencia de cambios en la comunicación entre rutas de señalización susceptibles de ser encontrados. Las interacciones encontradas podrían dar explicación a fenotipos resistentes, aparición de tolerancias o posibles reprogramaciones, que a su vez mejoren la manera de entender y tratar estas patologías.

Se han encontrado ejemplos previos a la realización de este trabajo en el que se pone de manifiesto la importancia de la comunicación entre rutas de señalización con respecto a las capacidades de las células y tejidos cancerígenos:

- Las rutas de señalización del ciclo celular y de p53 presentan actividad aumentada e interaccionan. Las alteraciones de estas rutas son importantes para muchos tipos de tumores (por ejemplo el cáncer de hígado) (Ito et al., 2000; Zhan et al., 1993).
- En cáncer de pecho, la comunicación entre IGF1R y EGFR aumenta el potencial metastático (van der Veeken et al., 2009; Riedemann et al., 2007).
- Se encuentran interacciones en 11 rutas de señalización entre sí en cáncer de próstata primario y 7 en cáncer de próstata metastático (Wang et al., 2012).
- Interacción de los genes JUN y FOS con ETS2, formando un heterotrímero que participa en la invasión y metástasis (Wang et al., 2016).

Por el motivo anterior (generación de grandes cambios en la señalización celular), relevancia del cáncer en la sociedad, disponibilidad de datos iniciales de calidad (origen: ICGC/TCGA), y compatibilidad con el software empleado, se ha decidido trabajar con muestras de diferentes tipos de cáncer como objeto de estudio en este trabajo. La información de origen son datos de expresión y también se dispone de los estadios en los que se encuentran las muestras en algunos de los cánceres, por lo que también son tratados en el trabajo.

Concluyendo todo lo anterior, para llevar a cabo el estudio de la señalización y poder integrar datos obtenidos por técnicas NGS teniendo en cuenta la definición de interacción presentada, se decidió emplear el software HiPathia, que combina la estructura conocida de cascadas de señalización con datos de expresión para obtener valores de activación de cada uno de los caminos (explicado en más detalle en el apartado de materiales y métodos). A diferencia de algunas modelizaciones introducidas en los anteriores párrafos, los resultados obtenidos mediante este acercamiento son de naturaleza dinámica ya que varían con los valores de expresión empleados.
Sabiendo que estos datos son específicos del tipo celular estudiado, la estrategia planteada para analizar las interacciones entre caminos consiste en una combinación de modelización, cálculo de correlaciones y filtrado. Se ha supuesto que una pareja de caminos tienen alguna interacción si sus valores de activación (obtenidos con la modelización) presentan un alto índice de correlación.

La metodología de este trabajo espera obtener resultados fiables y establecerse como una alternativa robusta para el estudio de las interacciones entre rutas de señalización, ampliando el control frente a otras estrategias mediante el empleo de datos reales y un filtrado estricto. Los datos obtenidos podrán complementar otras investigaciones y con suerte mejorar la comprensión de algunas enfermedades neoplásicas.

OBJETIVOS

- Planteamiento y validación de una metodología para obtener cambios significativos de interacciones entre rutas de señalización teniendo en cuenta datos de diferentes tipos celulares.

- Obtener listados de cambios en interacciones entre caminos para un conjunto de tipos de cáncer.

- Conseguir resultados de interacciones asociados a estadios de estos tipos de cáncer.

- Analizar los datos encontrados en conjunto para llegar a conclusiones a nivel de pan-cáncer.

MATERIALES Y MÉTODOS

Los análisis realizados en este trabajo se han llevado a cabo utilizando scripts escritos en R con el apoyo del software RStudio: R versión 3.2.1 (18/06/2015), RStudio versión 0.99.447 (2015). El actual lenguaje R es el resultado de un esfuerzo colaborativo que fue iniciado a mediados de 1997 por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland (Ihaka y Gentleman, 1996).

La estrategia empleada para estudiar las interacciones entre caminos presentes en diferentes tipos de cáncer consistió en observar las diferencias de las correlaciones encontradas en muestras normales frente a tumores y realizar un filtrado, obteniendo aquellas parejas de caminos cuyas correlaciones sufrían un cambio significativo. Estas interacciones deben dar respuestas con mayor probabilidad a las alteraciones mesurables en dichas células.

ANÁLISIS DE PAN-CÁNCER

1. Preparación de los datos y cálculo de la actividad de las rutas.

Los datos de partida de este proyecto vienen del TCGA, actualmente con información de más de 11000 pacientes pertenecientes a 33 tipos de cáncer, (y a su vez estadios), originados gracias a la colaboración de 20 instituciones. Se eligieron las muestras del TCGA por la gran cantidad de datos disponibles, (cosa que facilita el análisis), y por el interés que hay en el estudio de enfermedades neoplásicas (Tomczak et al., 2015).

Los datos de expresión empleados provienen de la "release 19" y fueron descargados directamente de los servidores del ICGC por los miembros del laboratorio receptor mediante FTP. De todos los datos descargados solo se usaron los del TCGA por la cantidad de tipos de cáncer que contiene y la alta fiabilidad de las variantes clínicas (estadios).
Aparte de esos datos, se descargó, (en "https://dcc.icgc.org/releases/release_19/Summary"), un fichero con información para cada muestra (sample.all_projects.tsv), indicando principalmente los diferentes identificadores asignados a cada una de ellas, relacionándolos con el proyecto y espécimen de origen , (se usó para clasificar las muestras). Por último, fue empleado un fichero con información clínica relativa a los individuos, generado por el grupo receptor a partir de los datos del TCGA tras ser procesados y adaptados para que resulte cómodo trabajar con ellos en R. Este último fichero se empleó posteriormente en los estudios de estadios.

A continuación, se utilizó la herramienta HiPathia (http://hipathia.babelomics.org/), para pasar los datos de expresión génica, (obtenidos mediante RNAseq), a datos de actividad de cascada de señalización para cada individuo. El software realiza el análisis de rutas de señalización asociadas a los siguientes componentes:

Tabla 1. Rutas de señalización analizadas por el software HiPathia. (Provenientes de los identificadores en inglés empleados por el software).

Ras	Rap1	MAPK
ErbB	Wnt	Notch
Hedgehog	TGF-beta	Hippo
VEGF	Jak-STAT	NF-kappa B
TNF	HIF-1	FoxO
Calcium	Sphingolipid	cAMP
cGMP-PKG	PI3K-Akt	AMPK
mTOR	Cell cycle	Oocyte meiosis
Apoptosis	p53	Focal adhesion
Adherens junction	Tight junction	Gap junction
Platelet activation	Toll-like receptor	NOD-like receptor

RIG-I-like receptor	Natural killer cell mediated cytotoxicity	T cell receptor
B cell receptor	Fc epsilon RI	Fc gamma R-mediated phagocytosis
Leukocyte transendothelial migration	Chemokine	Insulin
Glucagon	Adipocytokine	PPAR
GnRH	Estrogen	Progesterone-mediated oocyte maturation
Oxytocin	Thyroid hormone	Melanogenesis
Adrenergic signaling in cardiomyocytes	Vascular smooth muscle contraction	Neurotrophin
Pathways in cancer	Choline metabolism in cancer	Transcriptional misregulation in cancer
Proteoglycans in cancer	Gastric acid secretion	Hepatitis C

HiPathia analiza las rutas de señalización anteriores descomponiéndolas en diferentes caminos entre nodos de inicio o entrada (receptores de membrana, intracelulares...) y nodos finales o efectores.

Por ejemplo, la ruta de señalización siguiente se descompondría en 2 caminos efectores, uno por cada nodo final (proteína efectora). Cada uno de los caminos se traza desde todos los nodos de entrada (A, B y C) que llegan a cada nodo final en cada caso (D o E): (**A**/**B**/**C**→X→X→**D**, **A**/**B**/**C**→X→**E**).

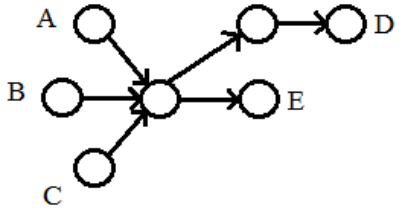


Figura 1. Esquema de una ruta de señalización sencilla.

A cada nodo de inicio se le asignó un valor, y para cada muestra se transmitió como si se tratase de la intensidad de una señal a lo largo del camino, siendo modulado teniendo en cuenta los datos de expresión. Como resultado se obtuvo una matriz que enfrentaba todos los caminos de todas las rutas de señalización con todas las muestras, y que contiene los respectivos valores de actividad.

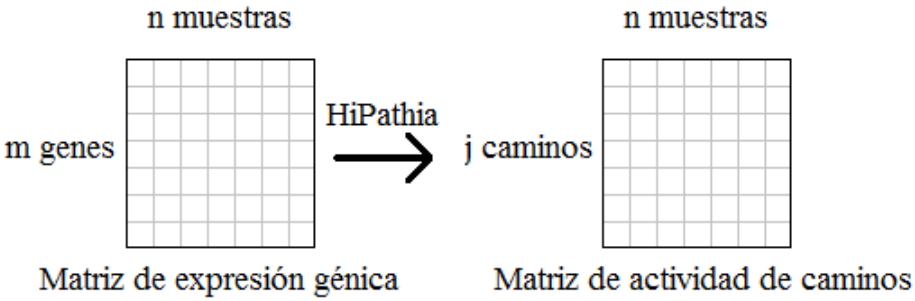


Figura 2. Esquema de las matrices iniciales; antes y después de emplear el software HiPathia.
Se introduce una matriz de datos de expresión con un número n de muestras que contiene los valores de expresión para cada uno de los m genes en cada muestra, y se obtiene otra matriz con las mismas muestras en una de las dimensiones de la matriz y con cada uno de los caminos que el software es capaz de analizar en la otra. El número de muestras y genes (n y m) vienen definidos por los datos de partida, mientras que el número de caminos depende de las rutas que HiPathia es capaz de analizar.

