UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCOLA TÈCNICA SUPERIOR D´ENGINYERIA AGRONÒMICA I DEL MEDI NATURAL

Uso de datos genómicos para la búsqueda de reguladores de importancia en cáncer

TRABAJO FIN DE GRADO EN BIOTECNOLOGÍA

ALUMNO: MATÍAS MARÍN FALCO

DIRECTOR: JOAQUÍN DOPAZO BLAZQUEZ

TUTOR: JOSÉ JAVIER FORMENT MILLET

Curso Académico: 2014-2015

VALENCIA, SEPTIEMBRE 2015

Datos del trabajo

Título del TFG:	Uso de datos genómicos para la búsqueda de reguladores de importancia en cáncer.
Autor:	Matías Marín Falco
Localidad y fecha:	Valencia, 1 de septiembre 2015
Tutor:	José Javier Forment Millet
Director:	Joaquín Dopazo Blazquez
Tipo de licencia:	Licencia Creative Commons. “Reconocimiento no Comercial – Sin Obra Derivada”.

Resumen

En las últimas décadas el cáncer ha adquirido una relevancia e impacto muy importante en la sociedad. Se trata de una enfermedad que produce una proliferación celular sin control, causada por una disfunción de las redes de transducción celular. Por lo tanto, estudiar los mecanismos reguladores que lo desarrollan es vital para entender la enfermedad y acercarse a la creación de alguna terapia que lo combata.

Recientemente, ha surgido un nuevo enfoque para el estudio del cáncer y sus mecanismos moleculares, el análisis de pan-cáncer. Este análisis consiste en la determinación de las alteraciones comunes entre diferentes linajes de tumores con el fin de diseñar terapias efectivas en un tipo de cáncer y poder extenderlas a otros perfiles tumorales similares.

Para este trabajo, se propuso el estudio de la influencia de los factores de transcripción, en varios tipos de cáncer, sobre diferentes variables clínicas, como son la supervivencia o el estadio del tumor. Para ello, se recurrió a los datos genómicos recogidos en distintas bases de datos (*Ensembl*, ICGC y TCGA), con los que se realizaron, mediante el uso del lenguaje de programación R, un análisis del estadio tumoral y otro de supervivencia. Tras aplicar a estos datos una serie de análisis estadísticos (expresión diferencia, GSEA y análisis de supervivencia), se observó que ninguno de los factores de transcripción analizados estan específicamente relacionados con la aparición de un estadio tumoral en los cánceres estudiados. En cambio, parece ser que la alteración de un factor de transcripción, si esta ocurre, se produce en todos los estadios. A su vez, también se determinaron, para cada cáncer estudiado, una serie de factores de transcripción que en conjunto influyen en la supervivencia, así como factores de transcripción que por sí solos son capaces de influir de forma significativa en la supervivencia de un individuo. Entre estos últimos se han observado algunos marcadores de buen pronóstico, es decir, factores de transcripción cuya alteración indica una mayor supervivencia.

Palabras clave

Cáncer, factor de transcripción, bioinformática, transcriptómica, pan-cáncer, análisis de supervivencia, expresión diferencial, GSEA.

Abstract

In the last decades, cancer has increased its relevance and impact on society. This disease produces an uncontrolled cell proliferation, which is triggered by a dysfunction of signal transduction networks that regulate molecular communications and cellular processes. Studying this regulatory mechanisms is an essential task for understanding and developing effective therapies against cancer.

Recently, a new approach for studying cancer and its molecular mechanisms has emerged, pan-cancer analysis. This analysis aims to examine the similarities and differences among the genomic alterations found across diverse tumor types, with the purpose of designing new therapies against cancer, being able to apply them to other similar tumor profiles.

In this project, it was proposed the study of the influence of some transcription factors, in several cancer types, on some clinical endpoints, such as vital status and tumor stage. In order to achieve it, genomic data was retrieved from different databases (*Ensembl*, ICGC y TCGA) and, using R programming language, a survival analysis and tumor stage analysis were performed. After a series of statistical analysis (survival analysis, differential expression analysis and GSEA), the results pointed that any of the studied transcription factors were not related with a certain tumor stage. In fact, the alteration of a transcription factor generally occurs in all stages. A group of transcription factors, for each cancer, were also selected as predictors of a patient survival, and some of them were capable of influencing the survival function just by themselves.

Índice general

CAPÍTULO 1

Introducción

1.1. Ciencias ómicas

La biología se ha convertido en las últimas décadas en una ciencia con una enorme cantidad de datos cuyo volumen continúa creciendo de manera acelerada debido, en parte, al abaratamiento de los costes de secuenciación y desarrollo de nuevas tecnologías de alto rendimiento, hechos que han promovido la aparición de las ciencias *ómicas* (Casado-Vela *et al*., 2011). La adición a diferentes estudios biológicos del sufijo "-oma", de origen latino y que significa "conjunto de", cubre las nuevas aproximaciones masivas en las que se está enfocando la biología recientemente, en lugar de analizar elementos individuales.

Entre estos nuevos campos de la biología destacan la genómica (estudio de la secuencia genética en su totalidad), transcriptómica (estudio del conjunto de ARNm transcritos en la célula), proteómica (estudio del conjunto proteico en un momento y tejido determinados) y metabolómica (cuantificación y estudio del conjunto de metabolitos presentes en la célula en un estado fisiológico específico).

Las ciencias *ómicas*, al basarse en el análisis de un gran volumen de datos, hacen necesario, en muchos casos, el uso de la bioinformática en la interpretación y procesamiento de los mismos (Altman y Miller, 2011). La bioinformática es una disciplina que utiliza la tecnología de la información para organizar, analizar y distribuir la información sobre biomoléculas con la finalidad de responder a preguntas complejas. El libre acceso a Internet permite que toda esta información esté al alcance de las manos de todo el mundo.

Este trabajo se centrará en el uso de la bioinformática para el análisis de datos del transcriptoma, o conjunto de ARNs codificados por el genoma de una célula u organismo específico, en pacientes con distintos tipos de cáncer para integrarlo posteriormente con una base de datos de factores de transcripción y estudiar su influencia en distintas variables clínicas.

1.2. Cáncer y sociedad

Cáncer es un término genérico para un largo grupo de enfermedades que pueden afectar a distintos órganos del cuerpo. Una característica definitoria del cáncer es la rápida

proliferación de células anormales que crecen por encima de los límites normales (Balmain, 2001), y que pueden invadir tejidos adyacentes y propagarse a otros órganos distantes, proceso conocido como metástasis. La metástasis es la mayor causa de muerte por cáncer (WHO, 2014).

Todos los cánceres surgen debido a alteraciones en el ADN. En algunos casos dichas alteraciones se pueden presentar en la línea germinal, y por lo tanto, son heredables y confieren un elevado riesgo de desarrollar cáncer a la descendencia, pero en la mayoría de los casos son mutaciones en las células somáticas producidas a lo largo de la vida de una persona.

La incidencia y mortalidad a causa del cáncer están creciendo globalmente como resultado del crecimiento y envejecimiento de la población. En 2012 se diagnosticaron cerca de 14,1 millones de nuevos casos de cáncer y hubo sobre 8,2 millones de muertes (Ferlay *et al.*, 2012). En general, las mayores tasas de incidencia están asociadas a los países más ricos, mientras que en los países menos desarrollados hay una incidencia menor (Figura 1.1). Hay algunas excepciones a este patrón, como el caso de Uruguay, debido a la alta frecuencia de cáncer de pulmón (y otros cánceres relacionados con el tabaco), o el caso de Mongolia donde el cáncer de hígado tiene una alta tasa debido a la alta incidencia de las infecciones con los virus de las hepatitis B y C.

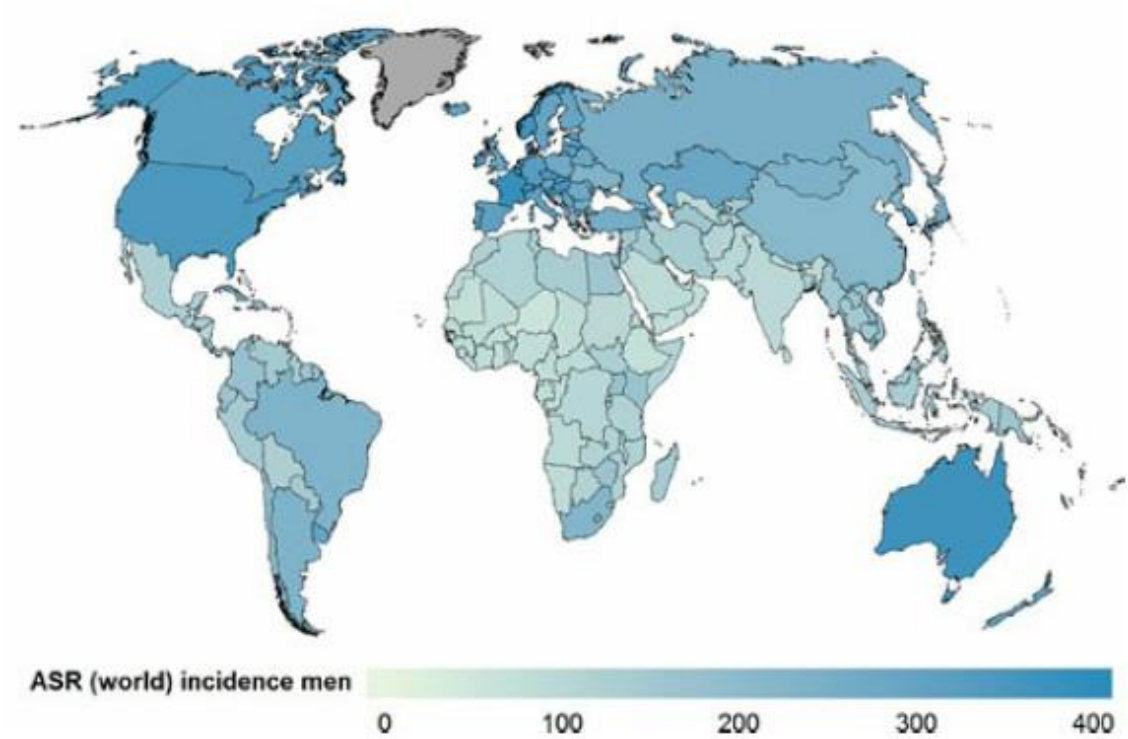

Figura 1.1: Tasas de incidencia por cada 100.000 habitantes, para todos los tipos de cáncer (a excepción del cáncer de piel no-melanómico), en hombres, en 2012. Fuente: Steward y Wild, 2014.

En la figura 1.2 se pueden observar las tasas de mortalidad estandarizadas por edad. Aunque las tasas de incidencia para todos los cánceres combinados eran el doble en los países más desarrollados, la diferencia en las tasas de mortalidad es solo de un 8% a un 15% mayor. Esto refleja las diferencias en la disponibilidad de tratamiento, detección y prevención (Edwards *et al.*, 2005). Además los cánceres asociados al estilo de vida de países industrializados, como los de pecho, colon o próstata, tienen un buen pronóstico, mientras que los cánceres de hígado, estómago y esófago, más comunes en países con bajos ingresos, tienen un pronóstico peor. En resumen, hay una mayor cantidad muertes debidas al cáncer por número de casos diagnosticados en países menos desarrollados.