Posteriormente, se extrajeron los identificadores asociados a muestras pertenecientes al proyecto TCGA, presentes en el fichero con los identificadores de muestra (sample.all_projects.tsv), y fueron clasificados por tipo de cáncer y tipo de muestra (normal/tumor). A continuación, se emplearon estos identificadores para dividir la matriz de actividad de caminos en otras más pequeñas con los datos para cada tipo de muestra de cada tipo de cáncer.
En detalle, el posterior análisis se realizó solo con los identificadores presentes en la matriz de caminos que también habían sido clasificados por tipo de cáncer y tipo de muestra. Por supuesto, en este punto fueron eliminados los tipos de cáncer que no contaban con muestras en la matriz de actividad de caminos.

2. Estimación del grado de comunicación entre cascadas de señalización.

Para valorar las interacciones entre rutas de señalización, se realizaron correlaciones con el método de Pearson de los valores de actividad entre cada par de caminos, para las muestras normales y tumores de cada tipo de cáncer (cada tipo de muestra por separado). La fórmula empleada fue la siguiente:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Ecuación 1.

Donde cov es la covarianza, σ_X es la desviación estándar de X, μ_X es la media de X y E es la esperanza.

Profundizando en los pasos seguidos, se extrajeron de la matriz de actividad de caminos todos aquellos pertenecientes a las rutas de señalización: "Hepatitis C", "Pathways in cancer", "Choline metabolism in cancer", "Transcriptional misregulation in cancer" y "Proteoglycans in cancer". Esto se debe a que esas rutas no aportan información coherente con un estudio de cáncer (hepatitis), o que tratan de partes muy específicas de la señalización, útiles en fenotipos concretos y que es mejor obviar para estudiar el comportamiento general ("... in cancer").
Además, a nivel de script se eliminaron las correlaciones de caminos con todos los valores iguales (la desviación estándar daba 0, y la falta de resultados causaba problemas en análisis posteriores).

A continuación se procedió a aplicar un estadístico que puntúe la diferencia entre dos correlaciones (sobre las correlaciones anteriores, que se almacenaron en forma de matrices para cada tipo de muestra y tipo de cáncer). Se empleó la siguiente fórmula (Fisher, 1921):

$$z = \frac{\left(\frac{1}{2}\ln\frac{1+r_1}{1-r_1}\right) - \left(\frac{1}{2}\ln\frac{1+r_2}{1-r_2}\right)}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}$$

Ecuación 2.

Donde r_1 y r_2 son los coeficientes de correlación para una misma posición en ambas matrices de correlación (normal y tumor), y n_1 y n_2 son el número de muestras (normales y tumores) con las que se obtuvieron las matrices de correlación respectivamente.

El estadístico se aplicó para todos los tipos de cáncer con un tamaño muestral mayor que 3, tanto para normales como tumores, ya que en caso contrario no es posible calcular el test de comparación, (además, sería una población de estudio demasiado pequeña).

De la aplicación de este test estadístico para cada par de caminos se recuperó una matriz con el primer conjunto de resultados (sin procesar). Las parejas de caminos sin resultado (NaN) o con

resultado infinito fueron descartadas porque son consecuencia de diferencias entre correlaciones de caminos pertenecientes a una misma ruta de señalización (entre ellos).

El análisis a continuación, hasta obtener el segundo conjunto de resultados (o resultados significativos finales), se realizó de manera paralela para todos los tipos de cáncer con un número de muestras correcto. Estos tipos de cáncer son los que se encuentran en la tabla 2 y se nombrarán en el trabajo con los identificadores del TCGA.

Tabla 2. Listado de los tipos de cáncer estudiados en este trabajo junto con los identificadores empleados.

BLCA	Cáncer de células de transición de pelvis renal y de uréter
BRCA	Cáncer de mama invasivo
COAD	Adenocarcinoma de colon
HNSC	Carcinoma de células escamosas de cabeza y cuello
KIRC	Carcinoma de células claras del riñón
KIRP	Carcinoma de células renales papilares
LIHC	Carcinoma hepatocelular
LUAD	Adenocarcinoma pulmonar
LUSC	Carcinoma pulmonar de células escamosas
PRAD	Adenocarcinoma de próstata
READ	Adenocarcinoma de recto
THCA	Cáncer tiroideo
UCEC	Cáncer endometrial

A partir de este punto, ya se disponía de resultados iniciales en forma de diferencias entre correlaciones de dos grupos (normales contra tumores), pero éstos debían ser recorridos para recuperar los significativos y eliminar un gran número de resultados no relevantes.

Para ello, se procedió a la obtención de un p-valor para saber si el valor del estadístico (que puntúa la diferencia) es significativo, aplicando la siguiente función a la matriz de resultados del test estadístico:

$$f(X) = 2 * FDA(-|X|)$$

Ecuación 3. Donde FDA es la función de distribución acumulada de una distribución normal (estándar en este caso).

El empleo de esta función, siendo X la matriz con los de resultados del test estadístico, devolvió otra matriz de igual tamaño con el p-valor correspondiente a cada pareja de caminos.

Sobre esta matriz con los p-valores se aplicó una corrección por test múltiple del p-valor con el método de Benjamini y Hochberg (False discovery rate, FDR)(Benjamini y Hochberg, 1995).
Se evaluaron los p-valores corregidos y se estableció un umbral de 0.05, eliminando las parejas de caminos cuya puntuación no se encontrase bajo este límite.

3. Filtrado de los resultados

A continuación se aplicaron tres filtros para centrar el estudio en aquellas parejas cuyas diferencias entre correlaciones resultan realmente informativas de acuerdo con el enfoque del trabajo.

En primer lugar, se evaluó para cada pareja de caminos si presentaban un valor por encima de 0.95 en ambas matrices de correlación (normal y tumor), eliminando los resultados que cumplieron esta condición. Esto se realizó porque la metodología está basada en la comparación de correlaciones, y sobre 0.95 ambas correlaciones son significativas, por lo que al final serían

desechadas debido a que no aportan el resultado diferencial considerado interesante en este trabajo. Además, observaciones posteriores de estos resultados revelaron que se tratan en algunos casos de caminos de una misma cascada de señalización entre ellos, que simplemente presentan valores altos en el test estadístico por tener los valores de correlación cercanos a 1, siendo no relevantes para el estudio.

A nivel de script, las parejas de caminos que superaron los controles hasta el filtro anterior (incluido) son el primer conjunto de resultados filtrados. Para cada uno de ellos fueron recopilados sus correspondientes valores de correlación (de normales y tumores), el resultado del test estadístico que evalúa la diferencia entre las correlaciones y los identificadores asociados. En detalle, fueron eliminadas las entradas que presentaban un valor de 0 exacto para una de las dos correlaciones (ya que el resultado sería erróneo por haber sido introducido ese cero en uno de los pasos anteriores).

El conjunto de datos obtenido ya eran resultados destacables. Sin embargo, para enriquecerlos en interacciones relevantes se aplicaron dos filtros más: se exigió una diferencia entre las matrices de correlación de al menos 0.5 y que cada resultado tuviera al menos una de las matrices de correlación (de normales o tumores) con un valor absoluto superior a 0.65. Las parejas que pasen esos filtros presentaron principalmente ganancias, pérdidas o inversiones de correlaciones y constituyeron el segundo conjunto de resultados.
Hay que destacar que en este punto acaba el filtrado, y que cada una de las parejas de caminos cuya diferencia entre correlaciones es significativa y que supera estos filtros es considerada como uno de los resultados significativos finales, independientemente de su reorganización o revisión posterior para llegar a otras conclusiones.

4. Obtención de matrices para el análisis de pan-cáncer

Posteriormente, como parte del análisis de pan-cáncer, se procedió a combinar en una misma tabla los resultados de todos los tipos de cáncer estudiados. Para ello, el script almacenó todas las parejas de caminos que pasaron los filtros de manera que cada pareja aparecía solo una vez (diferencia simétrica). Después, se creó una tabla en la que había cada uno de estos resultados significativos en una fila distinta, y cada columna contenía en forma de vector lógico (unos y ceros) la información sobre cuáles de ellos han superado los filtros para cada tipo de cáncer.

Aparte de esta tabla se creó otra conteniendo los valores del test estadístico para cada caso en lugar de vectores lógicos. Se empleó para observar si se mantenía la capacidad de agrupar tipos de cáncer similares y comprobar la coherencia de los resultados obtenidos.

Después se calculó, a partir de la anterior tabla, una nueva matriz lógica para los distintos tipos de cáncer calculando el p-valor para cada posición con la última función (Ecuación 3), y considerando que pasan el filtro los valores por debajo de 0.05. El objetivo de esto fue observar el número de resultados que habrían pasado los filtros si estos hubieran sido menos estrictos y si era posible agrupar tipos de cáncer relacionados.

Los análisis siguientes se realizaron por duplicado, partiendo de las tablas: lógica (de resultados por tipo de cáncer), y de valores del test estadístico filtrada por p-valor (presentadas en los tres últimos párrafos). De las matrices lógicas se puede extraer información para cualquiera de los tipos de cáncer que contiene.

En los siguientes apartados de la memoria son comentados los resultados encontrados en la primera matriz lógica generada, mientras que la de "valores del test filtrados por p-valor" es empleada simplemente para comprobar la robustez de la metodología y los cambios en los resultados al ser menos estrictos con el filtrado.

Como estudio de todos los datos se procedió a contar las veces que cada pareja de caminos había superado los filtros, y por tanto mostraba una diferencia significativa en sus correlaciones para los distintos tipos de cáncer, y se ordenaron de manera descendente. Estos resultados fueron analizados en busca de interacciones entre caminos que se mantuvieran entre tipos de cáncer.

Por último, se generaron nuevas tablas para clasificar como ha sido el cambio de correlación, a partir de la de resultados significativos finales frente a tipos de cáncer (una para cada tipo de cáncer). En cada una se encontraba, para cada pareja que ha pasado los filtros: los identificadores y los nombres comprensibles de las rutas, los valores de las dos matrices de correlación (de normales y tumores), la diferencia entre ambos y el tipo de cambio que ha ocurrido.
Para aclarar la explicación, un ejemplo de determinar el "tipo de cambio de correlación" sería el caso en el que dos caminos presentan un valor de correlación en muestras normales cercano a 0 y en torno a 0.8 en tumores, razón por la cual el cambio se consideraría una ganancia de correlación positiva.