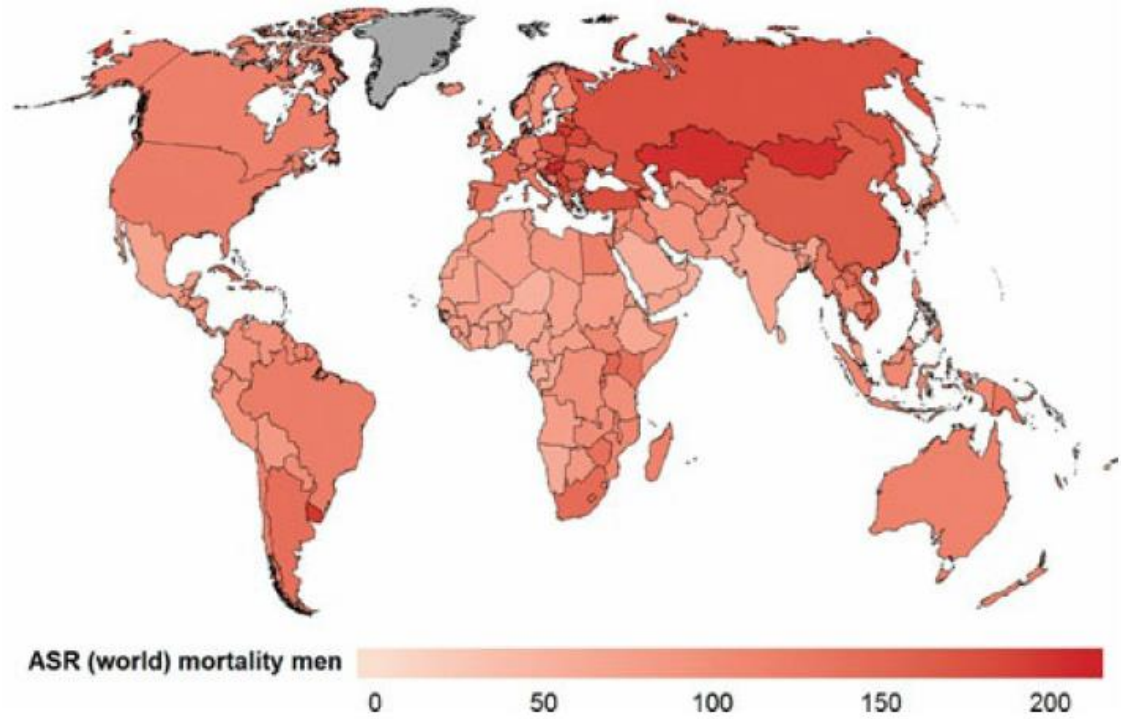

Figura 1.2: Tasas de mortalidad por cada 100.000 habitantes, para todos los tipos de cáncer (a excepción del cáncer de piel no-melanómico), en hombres, en 2012. Fuente: Steward y Wild, 2014.

En cuanto a la diferencia por sexos, entre los hombres, en 2012, los cinco tipos de cáncer más diagnosticados fueron el de pulmón (16,7% del total), próstata (15%), colon (10%), estómago (8,5%) e hígado (7,5%), siendo los causantes de mayor mortalidad el de pulmón (23,6% del total de muertes por cáncer), seguido por el de hígado (11,2%) y estómago (10,1%). Entre las mujeres, los cinco tipos de cáncer más comunes fueron el de pecho (25,2% del total), colon (9,2%), pulmón (8,7%), cérvix (7,9%) y estómago (4,8%), siendo los de mayor mortalidad el de pecho (14,7% del total), seguido por el de pulmón (13,8%) (Torre *et al*., 2015).

Las consecuencias del cáncer para los individuos, sus familias y la sociedad son enormes. El coste económico aproximado del cáncer según la Organización Mundial de la Salud

(OMS), incluyendo varios factores como prevención o tratamiento, es de 1,16 trillones de dólares americanos (Steward y Wild, 2014).

Por todo lo expuesto anteriormente el cáncer es una enfermedad, de gran relevancia en la sociedad, sobre la que se necesita continuar invirtiendo en prevención, diagnóstico y tratamiento para disminuir su prevalencia.

1.3. Regulación génica y cáncer

En organismos pluricelulares, prácticamente todas las células contienen la misma información genética, por tanto, la expresión génica diferencial es esencial para la formación y mantenimiento de los distintos tipos celulares. Las propiedades celulares vienen determinadas, tanto por la información genética codificada en su genoma, como por el conjunto de genes expresados, que varian en función de diferentes factores como el tejido o el momento del desarrollo celular. La activación o represión de genes seleccionados representa un delicado equilibrio para un organismo (homeostasis), ya que la expresión de un gen en un momento equivocado, en un tipo celular erróneo o en cantidades anormales puede conducir a un fenotipo deletéreo, como cáncer, aunque el gen sea totalmente normal (Klug *et al*., 2006).

Generalmente el desarrollo de un cáncer está ligado a la alteración en dos tipos generales de genes: los genes supresores de tumor, que normalmente actúan como "frenos" para inhibir el crecimiento y la división celular, y los protooncogenes, que normalmente actúan como "aceleradores". Las mutaciones que inhiben la actividad de los genes supresores de tumor (mutaciones de pérdida de función) o que sobreactivan protooncogenes (mutaciones de ganancia de función) pueden conducir a la producción de una célula cancerosa.

La primera mutación somática en un protooncogen causante de cáncer, fue descubierta en el gen humano HRAS (Reddy *et al*., 1982). Las mutaciones de ganancia de función, que convierten a los protooncogenes en oncogenes, actúan de manera dominante, es decir, el cambio de un alelo puede conducir a una proliferación celular descontrolada. Las mutaciones de ganancia de función pueden ser causadas por varias alteraciones, tanto por una mutación puntual que derive en una forma hiperactiva del protooncogen, como por la producción de copias extras del protooncogen y, por tanto, un mayor nivel de expresión del mismo, como por una mutación en la región promotora que produzca un aumento de su expresión (Chial, 2008).

Hay un grupo importante de protooncogenes que producen factores de transcripción. Un factor de transcripción es una proteína reguladora de la transcripción, que puede actuar de diferentes maneras, ya sea, reconociendo y uniéndose a secuencias concretas de ADN, uniéndose a otros factores, o uniéndose directamente a la ARN polimerasa. Los factores de transcripción son activados o desactivados selectivamente por señales citoplasmáticas, a menudo como paso final de la cadena de transmisión de señales intracelulares. Estos factores de regulación pueden actuar mediante mecanismos de control positivos (activación de la expresión) o negativos (represión de la expresión).

Hay factores de transcripción que ejercen papeles clave en la regulación de procesos de proliferación celular, inducción de apoptosis o reparación del ADN. Al tratarse de elementos reguladores que controlan la expresión de un elevado número de genes, son piezas extremadamente sensibles, cuya mutación puede suponer un gran impacto en la célula. Su activación oncogénica mediante una mutación de ganancia o pérdida de función determina que puedan actuar constitutivamente, sin necesidad de señales externas, estimulando así de manera continuada la síntesis de proteínas implicadas en la promoción del ciclo celular, dando lugar al crecimiento incontrolado de las células y procesos tumorales (IBMCC, 2015). Son numerosos los factores de transcripción cuya importancia ha sido demostrada en el cáncer, y es por esto que en este trabajo se utilizarán estos elementos reguladores como objeto de estudio. Entre los factores de transcripción que actúan como oncoproteínas cabe destacar algunos como MYC, MAX, MYB, FOS, JUN, REL o ETS (Chial, 2008).

En este trabajo se han realizado dos tipos de análisis cuyos resultados se presentarán por separado, aunque alguna parte de la metodología es común. Estos dos análisis corresponden al estudio de la influencia de los factores de transcripción en dos variables clínicas de los pacientes con cáncer: el estadio tumoral (*tumor stage*) y el estado vital (*vital status*).

El estadio tumoral es un indicador del desarrollo del cáncer que determina la extensión y gravedad de la enfermedad. Conocer el estadio en el que se encuentra un cáncer es importante ya que ayuda al médico a calcular el pronóstico y planificar el tratamiento apropiado de un paciente, además proporciona a los investigadores una terminología común para evaluar resultados de estudios clínicos y comparar entre diferentes estudios. El sistema de determinación de estadio tumoral más utilizado es el sistema TNM, que hace referencia al tamaño tumoral primario (T), a la afectación de ganglios linfáticos regionales o nodos linfáticos (N), y la diseminación a distancia del tumor primario o metástasis (M) (Byrd *et al*., 2010). Cada una de las letras anteriores va acompañada de un número que indica el estado del tumor del paciente (p.e. M0 indica que todavía no ha habido metastasis). Los ficheros clínicos de los cánceres elegidos en general emplean el sistema de determinación de estadio tumoral numérico, que utiliza el sistema TNM para agrupar a los pacientes en 4 estadios:

- Estadio I. Normalmente significa que el tumor es relativamente pequeño y está contenido dentro del órgano en el que comenzó. Se suele poder quitar mediante intervención quirúrgica.
- Estadio II. El tumor es más grande que en el estadio I y todavía no ha empezado a extenderse en los tejidos adyacentes. A veces, dependiendo del tipo de cáncer, el estadio II significa que las células cancerígenas se han expandido a los nódulos linfáticos cerca del tumor.
- Estadio III. El tamaño del tumor es mayor, y puede haber empezado a propagarse a los tejidos contiguos y haber llegado a los nódulos linfáticos de la zona.
- Estadio IV. El cáncer se ha propagado de donde empezó a otro órgano del individuo (metástasis).

Para el caso del estado vital, se realizó un análisis de supervivencia en el que se requería los datos del estado del paciente (si estaba vivo o ha fallecido) y los días que han pasado desde su diagnóstico.

1.4. Pan-cáncer

De los estudios realizados en las últimas décadas han emergido principios generales importantes sobre los distintos tipos de cáncer (Hanahan y Weinberg, 2011). Y es que hasta hace poco, la mayoría de la investigación sobre la naturaleza molecular, patológica y clínica del cáncer han sido agrupadas según el tipo de tumor. Solo hace falta mirar a la estructuración de departamentos de cualquiera de los mayores centros de cáncer para darse cuenta que la atención médica y quirúrgica están, en su gran mayoría, divididos por enfermedad según el tipo de órgano de origen (McDermott y Settleman, 2009). Los avances sobre el conocimiento de esta enfermedad demuestran que esta clasificación ha funcionado bien hasta ahora, pero también es verdad que cánceres de distintos órganos comparten algunas características similares, y contrariamente, algunos cánceres del mismo órgano pueden ser bastante distintos. Por ello, recientemente, se propuso la observación de las alteraciones comunes entre diferentes linajes de tumores con el fin de diseñar terapias efectivas en un tipo de cáncer y poder extenderlas a otros perfiles tumorales similares. Esto se conoce como análisis de pan-cáncer, concepto acuñado por la red de investigadores integrados en el proyecto del The Cancer Genome Atlas (TCGA) (Weinstein *et al*., 2013).

El análisis de pan-cáncer es un análisis complicado. Debido a la complejidad de la enfermedad y a que se recogen datos de distintos tipos tumorales y que pueden proceder de distintas plataformas, es difícil que los datos hayan sido obtenidos con los mismos criterios y por tanto, es difícil, a veces, extraer conclusiones de su análisis. Pero a pesar de dichas dificultades, ya ha habido importantes descubrimientos sobre procesos similares entre subtipos tumorales de diferentes tejidos. Por ejemplo, mutaciones oncogénicas en TP53 en distintos tipos de cáncer (ovario, endometrio y pecho), que comparten características transcripcionales y que convergen en rutas oncogénicas similares (Cancer Genome Atlas Research Network, 2013); o aberraciones en los genes de la familia NOTCH que tienen diferentes efectos dependiendo del órgano en el que surgen, inactivación en cánceres de cabeza y cuello (Stransky *et al*., 2011), pulmón, piel (Wang *et al*., 2011) y cérvix (Zagouras *et al*., 1995), pero activado por mutación en leucemias (Weng *et al*., 2004). Tales ejemplos muestran la importancia de desarrollar una visión que consiga integrar los datos procedentes de distintos órganos para así ayudar a identificar variaciones en las consecuencias de mutaciones en distintos tejidos, lo que puede conllevar importantes implicaciones terapéuticas (Weinstein *et al*., 2013).

En este contexto surge el International Cancer Genome Consortium (ICGC) (ICGC, 2015). Este consorcio intenta englobar la mayoría de proyectos internacionales que hay ahora sobre el cáncer, como pueden ser proyectos de secuenciación,

identificación de mutaciones, expresión génica, etc. El IGCG persigue cumplir una serie de objetivos:

- Crear almacén común donde se encuentran todos los datos sobre cáncer de los distintos proyectos que se realizan en distintas partes del mundo, y así hacerlos fácilmente disponibles para cualquier investigador
- Hacer posible la comparación de las diferencias moleculares entre subtipos de cáncer específicos que se encuentran en diferentes áreas geográficas.
- Proporcionar información de los métodos que han sido utilizados para producir, analizar e integrar los grandes conjuntos de datos.

1.5. Bioinformática y R

Como consecuencia de la aparición de las ciencias *ómicas* (apartado 1.1) y la gran cantidad de datos biológicos están siendo producidos (Reichhardt, 1999), los ordenadores se han convertido en indispensables en la mayoría de investigaciones biológicas, ya que permiten manejar y procesar tales cantidades de datos.

La bioinformática se define como la aplicación de técnicas computacionales para la gestión y análisis de datos biológicos, y persigue tres objetivos principales (Luscombe *et al*., 2001). El primer objetivo consiste en la organización de los datos de manera que se permita a los investigadores acceder a la información existente e introducir nuevos datos conforme se van produciendo. Para ello se han creado bases de datos o consorcios internacionales, como el NCBI (NCBI Resource Coordinators, 2013) o el ICGC (Apartado 1.4). El segundo es el desarrollo de herramientas y recursos que ayuden al análisis de los datos, como es el caso de la herramienta BLAST (Camacho *et al*., 2008). El tercer objetivo es usar dichas herramientas para analizar los datos e integrarlos con los conocimientos disponibles para interpretar los resultados de manera que tengan un significado biológico. De esta manera, la bioinformática también permite el estudio de los datos disponibles para buscar principios comunes a lo largo de distintos sistemas y así descubrir nuevas características.

Para el análisis bioinformático o la creación de nuevas herramientas es necesario utilizar distintos lenguajes de programación. Entre ellos destacan algunos como *python* (van Rossum, 2003), *Perl* (Wall *et al*., 2004) o la misma línea de comandos de Linux. Pero sin duda, uno de los lenguajes más utilizados en el área de la bioinformática, y el que mayoritariamente se emplea en este trabajo, es el lenguaje de programación R.

R es un software libre enfocado al análisis estadístico y gráfico, inspirado en el lenguaje S, uno de los lenguajes más utilizados en investigación por la comunidad estadística. Fue desarrollado inicialmente por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en 1993 (Ihaka y

Gentleman, 1996), y actualmente su desarrollo es responsabilidad del R Development Core Team.

El lenguaje y entorno de programación R se ha convertido en una herramienta muy empleada actualmente en el campo de la investigación biomédica (p.e. Tang *et al.*, 2013), bioinformática (p.e. He *et al.*, 2015), estadística y las matemáticas financieras (p.e. Spedicato, 2013).

La mayoría de este trabajo se ha realizado con el lenguaje de programación R, debido a sus características, que se exponen a continuación:

- Su manejo y almacenamiento de los datos es sencillo.
- Concede un excelente control de todos los parámetros de los gráficos para un análisis personal de los datos.
- Tiene la posibilidad de cargar diferentes bibliotecas o una gran variedad de paquetes (funciones, datos o códigos) con finalidades específicas de cálculo o gráfico.
- Es un software libre, disponible para diferentes tipos de sistemas operativos, bastante simple, completo y efectivo.
- Es de código abierto, de manera que se puede mirar el código fuente para arreglar cualquier error o añadirle nuevas funciones.

R emplea programación literaria, de manera que hace posible y más fácil realizar una investigación reproducible. Es importante en el análisis computacional de datos hacer una investigación reproducible (*reproducible research*) y es algo que se debe cuidar desde el principio de la investigación (Donoho, 2010). Es decir, es importante que estén disponibles e inteligibles las instrucciones del análisis de datos, de modo que cualquier investigador pueda recrear el mismo análisis con nuevos datos experimentales o verificar o actualizar el análisis con los mismos datos.

Los distintos paquetes de R y sus respectivos manuales se encuentran almacenados en varios repositorios. El repositorio de paquetes más importante es *CRAN* (*Comprehensive R Archive Network*), pero además de *CRAN*, en este trabajo, también se ha recurrido a paquetes del repositorio *Bioconductor*, que contiene paquetes destinados al análisis de datos genómicos obtenidos mediante tecnologías de alto rendimiento.

El uso de R es mucho más cómodo usando la interfaz *RStudio* (Racine, 2012), ya que permite realizar todos los pasos del análisis de manera fácil y visual. Aporta información sobre los objetos que se encuentran en el entorno de trabajo, permite ver en la misma interfaz las gráficas que se vayan haciendo y también consultar los manuales de los distintos paquetes, en definitiva, da muchas facilidades a la hora de realizar una investigación.

CAPÍTULO 2

Objetivos

Dada la importancia y el impacto del cáncer en la sociedad y la importancia para su desarrollo de los mecanismos reguladores, se ha planteado como objetivo principal realizar un estudio que mejore el conocimiento sobre la influencia de la alteración de ciertos factores de transcripción, extraídos de la base de datos de *Ensembl*, en diferentes variables clínicas, como son la supervivencia o el estadio del tumor, recogidas del TCGA (TCGA, 2015) y del ICGC.

Para ello se aplicarán diversos test estadísticos que permitan valorar las diferencias de expresión en cada tipo de cáncer. Seguidamente se evaluarán en conjunto, a fin de obtener aquellas diferencias más significativas y representadas a lo largo de todos los tipos de cáncer.

Para cumplir el objetivo principal será necesario desarrollar una metodología que permita obtener, limpiar, analizar y discutir los datos mencionados anteriormente. De manera que se plantean los siguientes objetivos:

- Seleccionar los cánceres a estudiar.
- Obtener la información necesaria de los cánceres seleccionados, cribarlos y cruzarlos.
- Realizar análisis estadísticos, como la expresión diferencial, en los diferentes tipos de cáncer.
- Llevar a cabo un análisis en conjunto (metaanálisis) que permita obtener resultados más robustos y extrapolables.

Por consiguiente, se espera obtener algunos factores de transcripción que estén relacionados con alguna variable clínica u observar algún patrón de comportamiento en los mismos, ayudando a esclarecer la implicación de los factores de transcripción en el cáncer.

También es un objetivo importante de este trabajo el seguir el concepto *Reproducible research* (Apartado 1.5) para que de este modo se pueda reproducir de manera fácil la metodología utilizada en una actualización de los datos utilizados o en nuevos datos en caso de que surjan.

CAPÍTULO 3

Materiales y métodos

El trabajo realizado sigue el esquema representado en la figura 3.1. De tal manera, que tras la obtención de los transcriptomas de los distintos cánceres procedentes del ICGC, se siguió una metodología que tiene como fin, estudiar la influencia de la alteración de factores de transcripción en dos variables clínicas (estadio tumoral y estado vital). Para lo cual se llevaron a cabo dos análisis con metodologías distintas, aunque con algunos elementos en común.

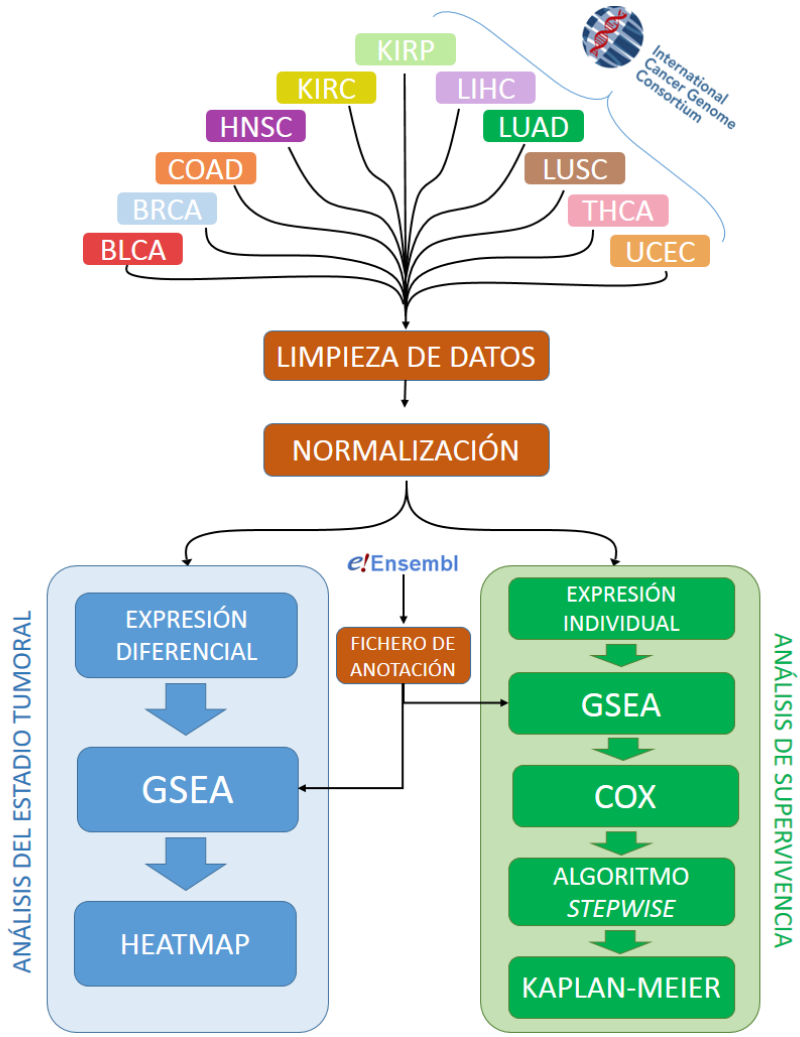

Figura 3.1. Esquema representativo de la metodología llevada a cabo para realizar el análisis de supervivencia y del estadio tumoral.

3.1. Recursos informáticos

Al trabajar con una gran cantidad de datos, fue necesario disponer de recursos informáticos con gran capacidad de computación. Para ello se utilizó el *cluster* (conjunto de máquinas que comparten *hardware* y están conectadas entre sí) del laboratorio de Genómica de Sistemas del Centro de investigación Príncipe Felipe (CIPF), que está compuesto por 36 máquinas, con 362 núcleos y 2,6 TB de memoria RAM en conjunto, y un almacenamiento en disco de 262 TB.Para la administración de tareas que se envían al *cluster* se usa un sistema de colas SGE (Gentzsch, 2001), que también permite correr procesos paralelamente en más de una máquina usando el protocolo MPI.