A nivel de script, este último elemento (tipo de cambio) se obtuvo a partir del procesado de las matrices de correlación que originaron esa entrada en los resultados. Se asumió sin correlación (0) todos aquellos valores de correlación que se encuentren entre 0.25 y -0.25. El resto se consideró una correlación directa (1) o inversa (-1), (por ejemplo, el paso de 0.86 a 0.18 aparecería como de 1 a 0). Tras simplificar el cambio de correlación a este sistema, los pocos que aparecían como "1 a 1" o "-1 a -1" fueron forzados a ir de "-1" o "1" a "0", considerando 0 la correlación cuyo valor absoluto sea la menor de las dos. Por último se sustituyeron estos códigos de cambio por etiquetas esquemáticas de lo que ocurrió (6 diferentes, referido siempre a correlaciones al pasar de normal a tumor): ganancia y pérdida positivas, ganancia y pérdida de negativas, y cambio de positivo a negativo o viceversa.

La figura 3 contiene un esquema de la metodología para facilitar su comprensión. La parte de la izquierda de la figura (flechas de color negro y gris) contiene los métodos expuestos en los apartados previos de la sección de materiales y métodos, mientras que la parte de la derecha (flechas azules) contiene brevemente los pasos que son explicados a continuación de la figura.

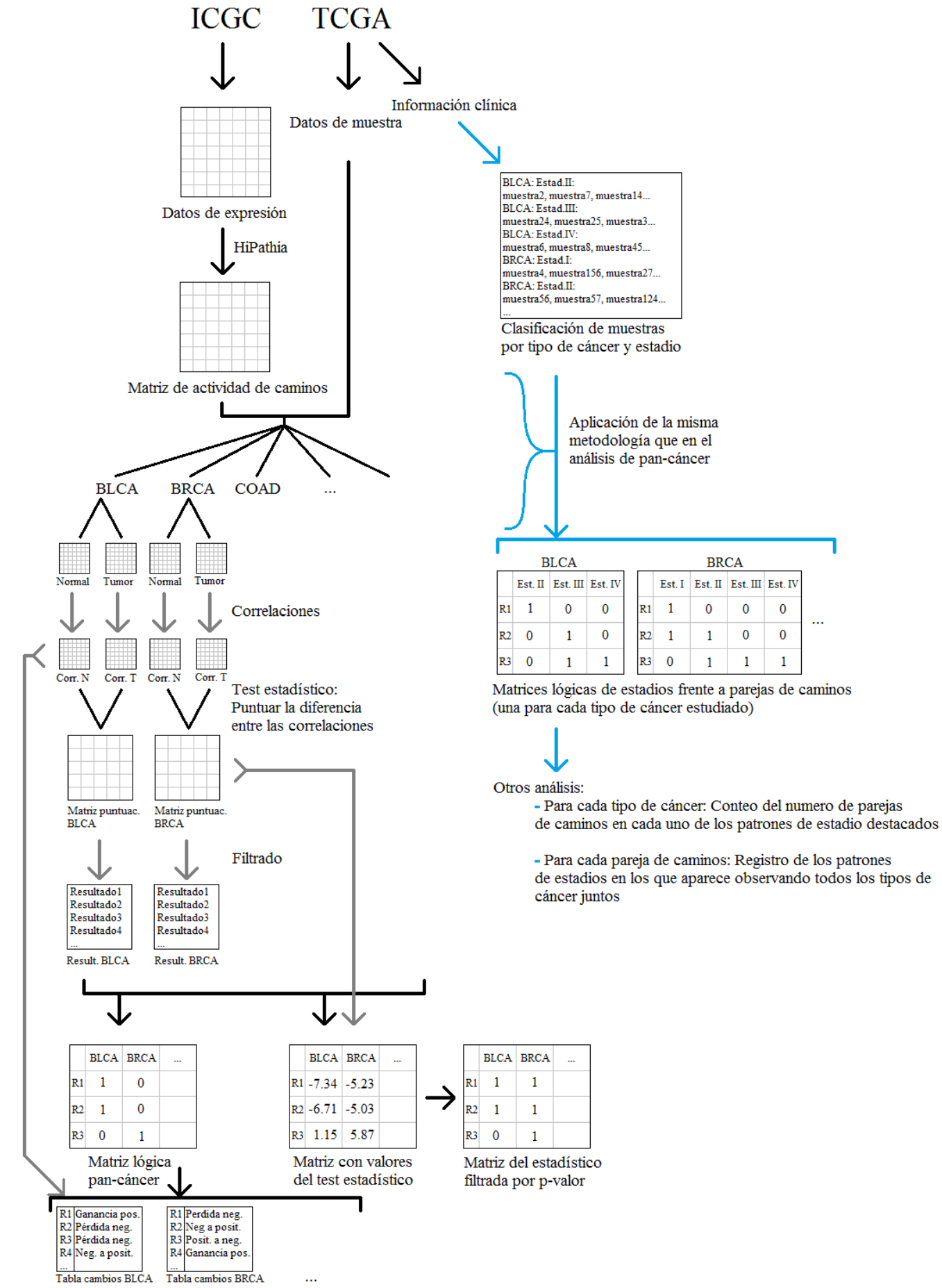


Figura 3. Esquema de la metodología.

ESTUDIO DE LAS INTERACCIONES ENTRE CAMINOS EN DIFERENTES ESTADIOS

Se tomaron los tipos de cáncer que superaron las exigencias del análisis anterior, (13 tipos, disponibles en la tabla 2), y fueron extraídos del fichero con información clínica los nombres de las distintas muestras de tumores, almacenándolas bajo su tipo de cáncer y estadio.
En detalle, el script eliminó, para cada tipo de cáncer, estadios no relevantes como "[Not Available]", "[Discrepancy]", "[Not Applicable]" o "Stage X". Después, se compararon las muestras de cada estadio con las muestras con datos para ese tipo de cáncer, para asegurar que se puede continuar el análisis. Los estadios de cáncer sin muestras asociadas y tipos de cáncer sin estadios fueron eliminados del análisis.

A continuación se simplificaron los estadios, ya que en algunos casos se dieron subestadios y su análisis queda fuera de los objetivos de este trabajo. Un número variable de estadios y subestadios complicaría la comparación de diferentes tipos de cáncer y la posterior búsqueda de coherencia o referencias en la bibliografía que confirmen lo encontrado. Por esto, se forzaron todos los estadios y subestadios a formar parte de un sistema de 3 a 4 estadios, véase: "I", "II", "III" y "IV".
Por la misma razón que en el análisis de pan-cáncer, se eliminaron los estadios que no tuvieran más de tres muestras, (por no poder calcular el test de comparación).

Una vez dispuestos los datos, se aplicó la misma metodología que en el análisis de pan-cáncer, iniciando el proceso con la eliminación de los caminos pertenecientes a las rutas de señalización de los tipos: "... in cáncer" y "Hepatitis C", y continuando con el procesamiento de los estadios de cada tipo de cáncer por separado. En cada caso, se compararon las muestras de cáncer de un estadio concreto contra las muestras normales asignadas a ese tipo de cáncer. Además, para valorar el efecto del ajuste del p-valor y el tamaño muestral en la obtención de resultados todo el análisis se repitió con y sin el ajuste.

Al final del proceso, se crearon matrices de estadios frente a parejas de caminos de varios tipos: lógico (con vectores de 1 y 0 para cada estadio indicando las parejas que han superado los filtros), la matriz de valores del test estadístico de cada posición y la matriz lógica recuperada de la anterior (de manera análoga al análisis de pan-cáncer).

De la primera y última matriz presentadas en el último párrafo fue posible obtener qué parejas de rutas presentan diferencias de correlación entre las muestras normales y las de un estadio o estadios en concreto. Gracias a esto, se procedió a procesar las dos matrices y contar el número de resultados que se daban en patrones de estadios relevantes, para dar una visión global de la distribución de las correlaciones. Sin embargo, en este trabajo solo se ha adjuntado y comentado la matriz lógica, (la primera de las anteriores), por considerarse los resultados deseados, reservándose la otra para futuros análisis.

En detalle, se contaron los resultados que se presentaban en uno o varios estadios siguiendo uno de estos patrones: Que se den solo en un estadio, que se inicien en el estadio 1 (o más bajo disponible) y se mantengan hasta uno de los estadios intermedios, o que se inicien en uno de los estadios intermedios y se mantengan hasta el estadio 4 (o más alto disponible). Los resultados que siguieron cualquiera de los otros patrones se descartaron por no ser relevantes y/o tratarse probablemente de falsos positivos. Las parejas de caminos comunes a todos los estadios de un tipo de cáncer se contaron aparte.

A partir de estos datos se generaron figuras que comparan entre distintos tipos de cáncer el número de resultados comunes a todos los estadios, y otras que contabilizan el número de resultados que se encuentran en cada uno de los patrones destacados presentados en el párrafo anterior.

Por otro lado, se realizó un análisis para observar si las interacciones entre caminos son específicas de estadios concretos en varios tipos de cáncer. Para ello se tomaron las matrices de estadios frente a parejas de caminos (solo las presentadas en párrafos anteriores como "tipo lógico") y se combinaron en una nueva tabla de resultados en la que cada fila corresponde a una pareja de caminos y las columnas a cada uno de los patrones de estadios relevantes anteriores (que la interacción ocurra en solo un estadio y los casos en los que se da desde el primer o último estadio hasta uno de los intermedios)

De esta tabla se puede extraer información como el número tipos de cáncer en los que se encuentra cada pareja y cuantas veces ocurre cada patrón de estadio en cada uno de los resultados, siendo objeto de estudio en los siguientes bloques del trabajo.

ESTUDIO DEL EFECTO DEL TAMAÑO DE MUESTRA Y DE LA CORRECCIÓN POR TEST MÚLTIPLE

La metodología empleada en este análisis fue la misma que en el estudio de pan-cáncer, con la diferencia de haberse incorporado todos los pasos en una función que los aplica solo sobre un tipo de cáncer y un determinado número de muestras elegibles en los argumentos. El fin de esta parte del análisis fue generar conjuntos de muestras aleatorios que en principio no deberían dar ninguna diferencia de correlación significativa. Con esto se midió variación en el número de falsos positivos generados al modificar el tamaño de muestra y aplicar una corrección por test múltiple.

A nivel de script, al emplear esta función, se introdujo el tipo de cáncer y tipo de muestra (normal o tumor) a emplear, además de dos números que eran la cantidad de muestras que debían ser elegidas aleatoriamente para comparar. Dicho de otra manera, en lugar de comparar para un tipo de cáncer concreto todas las muestras normales frente a las tumores, comparó un número n1 de muestras frente a n2 muestras, siendo ambos conjuntos elegidos aleatoriamente de un mismo tipo de cáncer y de muestra (los dos grupos contendrán solo muestras normales o tumores).

El proceso a continuación se inició comprobando la elección correcta de los tamaños de muestras a comparar (que el tipo de cáncer tuviera muestras suficientes de ese tipo y que fueran más de 3), seguido de un muestreo aleatorio de la cantidad y tipo elegido.

El siguiente paso fue obtener matrices de valores de actividad de caminos frente a los conjuntos de muestras aleatorias provenientes de la matriz con todas las muestras del análisis de pan-cáncer. El resto de análisis fueron exactamente igual que el de pan-cáncer excepto que el resultado que devolvió todo el proceso no fue una lista de resultados, sino cuantas parejas habían superado todos los filtros.

El análisis se concluyó repitiendo esta función para tipos de cáncer y muestra concretos, variando el tamaño de los dos conjuntos de muestras aleatorias comparadas y almacenando la media de la cantidad de resultados obtenida tras 30 réplicas.
Se compararon tamaños de muestras iguales de 5 a 50 en intervalos de 5, dejando un número de muestras fijo mientras se varió el otro (50 frente a una cantidad de 5 a 50 en incrementos de 5 muestras, todas aleatorias) y también todas las combinaciones de número de muestras de 5 a 50 en intervalos de 5 (incluyendo comparaciones de tamaños iguales). Toda la metodología se aplicó con y sin el ajuste de p-valor.

Por último, se generaron diagramas de cajas para facilitar la observación del número de falsos positivos en los distintos conjuntos comparados. Se incluyeron líneas en el gráfico correspondientes a umbrales de 0.05 y 0.01 del total de resultados.

RESULTADOS Y DISCUSIÓN

Análisis de pan-cáncer

Los primeros resultados a analizar son los que se obtuvieron tras completar los primeros tres apartados descritos en la parte de análisis de pan-cáncer, parte de los materiales y métodos. Se trata de una tabla de resultados para cada tipo de cáncer en la que cada una de las filas corresponde a una pareja de caminos cuya diferencia entre correlaciones ha superado todos los filtros.

A continuación se muestra con un diagrama de barras el número de resultados obtenido, (diferencias significativas entre caminos al comparar correlaciones de muestras normales y tumores), para los 13 tipos de cáncer estudiados:

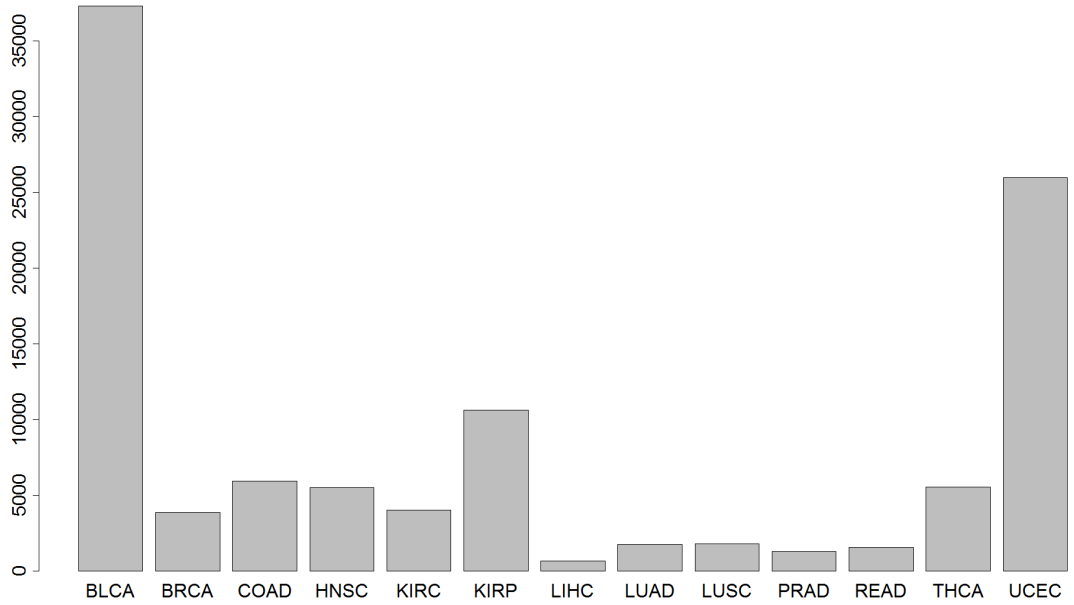


Figura 4. Diagrama de barras del numero de resultados frente a tipo de cáncer.

Para facilitar la observación de diagrama se incluye la siguiente tabla con los recuentos de cada tipo de cáncer empleados en la creación de la figura.

Tabla 3. Recuento de resultados por tipo de cáncer.

Tipo:	BLCA	BRCA	COAD	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	THCA	UCEC
Nº Result.	37314	3882	5943	5510	4030	10648	684	1789	1812	1284	1586	5552	25972

Debido al gran tamaño del listado de resultados, no es posible adjuntarlo al presente trabajo, pero para su posterior consulta y estudio el lugar donde se encuentran depositados se indica en el anexo 3.

El número de parejas de caminos que han superado los filtros es significativamente mayor en los tipos de cáncer BLCA, UCEC y KIRP, siendo en los dos primeros mucho más grande que en el resto. Este hecho no parece estar inversamente relacionado con el tamaño de muestra (cosa que favorecería la aparición de correlaciones espurias), ya que el tipo READ cuenta con un número muy bajo de muestras tipo normal y la cantidad de resultados encontrados es inferior a 2000.

Para su comparación, la tabla siguiente, (tabla 4), contiene el recuento de muestras de cada tipo de cáncer empleadas para obtener los resultados:

Tabla 4. Recuento de muestras de los distintos tipos de cáncer estudiados.

Tipo:	BLCA	BRCA	COAD	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	THCA	UCEC
Normal	17	113	41	42	72	32	48	55	45	52	9	58	23
Tumor	301	1057	451	480	526	222	294	486	428	379	153	500	516

Continuando la observación del tamaño muestral, el siguiente párrafo contiene resultados y discusión de los análisis descritos en la sección de materiales y métodos con el título: "Estudio del efecto del tamaño de muestra y de la corrección por test múltiple". En conjunto se realizaron para comprobar la robustez de la metodología.

Como era de esperar, tamaños de muestra menores daban un mayor número de falsos positivos si no se aplica ninguna corrección, dependiendo siempre del tamaño del conjunto de muestras menor entre los dos comparados. Este número de resultados erróneos siempre se encontraba por debajo del 5% del total, pero se consideró que dado el volumen de tests realizados era necesario aplicar una corrección.
Tras incluir una corrección por test múltiple del p-valor el número de falsos positivos se mantenía constante entre los diferentes tamaños de muestra y siempre muy por debajo del 0.5% del total. Los diagramas de cajas del número de falsos positivos encontrados en las diferentes condiciones se pueden encontrar en el anexo 5.

Desde un punto de vista biológico, el distinto número de resultados significativos encontrados (cambios en las interacciones entre caminos) puede deberse a que el grado de contribución de estas interacciones en las alteraciones varía entre los tipos de cáncer estudiados. Sin embargo, se debe destacar que el número de interacciones encontradas es relativamente bajo, considerando que el número de interacciones potenciales es superior a $8 \cdot 10^5$ (810000).

El gran número de resultados significativos dificulta la interpretación de las interacciones encontradas. Sin embargo, para comprobar la coherencia de los resultados se observarán brevemente algunos de los más significativos de varios tipos de cáncer.

Por ejemplo, observando uno de los resultados con mayor puntuación en el estadístico para el tipo KIRC se encuentra la pareja: HMOX1 con PFKFB3. El primero se encuentra altamente expresado en varios tumores sólidos, desempeñando un importante rol en el crecimiento rápido del tumor (Yim et al., 2011), mientras que el segundo es un enzima que en cáncer favorece la supervivencia de las células estimulando la glicólisis, media en el control circadiano de la carcinogénesis (las alteraciones en el reloj circadiano favorecen el crecimiento de las células cancerígenas) (Chen et al., 2016), e indirectamente favorece la angiogénesis (Xu et al., 2014). Dada su relación con el cáncer, a primera vista este resultado es coherente.

Elegido entre los resultados con mayor valor en el test estadístico, en el tipo LUSC se encuentra la pareja: RAC1 y CBLC. El primer gen está asociado entre otras cosas al movimiento de las células cancerígenas (Yang et al., 2012), mientras que el segundo participa en la reparación del DNA y está considerado un proto-oncogén. De nuevo es comprensible una interacción entre ambos. Además, el resto de los diez resultados con mayor puntuación de este tipo de cáncer se encuentran relacionados con las rutas de uniones estrechas y desmosomas en banda (motilidad y disgregación de los tumores), Ras, PI3K-Akt, ErbB, HIF-1, Apoptosis..., siendo todas estas rutas cuya señalización es muy relevante en cáncer.

Comentando uno más de los resultados, se observaron los de COAD de mayor puntuación, entre los que se encuentran genes de las rutas: Hedgehog, p53, fagocitosis mediada por Fc gamma R, Hippo, adhesión focal... El más significativo consiste en la interacción entre: PTCH1 y RPRM. El primero se encuentra inversamente relacionado con el potencial metastático de este tipo de cáncer (You et al., 2010), mientras que el segundo es un supresor de tumores, silenciado en algunos tipos de cáncer (Xu et al., 2012). De nuevo parece coherente que ambos caminos formen parte de este resultado y que haya superado los filtros.