Para desarrollar los programas que permitieron realizar los análisis de este estudio, se empleó, principalmente, el lenguaje de programación R debido a las grandes ventajas que presenta (Apartado 1.5). Aunque para algunas tareas fue necesario recurrir a *python* o la línea de comandos de Linux, que fue necesario para poder trabajar en el *cluster*. Los paquetes de R que se han utilizado son:

- *ff* (Adler *et al*., 2013)
- *ffbase* (de Jonge *et al*., 2015)
- *edgeR* (Robinson *et al*., 2010)
- *limma* (Ritchie *et al*., 2015)
- *survival* (Therneau, 2014)

3.2. Obtención de datos

3.2.1. Datos transcriptómicos

Los datos de los transcriptomas de los distintos cánceres fueron obtenidos del ICGC (Figura 3.1), de donde que se han extraídos dos tipos de archivos para cada cáncer, uno con los datos de la secuenciación del transcriptoma, y otro con los datos clínicos de los pacientes secuenciados.

Las matrices de expresión génica utilizadas fueron obtenidas a partir de experimentos realizados con la técnica de secuenciación RNA-Seq. En estos archivos se encuentran los datos de expresión de distintos pacientes, de los que generalmente se dispone una muestra tanto tumoral, como de tejido sano.

También se han utilizado los archivos clínicos proporcionados por el TCGA, de los que se obtuvieron la información del estadio del tumor (apartado 4.2) y del estado vital (apartado 4.3) de los pacientes.

3.2.2. Selección de los tipos de cáncer

De todos los proyectos recogidos en el ICGC se seleccionaron 11 tipos de cáncer atendiendo a 2 criterios: la disponibilidad de muestras pareadas para el análisis del estadio tumoral (muestras de transcriptoma tanto de tejido tumoral como de tejido normal adyacente al tumor), y la disponibilidad de datos clínicos (tanto de estadio del tumor como estado vital).

Los cánceres seleccionados y la disponibilidad de los datos mencionados se muestran en la siguiente tabla.

Tabla 3.1.Número de muestras disponibles en función de cada tipo de cáncer seleccionado.

	Tumor	Normal	Estadio I	Estadio II	Estadio III	Estadio IV	Vivos	Fallecidos
***Bladder Urothelial Cancer* [BLCA]**	294	17	1	95	99	98	221	73
***Breast Cancer* [BRCA]**	1039	113	177	591	237	17	937	98
***Colon Adenocarcinoma* [COAD]**	428	41	73	168	120	58	374	53
***Head and Neck Squamous Cell Carcinoma* [HNSC]**	480	42	26	74	72	245	320	158
***Kidney Renal Clear Cell Carcinoma* [KIRC]**	517	72	256	56	125	81	358	159
***Kidney Renal Papillary Cell Carcinoma* [KIRP]**	222	32	138	16	43	13	199	23
***Liver Hepatocellular carcinoma* [LIHC]**	294	48	132	66	71	5	222	72
***Lung Adenocarcinoma* [LUAD]**	473	55	255	116	81	24	355	118
***Lung Squamous Cell Carcinoma* [LUSC]**	426	45	217	128	75	6	290	136
***Head and Neck Thyroid Carcinoma* [THCA]**	500	58	282	54	110	52	481	14
***Uterine Corpus Endometrial Carcinoma* [UCEC]**	508	23	318	49	114	27	464	43

Todos los tipos de cáncer seleccionados fueron extraídos del repositorio oficial del ICGC, debido a las numerosas ventajas que ofrece (apartado 1.4), y también a que disponen de una interfaz web que permite obtener los datos con mayor facilidad. Estos cánceres proceden del proyecto del TCGA, ya que son los únicos que poseen datos transcriptómicos y clínicos de un número elevado de pacientes. Este hecho ha facilitado el análisis, ya que las muestras tenían el mismo formato y la información se ha obtenido empleando las mismas metodologías.

3.2.3. Datos de regulación génica

Los elementos reguladores seleccionados para el estudio son los factores de transcripción. Hubo que seguir una serie de pasos, que a continuación se exponen, para obtener la información sobre de los factores de transcripción que se requiere para los posteriores análisis (lista de factores y su sitio de unión en el genoma, y lista de genes y su posición en el genoma).

La información fue extraída de la base de datos de *Ensembl*, de su edición 81, que utiliza la versión del genoma humano 38 parche 3 (*GRCh38.p3*). El proyecto *Ensembl* es una fuente de datos relacionados con la secuencia del genoma de varios organismos, con especial énfasis en el humano, organismos modelos vertebrados y animales de granja, también es un sistema de software libre que puede usarse para la organización de dichos datos (Flicek *et al.*, 2013). Ensembl dispone de una herramienta web de extracción de datos (*BioMart*) que permite filtrar y obtener los datos requeridos de forma sencilla.

En este caso se descargaron con *BioMart* los sitios de unión de los 70 factores de transcripción disponibles en el la base de datos "*Ensembl Regulation 81*" en el set "*Homo sapiens Regulatory Evidence* (*GRCh38.p3*)" utilizando el filtro "*Transcription Factor*" y seleccionando los atributos "*Chromosome Name*", "*Start (bp)*", "*End (bp)*" y "*Feature type*". De esta manera se obtiene una lista con las coordenadas de los sitios de unión de los diferentes factores de transcripción disponibles.

Para determinar los genes regulados por cada factor de transcripción se obtuvieron las coordenadas de todos los genes se utilizó la base de datos "*Ensembl Genes 81*" con el set de datos "*Homo sapiens genes (GRCh38.p3)*" y filtrando para obtener los genes anotados y conocidos ("*Status (genes): KNOWN*"), seleccionando los atributos "*Chromosome Name*", "*Gene Start (bp)*", "*Gene End (bp)*", "*Strand*" y "*HGNC symbol*".

3.3. Limpieza de datos

3.3.1. Formación de tablas de expresión

Los archivos que contienen los transcriptomas de los distintos cánceres están formados por millones de líneas, en los que cada línea contiene el valor de expresión de un gen concreto en una muestra concreta. Para poner disponer mejor de los datos en los análisis posteriores, se procedió a la creación de una matriz, para cada tipo de cáncer, que contuviese en las columnas las distintas muestras y en las filas los distintos genes. A continuación se muestra la metodología seguida para obtener dichas matrices.

Primero, tras utilizar el lenguaje de programación *python* para la extracción de la lista de genes disponibles en los archivos de expresión y una lista de los pacientes, se eliminaron del muestreo aquellos pacientes que no tenían información en el archivo clínico, ya que a pesar de disponer de datos de expresión, si no se dispone de la información clínica, no

sería posible discernir si dichos datos pertenecen a la expresión de una célula sana o tumoral.

Posteriormente, para el análisis del estadio tumoral se miró, en el fichero clínico del ICGC, la columna "*specimen_type*", y se seleccionaron los pacientes con muestras pareadas, es decir, que hubiese al menos 2 muestras de un mismo paciente pertenecientes a una muestra de tejido tumoral ("*Primary tumour - solid tissue*") y otra muestra de tejido sano adyacente al tumor ("*Normal – tissue adjacent to primary*"). Esta metodología permite determinar más rigurosamente las diferencias de expresión debidas al estadio tumoral en sí, y se intenta evitar la influencia de otros factores como el tipo celular o la variabilidad genética (por ser las muestras que se comparan de distintos pacientes). Para realizar el análisis de supervivencia únicamente se contrastó que los datos del transcriptoma tuvieran su correspondiente información clínica.

Debido al tamaño de los archivos, y a que R consume muchos recursos al gestionar ficheros grandes, fue necesario utilizar paquetes específicos de R (*ff* y *ffbase*) para poder acceder a los datos transcriptómicos y manipularlos de forma eficiente. Para utilizar estos datos se creó una matriz vacía con las muestras seleccionadas como identificador de columnas y los genes disponibles en el archivo de expresión como nombre de filas. De esta manera, se pudo ir recorriendo secuencialmente el archivo de expresión fila por fila y rellenando cada celda con su valor correspondiente. En este caso las celdas fueron rellenadas con el valor de número de lecturas procedente de la secuenciación que ofrecía el archivo (*raw read count*).

3.3.2. Obtención del fichero de anotación

En el posterior análisis de enriquecimiento (Apartado 3.6) se requiere de un archivo de anotación. Este archivo consiste en un fichero que contiene todos los emparejamientos disponibles entre los factores de transcripción y los genes a los que se une (*targets*).

Para obtener este archivo se utilizaron los datos extraídos del *Ensembl*, en el que se dispone de las coordenadas de los sitios de unión de los factores de transcripción, o TFBS (del inglés, *Transcriptor factor binding site*), y de los genes.

En primer lugar, se miró la distancia en pares de bases (pb) entre los TFBS y los genes, para intentar definir la longitud de la zona promotora, con la que posteriormente se cruzarían las coordenadas de los TFBS para hacer una predicción de los *targets*. Para ello se utilizó la herramienta *BEDtools* (Quinlan y Hall, 2010) con el comando *closest*, en el que, a partir de las coordenadas introducidas, se devuelve un fichero indicando la distancia (en pb) entre las mismas. Finalmente, una vez se definieron las regiones promotoras, con la misma herramienta, pero utilizando el comando *intersect*, se cruzaron las coordenadas de los promotores con las de los TFBS y se creó el fichero de anotación que contiene los emparejamientos entre los factores de transcripción y sus *targets*.

3.4. Normalización

La normalización permite la comparación de niveles de expresión entre distintas muestras, e incluso dentro de las mismas muestras en réplicas distintas (comparación de la expresión de distintos genes en un individuo) (Marioni *et al*., 2008). Pretende asegurar que las diferencias en el número de lecturas obtenidas a la hora de comparar dos muestras realmente reflejen la expresión diferencial de los genes y no sesgos artificiales derivados de la secuenciación.

Existen varios métodos de normalización de datos genómicos, algunos de ellos como el ajuste según el tamaño de la biblioteca genómica (total de lecturas) son aproximaciones demasiado simples para varias aplicaciones biológicas, aunque sí que pueden servir para la normalización entre réplicas de muestras. El método de normalización elegido se trata de un método simple y efectivo para la estimación de los niveles de producción de ARN procedentes de datos de RNA-Seq, ya que baja la tasa de falsos positivos en el análisis de expresión diferencial (Robinson y Oshlack, 2010). Dicho método se trata de la normalización por TMM *(trimmed mean of M-values normalization*), que fue aplicada a los datos transcriptómicos preparados en matrices usando el paquete de R *edgeR*.

3.5. Expresión diferencial

El uso de tecnologías que miden la expresión génica es frecuente en el campo de la biología molecular para obtener una imagen de la actividad transcripcional en diferentes tejidos o poblaciones celulares. Estos perfiles después son comparados para identificar cambios en la expresión de genes asociados a un tratamiento o fenotipo de interés, lo que se conoce como análisis de la expresión diferencial.

A pesar de la complejidad, en los estudios de expresión génica, el número de muestras biológicas no suele ser muy elevado, lo que supone un reto estadístico para conseguir extraer la máxima información fiable a partir datos disponibles. Para ello se ha desarrollado el paquete de R *limma*, el cual es usado en este estudio para realizar el análisis de expresión diferencial entre dos grupos en todos los tipos de cánceres seleccionados. Primero entre las muestras normales y las tumorales pareadas, y posteriormente entre las muestras sanas y las muestras tumorales en los diferentes estadios.