A continuación, se crearon las tablas de resultados que aglutinan todos los tipos de cáncer para encontrar los que se mantienen entre ellos. En la sección de materiales y métodos se explica que se creó inicialmente una tabla lógica, sobre esta otra con los valores del test estadístico y a su vez otra matriz lógica simplemente calculando el p-valor y filtrando esta última. Los análisis realizados en el trabajo emplearán en cada caso los datos lógicos obtenidos directamente a partir de los resultados (la primera "tabla lógica" creada). Las posteriores conversiones a valores del test y valores lógicos se reservan para observar la distribución que siguen los resultados, como bien se especificó en la sección de materiales y métodos.

De esta tabla de resultados lógica, que contiene información sobre qué parejas presentan diferencias de correlación que superen los filtros se obtuvo el siguiente recuento:

Tabla 5. Recuento del número de resultados bajo la cantidad de tipos de cáncer en los que aparecen.

Tipos de cáncer	1	2	3	4	5	6
Número de resultados	70431	13362	2366	354	59	7

Aunque el nivel de exigencia del filtrado sea alto se han encontrado siete parejas de caminos que han superado los filtros en seis de los 13 tipos de cáncer estudiados. Estas parejas de caminos son las consideradas resultados finales de esta sección del trabajo.

Seguidamente, se han extraído cinco de las parejas de caminos presentes entre los resultados significativos del mayor número de tipos de cáncer y se han analizado de manera sistemática para comprobar la coherencia de dichas interacciones.
Las descripciones básicas de los genes o proteínas de cada pareja se han obtenido de las bases de datos: GeneCards (http://www.genecards.org), UniProt (http://www.uniprot.org) y Entrez Gene (http://www.ncbi.nlm.nih.gov/gene/).

Resultado 1: ERBB3 y TJP1 / CGN.

El resultado aparece en 6 de los 13 tipos de cáncer estudiados (BLCA, BRCA, HNSC, LUSC, READ y UCEC). En todos ocurre una pérdida de correlación positiva al pasar de células de tipo normal a tumor.
El gen ERBB3 codifica un receptor tipo tirosina kinasa de la familia de receptores del factor de crecimiento epidérmico. Por otro lado, TJP1 y CGN codifican proteínas relacionadas con la transmisión de señales en las uniones estrechas, (zonula occludens), y con la formación y regulación de la barrera de permeabilidad en el mismo lugar, respectivamente.
Se encuentran entradas en la literatura que respaldan este resultado, como relacionar la pérdida de uniones estrechas con la desdiferenciación de los tumores (Weinstein et al., 1976; Hoover et al., 1998), ocurriendo en los tipos de cáncer encontrados. En Bujko et al. (2015), se estudian los niveles de expresión en cáncer colorrectal, observando una disminución significativa en los niveles de CGN.
Por otro lado, en Ma et al. (2014), se defiende la naturaleza pro-oncogénica de ERBB3, incentivando la progresión tumoral en varias enfermedades y cuyos mecanismos subyacentes lo sitúan como una causa importante de los fallos en el tratamiento del cáncer principalmente por la activación de las rutas de señalización PI3K-Akt, MEK/MAPK y Jak/Stat, además de la kinasa Src. Una explicación de este resultado incluye a las ultimas rutas de señalización

mencionadas, ya que, como se comenta en Li et al. (2009), y Sheth et al. (2003), la ruta PI3K-Akt tiene la capacidad de controlar la integridad de las uniones estrechas.

Resultado 2: ACTB y GAB2.

Este resultado aparece en 6 de los 13 tipos de cáncer (BLCA, BRCA, KIRC, KIRP, LUAD y LUSC). En los 4 primeros tipos indicados se produce una pérdida de correlación positiva, mientras que en los dos de pulmón se produce una pérdida de correlación negativa.
El gen ACTB codifica una de las seis actinas, proteínas conservadas que participan principalmente en motilidad celular e integridad estructural. CAB2 pertenece a la familia de genes GAB, encargados de transmitir señales de respuesta a estímulos mediante citocinas y receptores de factores de crecimiento, así como receptores de antígenos de células B y T.
En Guo et al. (2013), se defiende la asociación de ACTB con tumores, jugando un rol en la patogénesis del cáncer. En cáncer renal se encuentran niveles de proteína alterados, superiores a los de las células normales. Por otro lado, aunque en pulmón se encuentran niveles superiores de expresión en células normales frente a tumores, análisis en serie de la expresión génica han revelado una diferencia entre estos niveles (en pulmón) y los del resto de tipos de cáncer (Chari et al., 2010). Puede que esto último explique el distinto tipo de cambio de correlación encontrado.
GAB2 también realiza un papel esencial en las rutas PI3K-Akt y ERK (ambas importantes en cáncer), entre otras cosas alterando la integridad celular y provocando una pérdida de adhesión y favorece el crecimiento independiente de anclaje (Adams et al., 2012).

Resultado 3: CCND1 y MYC.

Aparece en 6 de los 13 tipos de cáncer (BLCA, KIRC, KIRP, LUAD, LUSC, THCA), y en todos los casos se produce una ganancia de correlación positiva.
CCND1 codifica una proteína de la familia de las ciclinas, en concreto formando un complejo y actuando como subunidad reguladora de CDK4 o CDK6, cuya actividad es necesaria para la transición G1/S del ciclo celular.
MYC codifica una proteína involucrada en la progresión en el ciclo celular, apoptosis y transformaciones celulares. Actúa como factor de transcripción regulando genes específicos.
Existe literatura sobre ambos genes recalcando su relación con las neoplasias: CCND1 actúa como oncogén en diferentes neoplasias cuando se encuentra sobreexpresado (Moreno-Bueno et al., 2003; Musgrove et al., 2011), y MYC es un proto-oncogén reconocido, sobreexpresado en un gran porcentaje de los tumores (Evan y Littlewood, 1993; Hoffman y Liebermann, 2008).

La idea de que ambos genes presenten interacción en un contexto neoplásico es coherente dado el conocimiento que ya se tiene de ellos. Además ambos se encuentran sobreexpresados en cáncer, por lo que coincide con el tipo de cambio de correlación encontrado (ganancia positiva). Por último, se han encontrado artículos que indican que entre ambos genes se dan efectos sinergéticos (Nakagawa et al., 2011), y también que son susceptibles de verse afectados por reordenamientos cromosómicos similares (Roix et al., 2003).

Resultado 4: SPRY1 y CDC42.

Se encuentra en 6 de los 13 tipos de cáncer (BLCA, COAD, HNSC, READ, THCA, UCEC), y se produce en todos una pérdida de correlación negativa excepto en READ, que cambia de correlación negativa a positiva. La proteína codificada por SPRY1 es un regulador de la señalización del receptor tirosina kinasa, relacionado con el crecimiento, diferenciación y tumorigénesis, mientras que CDC42 codifica una GTPasa de la subfamilia Rho que regula rutas de señalización que afectan a la morfología celular, endocitosis, migración y progresión en el ciclo celular.

En la literatura se encuentra que CDC42 juega un papel clave promoviendo la metástasis (Reymond et al., 2012), mientras que SPRY1 parece actuar como un supresor de tumores (Masoumi-Moghaddam et al., 2014).
El comportamiento distinto en READ encontrado puede curiosamente explicarse porque en este tipo de cáncer se encuentra anormalmente sobreexpresado (Zhang et al., 2016; Holgren et al., 2010), y junto a los posiblemente también elevados niveles de CDC42 (Sahai y Marshall, 2002), pueden dar una respuesta a ese cambio del tipo de correlación de negativa a positiva. En el resto de cánceres de los resultados la correlación negativa se pierde probablemente porque los niveles de SPRY1 se mantienen bajos.

Resultado 5: GADD45G y RCHY1.

Este último resultado comentado se encuentra también en 6 de los 13 tipos de cáncer (BLCA, COAD, HNSC, LIHC, LUSC, UCEC), produciéndose en todos una ganancia de correlación positiva al pasar de tejido normal a tumor.
GADD45G participa en la respuesta a estrés ambiental activando la ruta p38/JNK mediante MTK1/MEKK4 y RCHY1 tiene actividad ubiquitina ligasa, regulando niveles de proteínas importantes y por extensión también el ciclo celular.
Se relaciona una falta de GADD45G con la progresión de enfermedades de tipo neoplásico por su capacidad de frenar el avance de la enfermedad (Tamura et al., 2012; Ju et al., 2009), y se esperan alteraciones del correcto funcionamiento de RCHY1 para permitir la presencia de cantidades aberrantes de proteínas en las células tumorales.
No se han encontrado referencias que relacionen ambos genes pero al menos el comportamiento de ambos parece similar en todos los tipos de cáncer, probablemente presentando niveles de expresión (y por lo tanto control de la inestabilidad celular) inferiores.

Como conclusión de esta exposición de resultados, hay que recalcar la dificultad de la comprobación de todos los resultados encontrados por su gran número. Sin embargo, el análisis de cinco de los resultados significativos más presentes entre los distintos tipos de cáncer plantea interacciones coherentes con la literatura y que respaldan la robustez de la metodología.

Llegado a este punto y siguiendo el orden expuesto en la sección de materiales y métodos cabe mencionar las otras tablas de resultados comunes entre tipos de cáncer generadas. Un heatmap de los valores de la tabla de resultados lógica no es capaz de detectar y agrupar tipos de cáncer parecidos, por lo que se pensó que la metodología había simplificado demasiado los resultados.
Posteriormente, al generar las tablas de valores del test y lógica a partir de ésta se recuperaron parte de las parejas de caminos, y el heatmap de éstos sí conseguía agrupar los tipos de cáncer similares. Por lo anterior se supuso que la metodología era correcta y la falta de capacidad de relacionar tipos de cáncer en un heatmap era debida al nivel de exigencia del filtrado, (que sumado a la naturaleza lógica de los datos empleados dificulta la correcta agrupación).

Para su consulta, los heatmaps de ambas tablas lógicas (antes y después de recuperar parte de los resultados) se encuentran en el anexo 1.