A partir de los datos normalizados, *limma* es capaz de realizar el análisis de expresión diferencial utilizando datos de distintas plataformas, incluida RNA-Seq. Esto se consigue creando una matriz de ceros y unos que sirva de referencia (matriz de diseño) e indique cuáles serán los casos de referencia (las muestras normales, a las que se les asociará el 0) y cuáles los casos a comparar (muestras de tumor o de distintos estadios). A continuación se usa la función v*oom*, la cual estima la relación de la varianza media de los logaritmos base dos de las lecturas normalizadas por millón y genera una precisa matriz de pesos para cada observación (Law *et al*., 2014). A partir de la transformación anterior se puede continuar con el análisis predeterminado de *limma*, en el que, mediante la función *lmFit*,

se ajusta la tabla obtenida a un modelo lineal y posteriormente se emplea el método empírico de Bayes para estimar un valor estadístico, el cual se utilizará para obtener una lista ordenada de genes en función de su expresión diferencial.

Por lo tanto, como resultado final de la expresión diferencial se extrae una lista, para cada tipo de cáncer seleccionado con dos parámetros estadísticos:

- P-valor. Es la probabilidad de obtener un resultado como el observado, asumiendo que la hipótesis nula es cierta. En este caso la hipótesis nula consiste en suponer que no hay diferencias en la expresión en los genes de las distintas muestras. De manera que si el p-valor es menor a un valor determinado (generalmente se establece 0,05), esto indica que un gen está diferencialmente expresado.

- T estadístico. Un valor positivo o negativo de *t* indica que un gen está más o menos expresado, respectivamente, que en la muestra de referencia, en este caso una muestra de tejido sano.

3.6. Análisis de enriquecimiento de grupos de genes

El análisis de expresión diferencial proporciona como resultado una larga lista de genes con unos valores asociados. Una manera de interpretar los resultados es mediante una visión holística, como la que proporciona la biología de sistemas, que es el campo de investigación que intenta explicar interacciones de procesos biológicos a través de modelos matemáticos. En el análisis habitual, los genes se ordenan en función del valor estadístico obtenido en el análisis de expresión diferencial y se seleccionan los genes se encuentran al principio y al final de la lista. Para ello había que asignar límites (*thresholds*) a las listas, y así poder decidir que formaba parte de estos extremos. El tener que asignar umbrales de confianza suele implicar bastantes complicaciones (Subramanian *et al*., 2005), y para evitarlas surgió el análisis de enriquecimiento de grupos de genes o GSEA (del inglés, *Gene Set Enrichment Analysis*).

El GSEA determina si un bloque, compuesto por un grupo definido de genes, en su mayoría se encuentra ubicado en alguno de los extremos de una lista de genes ordenada de acuerdo a algún criterio biológico (como la expresión diferencial obtenida al comparar dos condiciones) sin necesidad de recurrir a un umbral de confianza (Figura 3.2). Los grupos que conforman los bloques de genes, no confundir con la lista de genes obtenida en la expresión diferencial, se definen en base a un conocimiento biológico previo (Al-Shahrour *et al*., 2007). En el caso de este estudio, el fichero de anotación contendrá la información de los distintos bloques que serán empleados en el GSEA. Cada bloque corresponde al conjunto de *targets* de un factor de transcripción, por lo que habrá un bloque por cada factor de transcripción estudiado.

El análisis de enriquecimiento consigue amplificar y medir la influencia de un factor de transcripción más allá de su nivel de expresión, ya que éste no siempre es un indicador de su actividad celular (Schacht *et al*., 2014). La unión de un factor de transcripción a su TFBS no implica necesariamente un impacto regulatorio si el *target* no se ha expresado,

ya que la expresión de un gen no depende solo de la unión de un factor de transcripción. Además existen varios pasos entre la transcripción del ARNm y la regulación de los *targets*, y los factores de transcripción pueden sufrir modificaciones post-transcripcionales (metilación, ubiquitinación, fosforlación, etc.). Y se ha demostrado que dichas modificaciones pueden tener un impacto sustancial en la regulación de los *targets* (Filtz *et al*., 2014; Tootle y Rebay, 2005). Por lo tanto, la aplicación del GSEA en este estudio permite la evaluación indirecta de la influencia de un factor de transcripción a través de sus *targets*, y así realizar un análisis con mayor precisión.

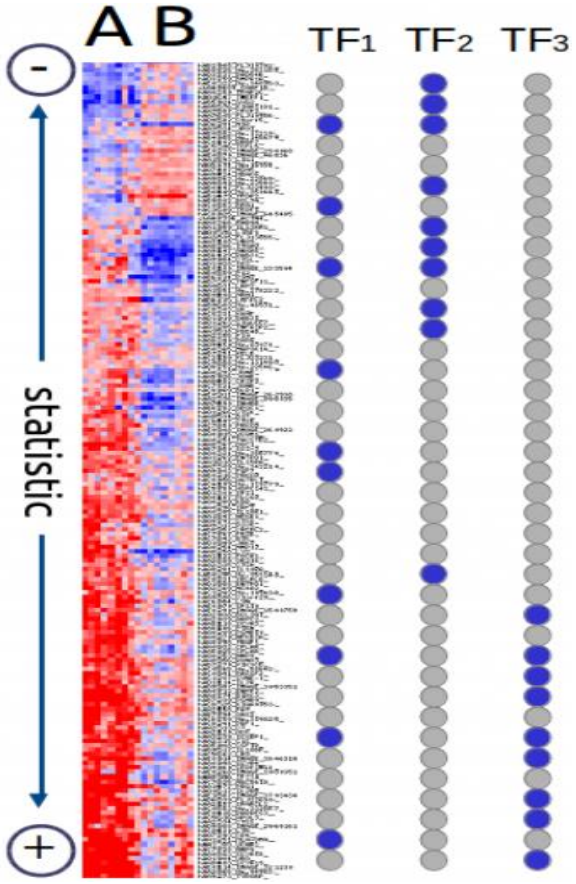

Figura 3.2. Esquema explicativo del funcionamiento del GSEA. Fuente: Al-Shahrour *et al*., 2007

La figura 3.2 muestra un ejemplo del análisis realizado, donde A y B son las muestras con distintas variables clínicas (p.e. caso-control) y sobre las que se ha realizado un análisis de expresión diferencial, del cual se extrae la lista de genes ordenados de menor a mayor valor estadístico. A la derecha se observa en qué posición de la lista quedan los *targets* de tres factores de transcripción que conforman 3 bloques de genes. En este ejemplo se observa claramente como los *targets* de los factores de transcripción TF2 y TF3 se ubican principalmente en los extremos de la lista, mientras que los *targets* de TF1 se encuentran distribuidos simétricamente por la lista. Por lo tanto el resultado esperado al aplicar el GSEA sería que los factores de transcripción TF2 y TF3 son significativos en la variable clínica medida en la expresión diferencial.

Con los datos de expresión diferencial producidos en el apartado anterior se realizó el GSEA utilizando la función *uvGsa*, que es la función que emplea para este análisis la herramienta Babelomics 5.0 (Alonso *et al*., 2015). El análisis se realiza con un modelo regresión logística, un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica en función de otras variables independientes (Montaner y Dopazo, 2010). Como resultado se obtiene una tabla en la que se indican para cada factor de transcripción el valor de varios parámetros estadísticos obtenidos en el análisis:

- El número de *targets* anotados en el bloque de cada factor de transcripción.
- El logaritmo de los ratios de probabilidad (*log odds ratio*) estimado para cada factor de transcripción. Este valor representa la ubicación de los *targets* en la lista de genes obtenida en la expresión diferencial.
- El p-valor, que indica la probabilidad de obtener aleatoriamente un resultado como el observado, asumiendo que la hipótesis nula es cierta. En este caso la hipótesis nula consiste en que el factor de transcripción no tiene peso en el proceso clínico, y como resultado la mayoría de sus *targets* no están diferencialmente expresados.
- El p-valor ajustado, reduce la probabilidad de encontrar resultados significativos por azar.

3.7. Análisis de supervivencia

El análisis de supervivencia es un tipo de análisis estadístico basado en el seguimiento de los individuos de un estudio desde una experiencia inicial o exposición hasta la ocurrencia de un evento. El evento observado suele ser una variable dicotómica, es decir, puede tomar dos valores. En este caso la variable observada fue el fallecimiento de los pacientes, evento que suele ser habitual en este análisis, ya que de ahí procede el nombre de análisis de supervivencia. Este tipo de análisis es ampliamente usado en la investigación clínica y epidemiológica (p.e. para el estudio de enfermedades o de efectos de fármacos), pero también es utilizado en otros campos que no están relacionados con la salud, como la ingeniería o el sector financiero. Un aspecto importante del análisis de supervivencia es que tiene en cuenta los periodos de seguimiento de los individuos, por lo que no tendrá el mismo peso en el análisis un evento que ocurre a la semana de comenzar el estudio que uno que ocurre al final. También permite incorporar al modelo los datos de individuos que se incorporan tarde al estudio o que nunca sufren el evento de interés, ya sea debido a que se pierde su seguimiento en el estudio, porque el paciente sufre otro evento distinto al estudiado o porque no ocurra el evento durante el estudio (datos censurados) (Flynn, 2012).

Para realizar el análisis de supervivencia, primero hubo que construir un modelo que permita discernir aquellos factores de transcripción cuya alteración afecte significativamente a la supervivencia de los individuos en cada cáncer. Entre los diferentes tipos de modelos multivariantes, uno de los más extendidos en medicina es el modelo de riesgos proporcionales, también conocido como modelo de Cox (Cox, 1972). La regresión de Cox busca cuales de las variables independientes introducidas en el modelo se relacionan con variaciones en la función de supervivencia, para lo cual calculará, para cada variable (factores de transcripción), la probabilidad de que no influya

en el modelo construido (p-valor), y también un coeficiente que indique el peso de las variables en el modelo.

Para poder aplicar el análisis de supervivencia al estudio que se pretende hacer, fue necesario desarrollar una metodología que permitiese asignar un valor de las variables independientes, a cada individuo, ya que los valores obtenidos en la expresión diferencial y posteriormente el GSEA son valores globales, calculado a partir de todo el conjunto de individuos. Dicha metodología consiste en realizar un GSEA para cada uno de los individuos.

Primeramente hubo que obtener la lista ordenada de genes en cada individuo, necesaria para la aplicación del GSEA. Dicha lista se obtuvo comparando el valor normalizado de expresión de cada gen (v) frente a la media de la expresión genética del conjunto de las muestras normales (m_n) y dividido por la desviación estándar de la misma (sd_n) (ecuación 4.1), de esta manera se consigue un valor que mide la distancia de cada gen en un individuo con respecto al conjunto de genes de las muestras normales. Una vez ordenada la lista de genes se realiza el GSEA con el fichero de anotación (apartado 3.3.2), y se determina en cada individuo si cada uno de los factores de transcripción estudiados están alterados (si el p-valor obtenido del GSEA es menor a 0,05). De esta manera cada individuo dispondrá de una variable dicotómica (con solo dos valores posibles) independiente e individual para cada uno de los factores de transcripción.

$$\frac{v-m_n}{sd_n} \qquad \text{(Ecuación 4.1)}$$

Una vez obtenido el valor, para cada individuo, de las variables independientes se aplicó el modelo de regresión de Cox utilizando la función *coxph* del paquete de R *survival*.