Para concluir esta parte de la metodología, se tomaron las parejas resultantes y se generaron tablas para observar el tipo de cambio de correlación. A partir de estas, se ha creado otra con el recuento de los cambios para todos los tipos de cáncer estudiados.

Tabla 6. Recuento del número de resultados en cada tipo de cambio para cada tipo de cáncer.

Tipo:	BLCA	BRCA	COAD	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	THCA	UCEC
Pérdida neg.	18590	1798	2560	2552	1726	4132	249	924	812	384	534	2344	11040
Pérdida pos.	17365	1883	2861	2757	1976	5961	354	843	789	564	855	3101	14352
Ganancia neg.	33	12	28	25	46	34	8	4	44	139	0	16	32
Ganancia pos.	253	149	164	130	148	108	58	13	156	166	2	39	208
Neg. a posit.	674	30	241	22	83	282	10	1	6	18	86	37	208
Posit. a neg.	399	10	89	24	51	131	5	4	5	13	109	15	132

Se observa que la gran mayoría de cambios en las interacciones entre caminos suponen la pérdida de estas correlaciones al pasar de normal a tumor. Se puede argumentar que esto es lógico por el gran número de mutaciones incontroladas que se dan en las células neoplásicas, por lo que muy probablemente la mayoría de cambios en el genoma alteren negativamente la expresión de genes no relacionados con el cáncer. Así mismo, las pocas interacciones que se ganan probablemente serán interesantes, ya que estarán asociadas a genes que necesita el cáncer para su desarrollo. Un ejemplo de estas ganancias de interacciones asociadas a oncogenes y pérdidas asociadas a supresión de tumores se puede observar en el resultado 3 de los comentados previamente (interacción entre CCND1 y MYC). Estos dos genes están sobreexpresados en cáncer y de manera coherente aparecen como ganancia de correlación en el estudio de cambios en las correlaciones.

Estudio de las interacciones entre caminos en diferentes estadios.

De la metodología introducida en la sección de materiales y métodos con el mismo título que este apartado, se obtuvieron para cada tipo de cáncer una tabla lógica con información referente a los estadios en los que está presente cada una de las parejas de caminos de los resultados.

Estos resultados de dividieron en dos: las parejas que se encontraban en todos los estadios y las que no lo hacen. El diagrama del anexo 2 contiene el recuento de resultados comunes a todos los estadios encontrados en los distintos tipos de cáncer.
Por otro lado, el resto de resultados fueron cribados como se explica en los métodos y se realizaron recuentos del número de resultados (disponibles en el anexo 4).

Se debe destacar que un gran número de resultados significativos (cambios en las correlaciones entre muestras tipo normal y tumor) encontrados aparecen en todos los estadios de sus respectivos tipos de cáncer. La distribución de resultados entre estadios en los distintos tipos de cáncer no sigue ningún patrón concreto, encontrándose picos en cualquiera de los estadios (aunque más frecuentemente estadios 2 y 4), y también variando la exclusividad de los resultados por estadios concretos. Sin embargo, a grandes rasgos se puede decir que el número de resultados encontrados aumenta al avanzar en los estadios en todos los tipos de cáncer estudiados excepto en KIRC.

CONCLUSIONES

La combinación de datos provenientes de un número creciente de tecnologías ómicas con modelizaciones se presenta como una alternativa muy prometedora y que sin duda será empleada por los distintos grupos de investigación en el futuro próximo. La creciente capacidad de gestión de datos por parte de la bioinformática puede llegar a dejar atrás el clásico enfoque

reduccionista y encontrar respuestas más complejas, o que en el caso contrario serían descartadas al hacer la investigación sobre una parte del conjunto.

El análisis realizado permite la detección de alteraciones coherentes con la literatura y permite la detección de interacciones que de otra manera no serían observadas. El enfoque empleado, que combina modelización, datos reales, correlaciones y filtrado puede ser el origen de un nuevo marcador abstracto que va mucho más allá de la alteración (detectar el comportamiento e interacciones entre componentes con niveles de expresión aberrantes).

El acercamiento realizado, clasificable como de biología de sistemas, aporta resultados significativos coherentes y en grandes rasgos permite ver que la expresión diferencial entre cánceres es muy variable, así como encontrar respuestas a nivel de pan-cáncer.

Las líneas de investigación futuras deberán continuar el análisis del grandísimo número de resultados obtenidos para ver cómo afectan éstas interacciones a los procesos celulares. También deberán mejorar la metodología, modificando los parámetros de filtrado, progresivamente ir incluyendo más datos de partida y con el tiempo emplear información de otros tipos (como por ejemplo tener en cuenta datos de supervivencia).
Por último, se deberá seguir trabajando en el estudio de los resultados de pan-cáncer y estadios.

REFERENCIAS BIBLIOGRÁFICAS

ADAMS, S.J.; AYDIN, I.T.; CELEBI, J.T. (2012). GAB2--a scaffolding protein in cancer. Mol Cancer Res. 10(10):1265-1270.

BENJAMINI, Y.; HOCHBERG, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Stat Methodol. 57(1):289-300.

BISHOP, J.M.; WEINBERG, R.A. (1996). Molecular Oncology. New York: Scientific American, Inc.

BUJKO, M.; KOBER, P.; MIKULA, M.; LIGAJ, M.; OSTROWSKI, J.; SIEDLECKI, J.A. (2015). Expression changes of cell-cell adhesion-related genes in colorectal tumors. Oncol Lett. 9(6):2463-2470.

CHARI, R.; LONERGAN, K.M.; PIKOR, L.A.; COE, B.P.; ZHU, C.Q.; CHAN, T.H.; MACAULAY, C.E.; TSAO, M.S.; LAM, S.; NG, R.T. et al., (2010). A sequence-based approach to identify reference genes for gene expression analysis. BMC Med Genomics. 3.

CHEN, L.; ZHAO, J.; TANG, Q.; LI, H.; ZHANG, C.; YU, R.; ZHAO, Y.; HUO, Y.; WU, C. (2016). PFKFB3 Control of Cancer Growth by Responding to Circadian Clock Outputs. Sci Rep. 6.

DONATO, M.; XU, Z.; TOMOIAGA, A.; GRANNEMAN, J.G.; MACKENZIE, R.G.; BAO, R.; THAN, N.G.; WESTFALL, P.H.; ROMERO, R.; DRAGHICI, S. (2013). Analysis and correction of crosstalk effects in pathway analysis. Genome Res. 23(11):1885-1893.

DOTAN-COHEN, D.; LETOVSKY, S.; MELKMAN, A.A.; KASIF, S. (2009). Biological process linkage networks. PLoS One. 4(4).

EVAN, G.I.; LITTLEWOOD, T.D. (1993). The role of c-myc in cell growth. Curr Opin Genet Dev. 3(1):44-49.

FISHER, R.A. (1921). On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron. 1:3-32.

FOSSETT, N. (2013). Signal transduction pathways, intrinsic regulators, and the control of cell fate choice. Biochim Biophys Acta. 1830(2):2375-2384.

GUO, C.; LIU, S.; WANG, J.; SUN, M.Z.; GREENAWAY, F.T. (2013). ACTB in cancer. Clin Chim Acta. 417():39-44.

HAN, J.; LI, C.; YANG, H.; XU, Y.; ZHANG, C.; MA, J.; SHI, X.; LIU, W.; SHANG, D.; YAO, Q. et al., (2015). A novel dysregulated pathway-identification analysis based on global influence of within-pathway effects and crosstalk between pathways J R Soc Interface. 12(102).

HANAHAN, D.; WEINBERG, R.A. (2000). The Hallmarks of Cancer. Cell. 100:57-70.

HANAHAN, D.; WEINBERG, R.A. (2011). Hallmarks of cancer: the next generation. Cell. 144(5):646-674.

HOFFMAN, B.; LIEBERMANN, D.A. (2008). Apoptotic signaling by c-MYC. Oncogene. 27(50):6462-6472.

HOLGREN, C.; DOUGHERTY, U.; EDWIN, F.; CERASI, D.; TAYLOR, I.; FICHERA, A.; JOSEPH, L.; BISSONNETTE, M.; KHARE, S. (2010). Sprouty-2 controls c-Met expression and metastatic potential of colon cancer cells: sprouty/c-Met upregulation in human colonic adenocarcinomas. Oncogene. 29(38):5241-5253.

HOOVER, K.B.; LIAO, S.Y.; BRYANT, P.J. (1998). Loss of the tight junction MAGUK ZO-1 in breast cancer: relationship to glandular differentiation and loss of heterozygosity. Am J Pathol. 153(6):1767-1773.

HOU, Y.; HOU, Y.; HE, S.; MA, C.; SUN, M.; HE, H.; GAO, N. (2014). The merged basins of signal transduction pathways in spatiotemporal cell biology. J Cell Physiol. 229(3):287-291.

IHAKA, R.; GENTLEMAN, R. (1996). R: A Language for Data Analysis and Graphics. J Comp Graph Stat. 5(3):299-314.

ITO, T.; SHIRAKI, K.; SUGIMOTO, K.; YAMANAKA, T.; FUJIKAWA, K.; ITO, M.; TAKASE, K.; MORIYAMA, M.; KAWANO, H.; HAYASHIDA, M. et al. (2000). Survivin promotes cell proliferation in human hepatocellular carcinoma. Hepatology. 31(5):1080-1085.

JU, S.; ZHU, Y.; LIU, L.; DAI, S.; LI, C.; CHEN, E.; HE, Y.; ZHANG, X.; LU, B. (2009). Gadd45b and Gadd45g are important for anti-tumor immune responses. Eur J Immunol. 39(11):3010-3018.

LANDRY, B.D.; CLARKE, D.C.; LEE, M.J. (2015). Studying Cellular Signal Transduction with OMIC Technologies. J Mol Biol. 427(21):3416-3440.

LI, Y.; AGARWAL, P.; RAJAGOPALAN, D. (2008). A global pathway crosstalk network. Bioinformatics. 24(12):1442-1447.

LI, N.; NEU, J. (2009). Glutamine deprivation alters intestinal tight junctions via a PI3-K/Akt mediated pathway in Caco-2 cells. J Nutr. 139(4):710-714.