El modelo de Cox obtenido puede ser mejorado, ya que *a priori* no se conoce qué variables (factores de transcripción) están relacionadas con la supervivencia, y al introducirlas todas en el modelo de regresión, se aumenta el ruido y la precisión del mismo. Para esta mejora se utilizó la función de R *step*, mediante la cual se mejora el modelo utilizando un algoritmo *stepwise*, que selecciona, de entre todas las variables candidatas a ser explicativas de la variable dependiente, un subconjunto que resulte suficientemente explicativo y también no demasiado complejo. Como resultado se obtiene un grupo de factores de transcripción en cada tipo de cáncer seleccionados cuya alteración afecta significativamente a la supervivencia.

Como último paso se hizo un análisis univariante, para determinar si la alteración de alguno de los factores de transcripción que eran significativos en el modelo anterior, por sí solo, era capaz de predecir significativamente la supervivencia en un cáncer. Para ello se empleó una de las técnicas más realizadas en los análisis de supervivencia, la función de Kaplan-Meier (Kaplan y Meier, 1958). Se trata de una técnica que considera en distintos puntos el número de pacientes que permanecen en la población y el número de eventos acumulados que han ocurrido hasta ese punto, a partir de lo cual, va calculando la probabilidad acumulada de supervivencia en los distintos momentos del estudio. El

método más habitual de observar los resultados es mediante la representación gráfica de la funcion de supervivencia Kaplan-Meier frente al tiempo, donde generalmente se suelen representar conjuntamente los distintos grupos a estudiar, para comparar las curvas y ver si existen diferencias. En este caso se representaron conjuntamente la supervivencia de aquellos individuos cuyo factor de transcripción estaba alterado y los que no. Finalmente para argumentar estadísticamente si existe una diferencia significativa entre dos o varias curvas de supervivencia se suele aplicar un test (*log-rank test*), de donde se obtiene un p-valor a partir del cual se determina si existe dicha diferencia significativa entre las curvas. Este test fue realizado mediante la función *survdiff* del paquete *survival*.

CAPÍTULO 4

Resultados y discusión

Como paso preliminar se quiso observar, con los datos de la expresión diferencial entre muestras pareadas sanas y enfermas, el comportamiento general de la alteración de los factores de transcripción. Para ello se realizó un *heatmap* que indicara, para todos los factores de transcripción, si éstos estaban alterados en cada cáncer (Figura 4.1), donde las celdas en gris indican que en ese cáncer el factor de transcripción no está alterado, y por el contrario, las celdas rojas indican que están sobreexpresados y las azules que están infraexpresados. En la figura 4.1 se ha indicado también el nivel de expresión de cada factor de transcripción en el tejido en el que se desarrolla cada cáncer. Las celdas con un 0 indican que no había disponible información para dicho factor de transcripción o tejido.

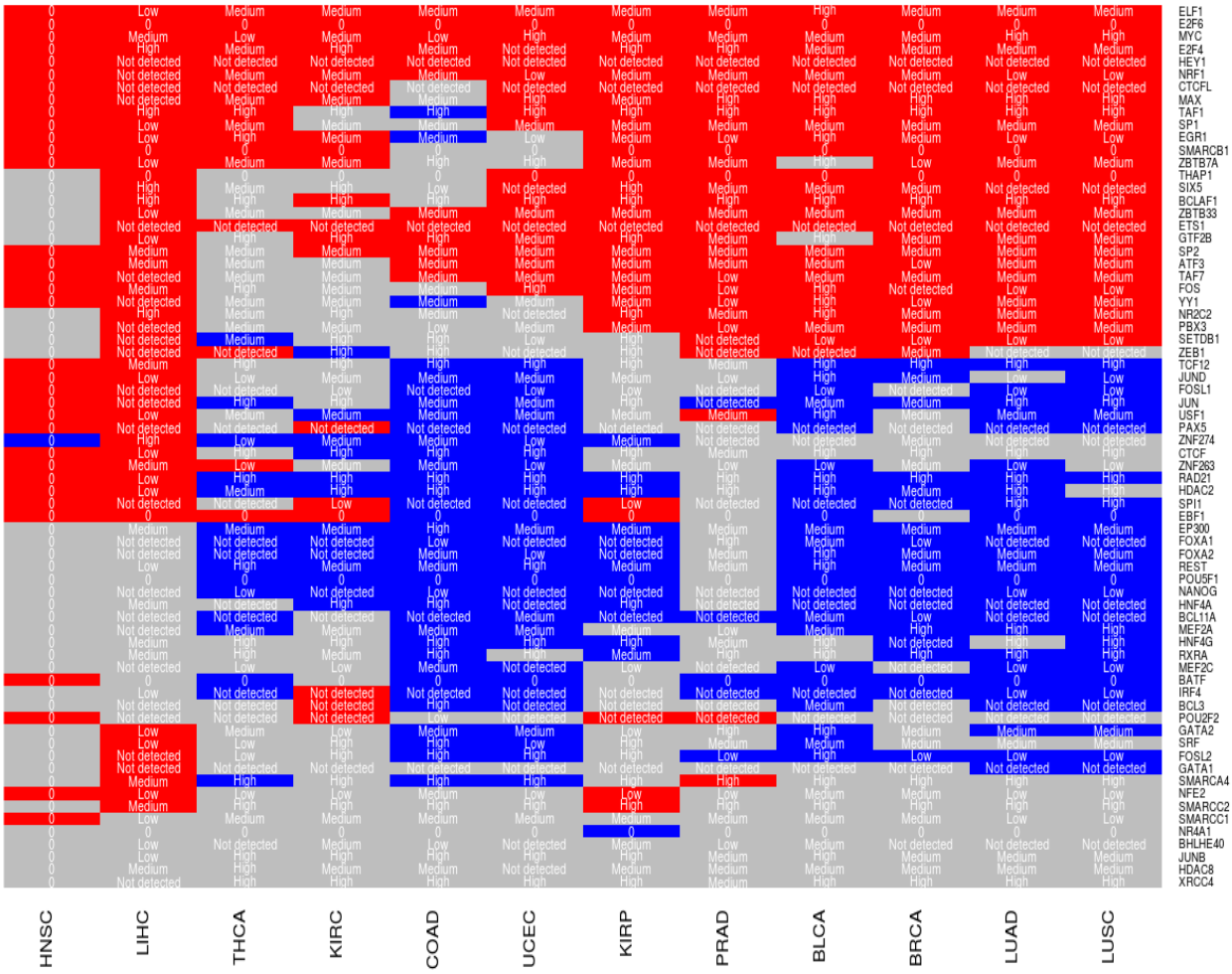

Figura 4.1. Esquema de la alteración de los factores de transcripción en los diferentes cánceres.

Como se puede apreciar en la figura 4.1, hay una serie de factores de transcripción (ELF1, E2F6, MYC, E2F4, HEY1 y NRF1) que se encuentran alterados (sobreexpresados) en todos los cánceres. Este hecho apunta a que dichos factores de transcripción son protooncogenes con un papel esencial en la carcinogénesis (proceso de producción y desarrollo de cáncer). Algunos de los factores de transcripción nombrados anteriormente tienen una relevancia contrastada en los procesos de iniciación y mantenimiento de cáncer. Como es el caso de MYC (Gabay *et al*., 2014; Wolf *et al*., 2015), o de las proteínas de la familia E2F (Nevins, 2001; Chen *et al*., 2009), cuya sobreexpresión induce una proliferación celular descontrolada debido a que son factores de transcripción ubicados al principio de rutas celulares importantes, como el control del ciclo celular.

A su vez, se puede observar que todos los cánceres estudiados tienen un elevado número de factores de transcripción alterados y que la mayoría de factores de transcripción se encuentran alterados en muchos tipos de cáncer. Esto refleja la compleja naturaleza de esta enfermedad y del delicado equilibrio de las redes de transducción de señales (Kolch *et al*., 2015), ya que la alteración de un elemento regulador suele conllevar la alteración en cascada de varios genes y otros elementos reguladores.

Junto con el *heatmap* se realizó un agrupamiento de los cánceres (*clustering*) en función de los resultados obtenidos (Figura 4.2).

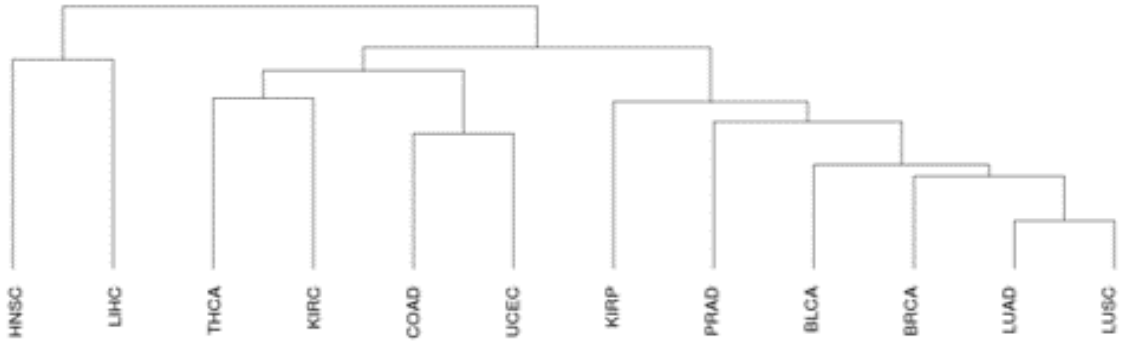

Figura 4.2. Agrupación de cánceres según la alteración de los factores de transcripción.

En esta figura se puede observar que las alteraciones de los factores de transcripción en los dos tipos de cáncer de pulmón (LUAD y LUSC) analizados son muy similares, ya que en el *cluster* se encuentran agrupados. Sin embargo, los dos tipos de cáncer de riñón analizados (KIRC y KIRP) se encuentran en partes distantes del *cluster*, lo que demuestra que, a pesar de que dos cánceres se produzcan en un órgano común, esto no implica que las rutas moleculares que están detrás de la formación de dichos cánceres sean las mismas.

4.1. Fichero de anotación

Para la obtención del archivo de anotación hubo que definir la zona promotora para posteriormente cruzar estas coordenadas con las de los TFBS y sacar los *targets*. Por este motivo, se observó primero la distancia entre las coordenadas de los TFBS y las de los genes, ambas obtenidas de Ensembl. En el 98,6% de los casos dicha distancia era de 0 pb. Este porcentaje no cambia mucho si se mira que coordenadas están a 2000 pb, ya que los casos aumentan solo en un 0.5%.

Para buscar una explicación a que la distancia entre genes y TFBS fuera de 0 pb, se realizó un análisis de la longitud de los sitios de unión. Se observó que dicha longitud, en la mayoría de los casos, era mayor de 300 nucleótidos (Figura 4.3), mucho mayor de los 10 nucleótidos esperados que suelen reconocer los factores de transcripción (Stewart *et al*., 2012).