MA, J.; LYU, H.; HUANG, J.; LIU, B. (2014). Targeting of erbB3 receptor to overcome resistance in cancer treatment. Mol Cancer. 13.

MASOUMI-MOGHADDAM, S.; AMINI, A.; MORRIS, D.L. (2014). The developing story of Sprouty and cancer. Cancer Metastasis Rev. 33(2-3):695-720.

MC MAHON, S.S.; SIM, A.; FILIPPI, S.; JOHNSON, R.; LIEPE, J.; SMITH, D.; STUMPF, M.P. (2014). Information theory and signal transduction systems: from molecular information processing to network inference. Semin Cell Dev Biol. 35:98-108.

MORENO-BUENO, G.; RODRÍGUEZ-PERALES, S.; SÁNCHEZ-ESTÉVEZ, C.; HARDISSON, D.; SARRIÓ, D.; PRAT, J.; CIGUDOSA, J.C.; MATIAS-GUIU, X.; PALACIOS, J. (2003). Cyclin D1 gene (CCND1) mutations in endometrial cancer. Oncogene. 22(38):6115-6118.

MUSGROVE, E.A.; CALDON, C.E.; BARRACLOUGH, J.; STONE, A.; SUTHERLAND, R.L. (2011). Cyclin D as a therapeutic target in cancer. Nat Rev Cancer. 11(8):558-572.

NAKAGAWA, M.; TSUZUKI, S.; HONMA, K.; TAGUCHI, O.; SETO, M. (2011). Synergistic effect of Bcl2, Myc and Ccnd1 transforms mouse primary B cells into malignant cells. Haematologica. 96(9):1318-1326

PE'ER, D.; HACOHEN, N. (2011). Principles and strategies for developing network models in cancer. Cell. 144(6):864-873.

REYMOND, N.; IM, J.H.; GARG, R.; VEGA, F.M.; BORDA D'AGUA, B.; RIOU, P.; COX, S.; VALDERRAMA, F.; MUSCHEL, R.J.; RIDLEY, A.J. (2012). Cdc42 promotes transendothelial migration of cancer cells through β1 integrin. J Cell Biol. 199(4):653-668.

RIEDEMANN, J.; TAKIGUCHI, M.; SOHAIL, M.; MACAULAY, V.M. (2007). The EGF receptor interacts with the type 1 IGF receptor and regulates its stability. Biochem Biophys Res Commun. 355(3):707-714.

ROIX, J.J.; MCQUEEN, P.G.; MUNSON, P.J.; PARADA, L.A.; MISTELI, T. (2003). Spatial proximity of translocation-prone gene loci in human lymphomas. Nat Genet. 34(3):287-291.

SAHAI, E.; MARSHALL, C.J. (2002). RHO-GTPases and cancer. Nat Rev Cancer. 2(2):133-142.

SHETH, P.; BASUROY, S.; LI, C.; NAREN, A.P.; RAO, R.K. (2003). Role of phosphatidylinositol 3-kinase in oxidative stress-induced disruption of tight junctions. J Biol Chem. 278(49):49239-49245.

TAMURA, R.E.; DE VASCONCELLOS, J.F.; SARKAR, D.; LIBERMANN, T.A.; FISHER, P.B.; ZERBINI, L.F. (2012). GADD45 proteins: central players in tumorigenesis. Curr Mol Med. 12(5):634-651.

TEGGE, A.N.; SHARP, N.; MURALI, T.M. (2016). Xtalk: a path-based approach for identifying crosstalk between signaling pathways. Bioinformatics. 32(2):242-251.

TOMCZAK, K.; CZERWIŃSKA, P.; WIZNEROWICZ, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 19(1A):68-77.

VAN DER VEEKEN, J.; OLIVEIRA, S.; SCHIFFELERS, R.M.; STORM, G.; VAN BERGEN EN HENEGOUWEN, P.M.; ROOVERS, R.C. (2009). Crosstalk between epidermal growth factor receptor- and insulin-like growth factor-1 receptor signaling: implications for cancer therapy. Curr Cancer Drug Targets. 9:748-760.

WANG, T.; GU, J.; YUAN, J.; TAO, R.; LI, Y.; LI, S. (2013). Inferring pathway crosstalk networks using gene set co-expression signatures. Mol Biosyst. 9(7):1822-1828.

WANG, J.M.; WU, J.T.; SUN, D.K.; ZHANG, P.; WANG, L. (2012). Pathway crosstalk analysis based on protein-protein network analysis in prostate cancer. Eur Rev Med Pharmacol Sci. 16:1235-1242.

WANG, L.; LI, J.; ZHAO, H.; HU, J.; PING, Y.; LI, F.; LAN, Y.; XU, C.; XIAO, Y.; LI, X. (2016). Identifying the crosstalk of dysfunctional pathways mediated by lncRNAs in breast cancer subtypes. Mol Biosyst. 12(3):711-720.

WEINSTEIN, R.S.; MERK, F.B.; ALROY, J. (1976). The structure and function of intercellular junctions in cancer. Adv Cancer Res. 23:23-89.

XU, Y.; AN, X.; GUO, X.; HABTETSION, T.G.; WANG, Y.; XU, X.; KANDALA, S.; LI, Q.; LI, H.; ZHANG, C. et al., (2014). Endothelial PFKFB3 plays a critical role in angiogenesis. Arterioscler Thromb Vasc Biol. 34(6):1231-1239.

XU, M.; KNOX, A.J.; MICHAELIS, K.A.; KISELJAK-VASSILIADES, K.; KLEINSCHMIDT -DEMASTERS, B.K.; LILLEHEI, K.O.; WIERMAN, M.E. (2012). Reprimo (RPRM) is a novel tumor suppressor in pituitary tumors and regulates survival, proliferation, and tumorigenicity. Endocrinology. 153(7):2963-2973.

YANG, W.H.; LAN, H.Y.; HUANG, C.H.; TAI, S.K.; TZENG, C.H.; KAO, S.Y.; WU, K.J.; HUNG, M.C.; YANG, M.H. (2012). RAC1 activation mediates Twist1-induced cancer cell migration. Nat Cell Biol. 14(4):366-374.

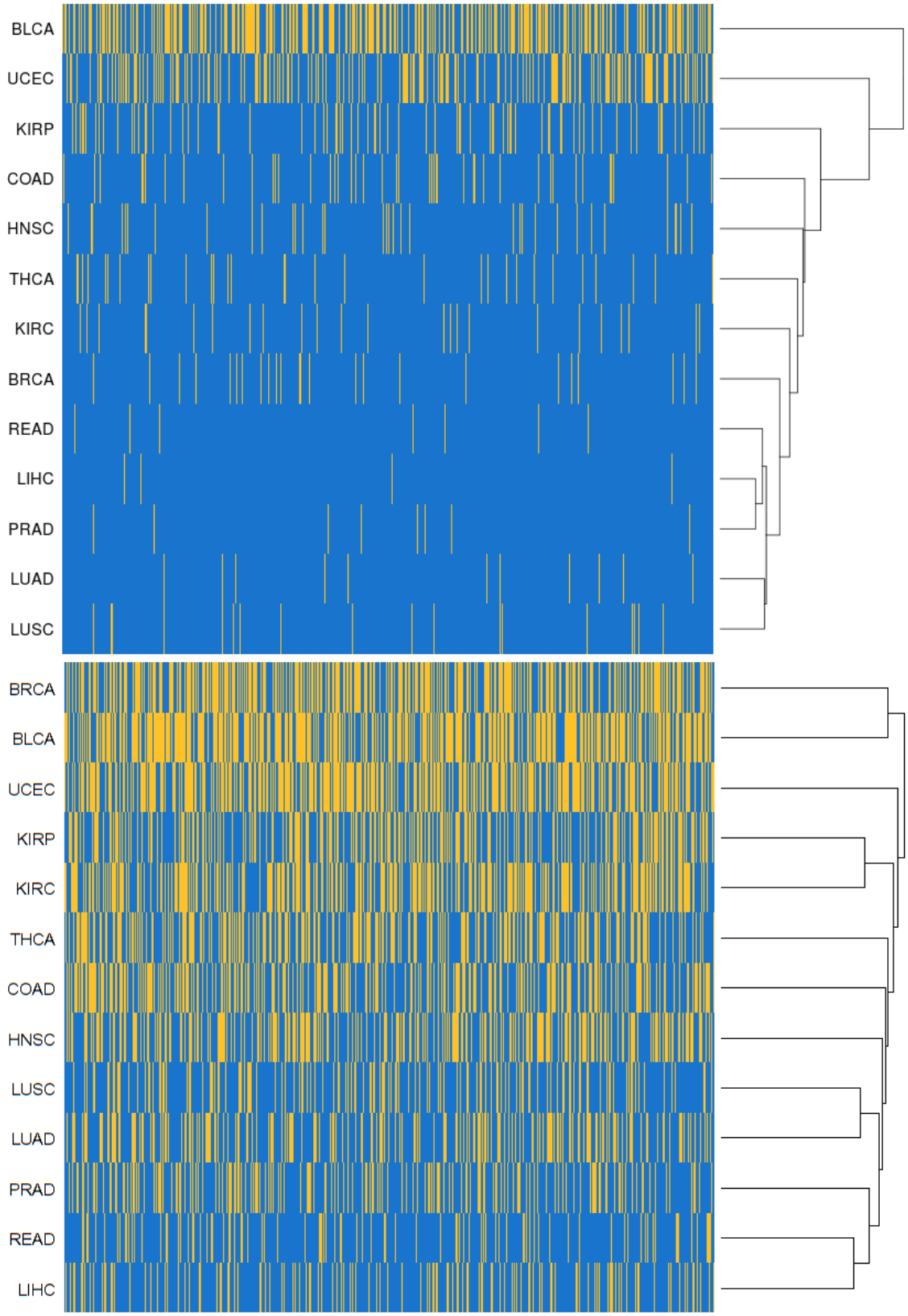
YIM, M.S.; HA, Y.S.; KIM, I.Y.; YUN, S.J.; CHOI, Y.H.; KIM, W.J. (2011). HMOX1 is an important prognostic indicator of nonmuscle invasive bladder cancer recurrence and progression. J Urol. 185(2):701-705.

YOU, S.; ZHOU, J.; CHEN, S.; ZHOU, P.; LV, J.; HAN, X.; SUN, Y. (2010). PTCH1, a receptor of Hedgehog signaling pathway, is correlated with metastatic potential of colorectal cancer. Ups J Med Sci. 115(3):169-175.

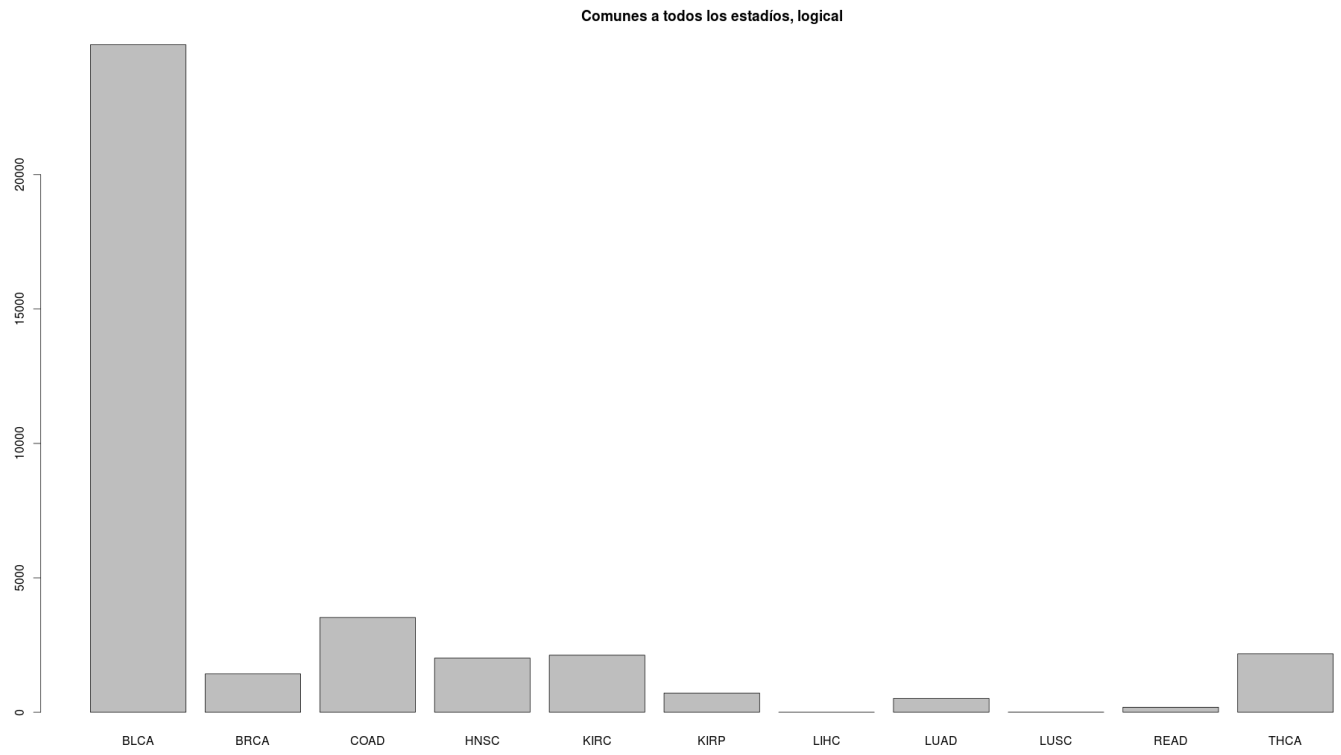
ZHAN, Q.; CARRIER, F.; FORNACE, A.J. (1993). Induction of cellular p53 activity by DNA-damaging agents and growth arrest. Mol Cell Biol. 13(7):4242–4250.

ZHANG, Q.; WEI, T.; SHIM, K.; WRIGHT, K.; XU, K.; PALKA-HAMBLIN, H.L.; JURKEVICH, A.; KHARE, S. (2016). Atypical role of sprouty in colorectal cancer: sprouty repression inhibits epithelial-mesenchymal transition. Oncogene. 35(24):3151-3162.

ANEXO 1. Heatmap de la tabla lógica empleada en el trabajo (arriba), y de la generada tras recuperar resultados con los valores del test estadístico y filtrarlos por p-valor (abajo).
En el heatmap superior no se observa ninguna agrupación, mientras que en el inferior aparecen juntos tipos de cáncer relacionados, (por ejemplo con el mismo origen), como KIRP/KIRC (riñón) o LUAD/LUSC (pulmón).



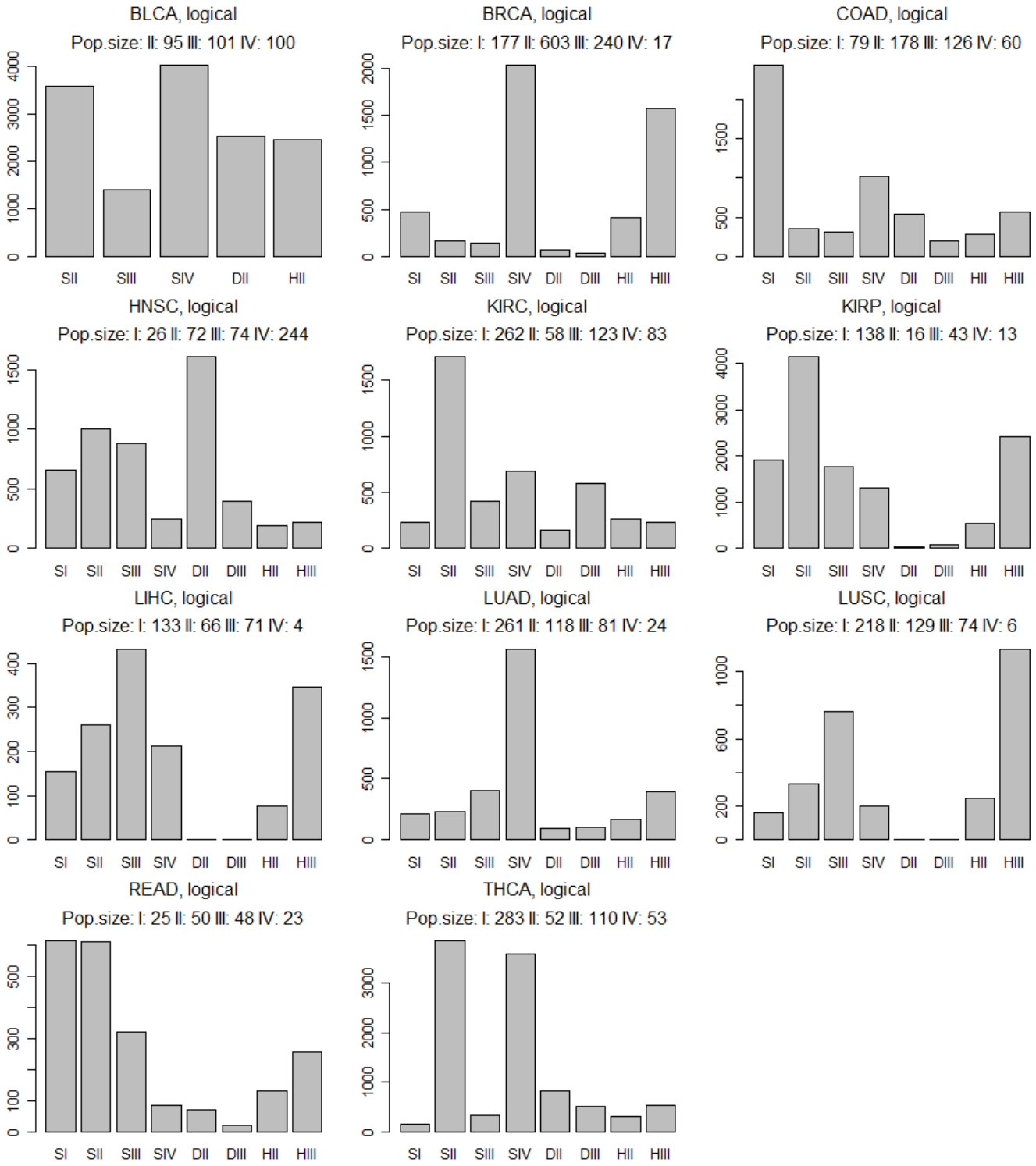
ANEXO 2. Diagrama de barras del recuento de resultados comunes a todos los estadios encontrados en los distintos tipos de cáncer.



ANEXO 3. Dirección web donde se depositarán las últimas versiones de los scripts y los resultados obtenidos (o el enlace a la base de datos donde se encuentren):

https://github.com/nebforpau/crosstalk_analysis

ANEXO 4. Diagramas de barras del recuento del número de resultados en cada uno de los patrones de estadio relevantes para cada tipo de cáncer. En cada uno aparece el tipo de cáncer de los datos de origen y "logical" por haberse generado a partir de las tablas de resultados significativos lógicas. Posteriormente, se incluye el número de muestras de cada estadio utilizadas, y por último bajo cada barra se encuentra un identificador del estadio o patrón de estadios cuyo recuento ha definido la altura de la barra. Los patrones de estadios son: de SI a SIV para los resultados que se encuentran solo en los estadios uno a cuatro respectivamente, DII y DIII para los resultados que se inician en el estadio dos o tres y se mantienen hasta el estadio 4 (incluido), y HII y HIII para aquellos resultados que se inician en el estadio uno y se mantienen hasta los estadios dos o tres respectivamente.



ANEXO 5. Diagramas de cajas del número de falsos positivos encontrados. La figura superior corresponde al análisis sin un ajuste por test múltiple, mientras que en la inferior si se aplicó el ajuste en su obtención. La prueba se realizó con muestras de KIRC, tipo tumor. Las líneas rojas de la figura superior marcan el 5% (40500) y 1% (8100) del total de resultados posibles. (Revelan las muestras que serían tenidas en cuenta si se filtrase exigiendo un p-valor del 0.05 o 0.01 respectivamente)