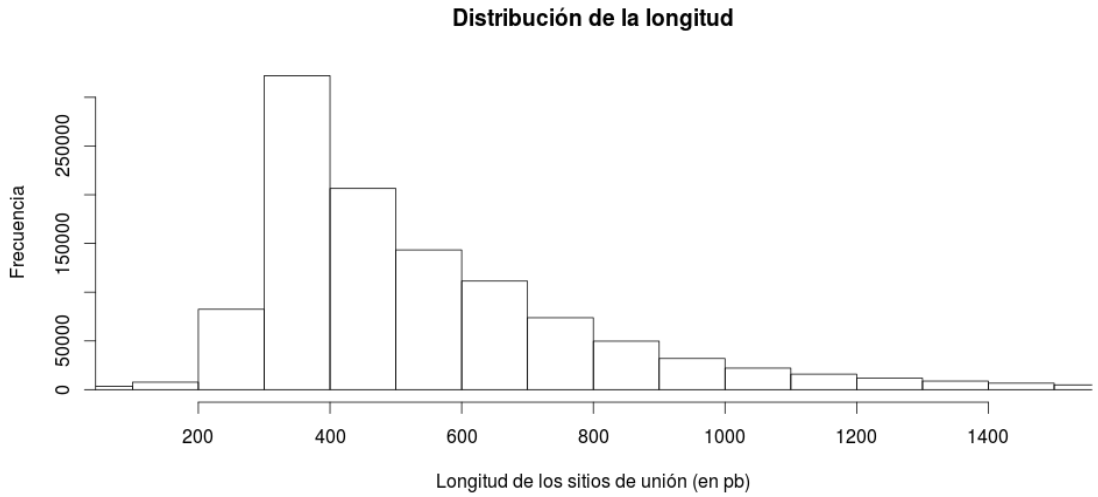

Figura 4.3. Frecuencia de la longitud de los sitios de unión de los factores de transcripción

Este hecho, se debe a que Ensembl obtiene la información de los TFBS, a partir de experimentos ChIP-seq, de fuentes como ENCODE o RoadMap Epigenomics (Ensembl, 2015). El análisis ChIP-seq consiste en un método para observar las interacciones entre proteínas y el ADN, y su nombre procede de las técnicas utilizadas para llevar a cabo dicho análisis, inmunoprecipitación de cromatina (ChIP) y secuenciación masiva de ADN paralela (seq). Como resultado final de esta técnica se obtienen ciertos picos que indican probabilidad de que una proteína se una a cierta zona de ADN, y mediante una serie de métodos y algoritmos (*peak-calling methods*) se consigue delimitar la zona de unión. En el caso de Ensemb usan dos tipos de algoritmos, SWEMBL y CCAT (Flicek *et al*., 2014; Xu *et al*., 2010), que define las coordenadas del TFBS según los picos obtenidos de los experimentos ChIP-seq. Las coordenadas obtenidas son un poco borrosas por la misma naturaleza del análisis, y dichas coordenadas no serán el TFBS en sí, si no, la zona del genoma por donde éstos se pueden encontrar.

En conclusión, debido a la longitud de los sitios de unión, y a que en el 98,6% de los casos la distancia entre TFBS y genes era de 0 pb, se decidió cruzar directamente ambas coordenadas obtenidas en el Ensembl para ver cuales intersectan y así definir las parejas

factor de transcripción-*target*. Por lo tanto, se definió como si fuese región promotora el mismo gen ya que si se hubiese definido, por ejemplo, una región promotora de 2000 Kb desde el inicio del gen solo se hubiese aumentado un 0,5% los *targets* elegidos, pero conllevaría también un descenso de la especificidad y un aumento de los falsos positivos, que en el análisis de enriquecimiento pueden influir más que los falsos negativos (Simmons *et al*., 2011).

Como resultado, se obtuvo un fichero con miles parejas de factor de transcripción-*target*, que servirán como base para los posteriores análisis. En dicho fichero hay un total de 70 factores de transcripción distintos disponibles, los cuales se nombran a continuación:

ATF3, BATF, BCL3, BCL11A, BCLAF1, BHLHE40, CTCF, CTCFL, E2F4, E2F6, EBF1, EGR1, ELF1, EP300, ETS1, FOS, FOSL1, FOSL2, FOXA1, FOXA2, GATA1, GATA2, GTF2B, HDAC2, HDAC8, HEY1, HNF4A, HNF4G, IRF4, JUN, JUNB, JUND, MAX, MEF2A, MEF2C, MYC, NANOG, NFE2, NR2C2, NR4A1, NRF1, PAX5, PBX3, POU2F2, POU5F1, RAD21, REST, RXRA, SETDB1, SIX5, SMARCA4, SMARCB1, SMARCC1, SMARCC2, SP1, SP2, SPI1, SRF, TAF1, TAF7, TCF12, THAP1, USF1, XRCC4, YY1, ZBTB7A, ZBTB33, ZEB1, ZNF263, ZNF274.

Se puede acceder al fichero de anotación obtenido en el siguiente repositorio de GitHub: https://github.com/1mati1/pan-cancer-analysis.

4.2. Análisis del estadio tumoral

Como se puede observar en la figura 3.1, el análisis del estadio tumoral consta de tres pasos. Tras la normalización de los datos se realizó la expresión diferencial cuatro veces por cada cáncer, comparando los transcriptomas del conjunto de las muestras normales con los del conjunto de muestras tumorales pertenecientes a cada estadio. Posteriormente se realizó un GSEA para cada una de las listas obtenidas y ordenadas en función del estadístico. Y por último se realizó un *heatmap* para observar los resultados (Figura 4.4), donde se agruparon las representaciones de todos los factores de transcripción.

En la figura 4.4 se puede observar el comportamiento de cada factor de transcripción en los diferentes tipos de cáncer y estadios tumorales, donde la sobreexpresión o infraexpresión de los factores de transcripción se indica en rojo y azul, respectivamente, y si el factor de transcripción no ha sido alterado con respecto al conjunto de las muestras normales se indica en gris. A su vez, en cada celda se encuentra representado el p-valor ajustado que se obtuvo en el GSEA en cada caso.

Cabe indicar que los resultados obtenidos para el BLCA en el estadio I, y el LIHC y LUSC en estadio IV no son determinantes debido al bajo número de muestras en dichos estadios (Tabla 3.1). Pero a pesar de este hecho, los datos de estos cánceres fueron utilizados ya que sí aportan información relevante en los otros estadios.

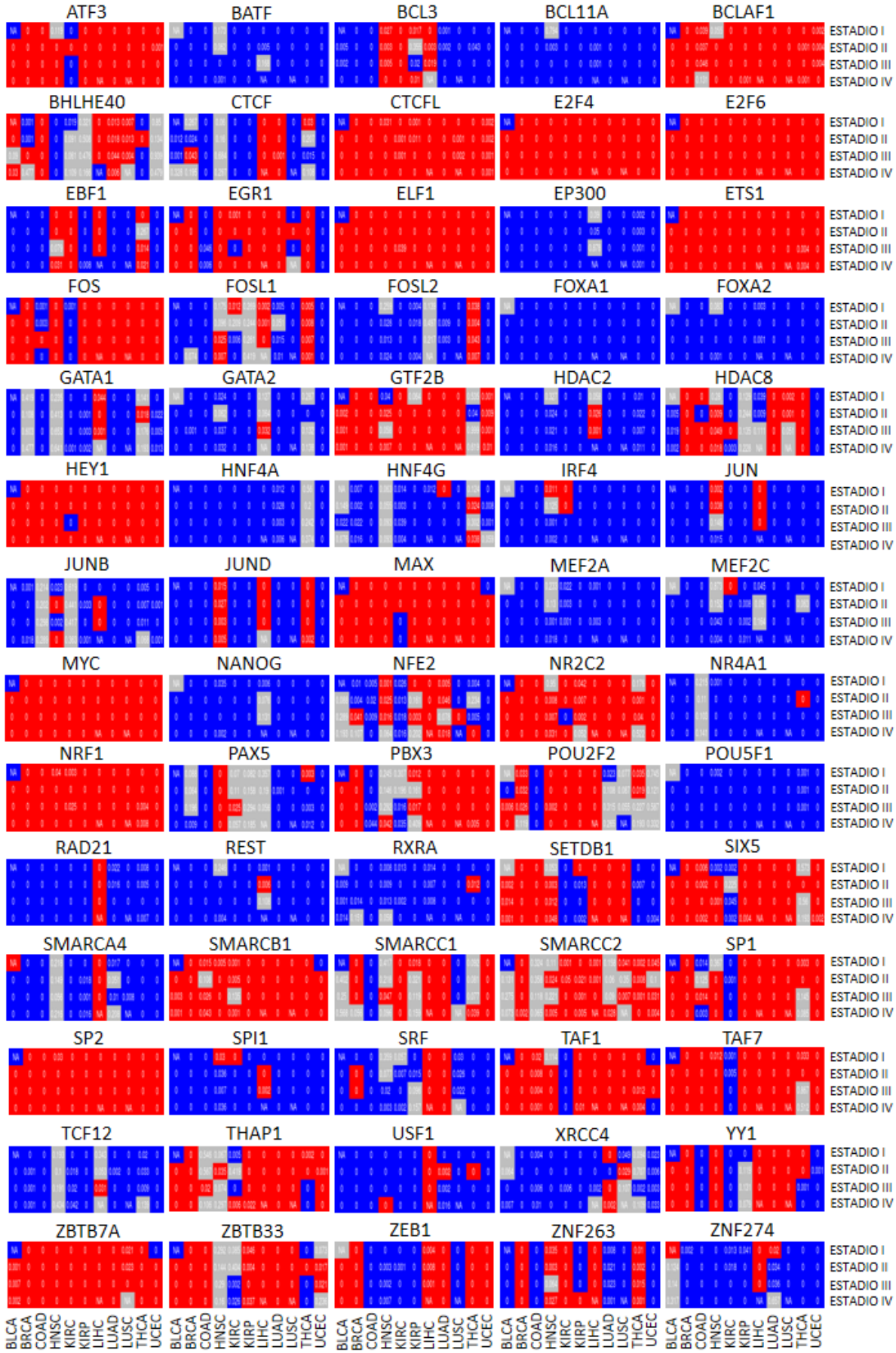

Figura 4.4. Alteración de los factores de transcripción en los distintos cánceres y estadios.

Antes de realizar el análisis de pan-cáncer se esperaba observar algún patrón común en todos los cánceres, como por ejemplo la alteración de algún factor de transcripción en un estadio concreto, que lo relacionara con la aparición de dicha fase tumoral. Aunque no se han obtenido los resultados esperados *a priori*, sí que se ha observado un patrón común en la mayoría de los cánceres con respecto al estadio tumoral, y es que cuando se altera un factor de transcripción suele estar alterado en todos los estadios. Para apoyar esta idea se realizó la figura 4.5, en la que se observa que el número de casos en los que un factor de transcripción está alterado en los cuatro estadios es mayor.

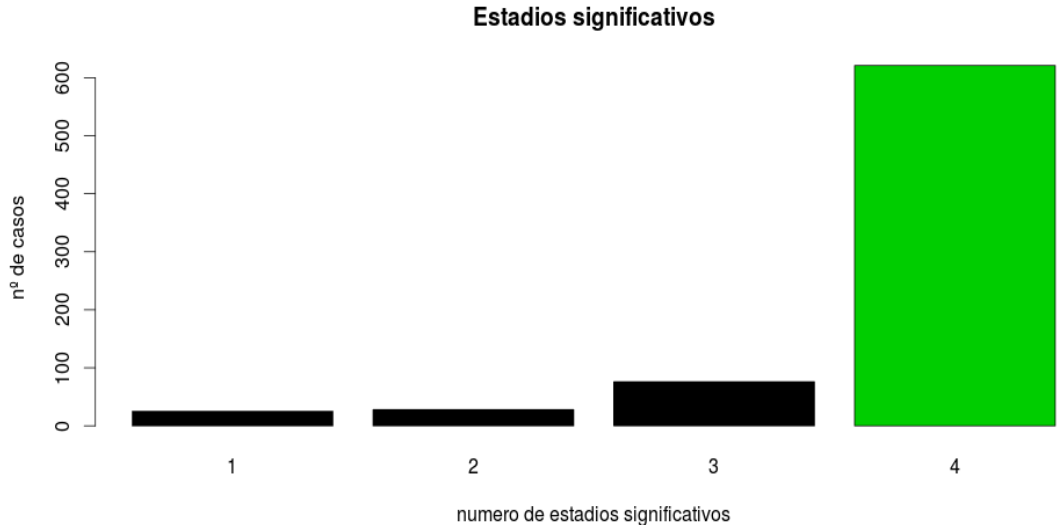

Figura 4.5. Frecuencia según el número de estadios significativos.

Además, en la figura 4.4 se observa la alteración, en todos los cánceres y todos los estadios, de los mismos factores de transcripción que en la figura 4.1 (ELF1, E2F6, MYC, E2F4, HEY1 y NRF1), lo que refuerza la hipótesis de que son protooncogenes con un papel esencial en la carcinogénesis. A esta lista de protooncogenes se les puede añadir algunos factores de transcripción más (CTCFL, ETS1, FOS, RAD21, SP2, SPI1, USF1) que también en esta figura se encuentran siempre alterados.

4.3. Análisis de supervivencia

La realización de la regresión de Cox, y la posterior mejora de cada modelo, en los distintos cánceres, mediante el algoritmo *stepwise*, reveló aquellos factores de transcripción para cada cáncer cuya alteración afectaba a la supervivencia de los pacientes (Tabla 4.1). En el cáncer de tiroides (THCA) y uno de los de riñón (KIRP) no fue posible construir un modelo de regresión, debido al bajo número de pacientes fallecidos (Tabla 3.1).

Tabla 4.1. Factores de transcripción significativos en el modelo de regresión.

	Cáncer	Factor de transcripción	P-valor	Estadístico	Coeficiente
1	BLCA	CTCF	2,39E-02	2,26	0,82
2	BLCA	RXRA	2,84E-03	-2,99	-7,79
3	BLCA	EP300	6,80E-04	-3,40	-1,41
4	BLCA	SRF	2,65E-02	2,22	0,81
5	BLCA	TCF12	1,04E-03	-3,28	-1,60
6	BLCA	GATA2	4,63E-02	1,99	0,91
7	BLCA	PAX5	4,96E-03	-2,81	-1,10
8	BLCA	TAF7	4,69E-03	-2,83	-0,96
9	BLCA	FOS	1,51E-04	-3,79	-3,71
10	BLCA	IRF4	1,86E-02	2,35	0,83
11	BLCA	SMARCA4	3,23E-03	2,95	1,11
12	BLCA	SMARCB1	2,39E-04	3,67	2,03
13	BLCA	JUNB	1,49E-02	2,44	5,73
14	BLCA	POU2F2	3,30E-02	2,13	1,27
15	BLCA	POU5F1	8,83E-03	-2,62	-0,88
16	BLCA	ZEB1	4,39E-02	-2,02	-1,65
17	BLCA	GATA1	8,08E-03	2,65	2,16
18	BLCA	NR2C2	4,66E-03	2,83	2,77
19	BRCA	JUND	5,50E-04	-3,46	-1,37
20	BRCA	EP300	1,45E-02	2,45	0,78
21	BRCA	REST	1,72E-03	3,14	1,02
22	BRCA	BCLAF1	1,72E-02	2,38	1,57
23	BRCA	PAX5	5,46E-03	-2,78	-1,00
24	BRCA	IRF4	1,32E-02	-2,48	-0,96
25	BRCA	SIX5	7,86E-03	-2,66	-1,93
26	BRCA	FOSL1	4,32E-02	-2,02	-1,30
27	BRCA	POU5F1	2,66E-04	-3,65	-0,96
28	BRCA	BCL11A	1,72E-04	3,76	1,51
29	BRCA	ZBTB33	3,30E-02	2,13	0,82
30	BRCA	HDAC8	1,51E-07	5,25	6,13
31	COAD	CTCF	1,98E-05	-4,27	-4,04
32	COAD	HNF4A	6,07E-06	-4,52	-3,77
33	COAD	JUN	6,83E-10	6,17	4,58
34	COAD	JUND	4,83E-03	-2,82	-4,31
35	COAD	ELF1	1,18E-08	-5,70	-5,96
36	COAD	REST	1,46E-03	-3,18	-2,38
37	COAD	USF1	4,67E-06	4,58	4,51
38	COAD	SRF	3,74E-04	3,56	3,13
39	COAD	EBF1	2,28E-02	2,28	0,99
40	COAD	EGR1	1,44E-04	3,80	4,27
41	COAD	TCF12	2,18E-02	-2,29	-1,79
42	COAD	BCLAF1	9,64E-06	-4,43	-3,57
43	COAD	E2F4	6,06E-08	-5,42	-5,66
44	COAD	GATA2	1,10E-02	2,54	1,75
45	COAD	HEY1	2,09E-03	3,08	1,89
46	COAD	SIX5	3,70E-02	2,09	3,01
47	COAD	SMARCB1	2,42E-08	5,58	7,11
48	COAD	POU2F2	2,78E-02	-2,20	-1,28
49	COAD	ETS1	1,37E-04	3,81	4,20
50	COAD	PBX3	1,42E-02	-2,45	-3,51
51	COAD	SP2	4,35E-03	2,85	2,80
52	COAD	BCL3	9,34E-05	3,91	3,67
53	COAD	ZBTB7A	1,12E-02	-2,54	-1,94
54	COAD	ZEB1	1,24E-02	-2,50	-2,11
55	COAD	BCL11A	3,03E-03	-2,97	-1,78
56	COAD	HNF4G	4,50E-06	-4,59	-5,48
57	COAD	ZNF263	5,20E-03	2,79	2,02
58	COAD	GATA1	6,87E-03	-2,70	-2,18
59	COAD	ZBTB33	6,26E-07	4,98	3,73

	Cáncer	Factor de transcripción	P-valor	Estadístico	Coeficiente
60	COAD	NR2C2	1,90E-02	-2,35	-1,68
61	COAD	XRCC4	1,72E-03	3,14	3,92
62	HNSC	TAF1	7,92E-03	2,66	0,68
63	HNSC	SRF	5,34E-03	-2,79	-0,83
64	HNSC	HEY1	2,27E-02	-2,28	-0,45
65	HNSC	BCL3	2,31E-02	-2,27	-0,76
66	KIRC	HNF4A	1,58E-03	3,16	0,89
67	KIRC	RXRA	1,35E-02	-2,47	-0,68
68	KIRC	FOXA2	3,04E-04	-3,61	-1,11
69	KIRC	REST	1,24E-02	-2,50	-0,67
70	KIRC	SRF	5,46E-03	2,78	0,79
71	KIRC	EBF1	2,85E-03	-2,98	-0,70
72	KIRC	BCLAF1	1,29E-02	2,49	0,81
73	KIRC	HEY1	1,15E-03	-3,25	-1,01
74	KIRC	SP1	1,72E-03	-3,14	-0,83
75	KIRC	FOSL2	3,46E-02	-2,11	-0,52
76	KIRC	GTF2B	2,62E-02	-2,22	-0,74
77	KIRC	ATF3	4,21E-05	4,10	1,25
78	KIRC	JUNB	9,44E-03	2,60	1,50
79	KIRC	POU5F1	3,19E-02	-2,15	-0,57
80	KIRC	ETS1	2,92E-03	-2,98	-1,15
81	KIRC	ZEB1	1,80E-02	2,37	0,56
82	KIRC	ZNF274	7,90E-03	2,66	0,48
83	KIRC	HNF4G	1,16E-03	3,25	0,87
84	KIRC	NANOG	1,36E-04	3,82	0,97
85	KIRC	MEF2C	4,54E-04	3,51	0,97
86	KIRC	NFE2	4,43E-02	-2,01	-0,50
87	KIRC	BHLHE40	9,57E-03	2,59	0,78
88	LIHC	CTCF	2,22E-03	3,06	1,53
89	LIHC	JUN	1,18E-03	3,24	1,74
90	LIHC	EP300	1,44E-02	2,45	3,21
91	LIHC	RAD21	4,20E-03	-2,86	-1,65
92	LIHC	ELF1	1,29E-04	3,83	1,29
93	LIHC	E2F4	2,41E-04	-3,67	-7,70
94	LIHC	SPI1	2,72E-03	3,00	2,31
95	LIHC	FOSL2	1,36E-02	-2,47	-2,96
96	LIHC	IRF4	3,46E-03	-2,92	-1,44
97	LIHC	SMARCB1	2,36E-03	-3,04	-5,17
98	LIHC	POU2F2	5,21E-06	4,56	4,80
99	LIHC	ETS1	1,66E-03	3,15	13,99
100	LIHC	SP2	1,52E-04	3,79	8,11
101	LIHC	BCL3	3,98E-03	2,88	2,67
102	LIHC	ZNF274	1,98E-02	2,33	1,10
103	LIHC	NANOG	8,01E-03	-2,65	-4,22
104	LIHC	MEF2C	5,49E-04	-3,46	-4,01
105	LIHC	NFE2	9,87E-03	2,58	1,65
106	LIHC	ZBTB33	4,70E-02	1,99	2,04
107	LUAD	EGR1	6,85E-03	2,70	0,97
108	LUAD	BCLAF1	2,01E-02	2,32	1,10
109	LUAD	MYC	3,15E-03	2,95	1,01
110	LUAD	SMARCA4	5,28E-03	2,79	1,05
111	LUAD	FOSL1	1,22E-02	-2,51	-1,06
112	LUAD	SMARCB1	3,91E-03	-2,89	-1,47
113	LUAD	ZEB1	9,88E-03	2,58	1,52
114	LUAD	ZNF274	4,24E-04	3,52	0,77
115	LUSC	CTCF	3,80E-03	2,89	0,99
116	LUSC	EP300	5,60E-03	-2,77	-0,86
117	LUSC	RAD21	3,02E-02	-2,17	-0,75
118	LUSC	FOXA2	3,48E-02	2,11	0,61
119	LUSC	REST	2,39E-02	2,26	0,80
120	LUSC	BATF	3,58E-02	-2,10	-0,43

	Cáncer	Factor de transcripción	P-valor	Estadístico	Coeficiente
121	LUSC	EGR1	1,30E-02	-2,48	-0,55
122	LUSC	TCF12	3,38E-03	2,93	0,82
123	LUSC	GATA2	3,13E-02	-2,15	-0,51
124	LUSC	PAX5	4,87E-02	1,97	0,47
125	LUSC	TAF7	2,43E-03	3,03	0,73
126	LUSC	FOSL2	9,64E-03	-2,59	-0,64
127	LUSC	ETS1	2,59E-02	2,23	0,81
128	LUSC	ZNF263	1,32E-02	-2,48	-0,94
129	UCEC	CTCF	3,30E-02	-2,13	-3,07
130	UCEC	HNF4A	3,84E-02	-2,07	-2,09
131	UCEC	MAX	4,99E-02	-1,96	-1,95
132	UCEC	JUND	3,64E-02	-2,09	-2,30
133	UCEC	FOXA2	2,74E-04	3,64	3,68
134	UCEC	SRF	1,77E-03	3,13	2,76
135	UCEC	EBF1	1,36E-02	2,47	2,78
136	UCEC	BCLAF1	2,33E-02	2,27	3,09
137	UCEC	E2F4	4,69E-04	-3,50	-3,44
138	UCEC	SPI1	2,07E-02	-2,31	-2,07
139	UCEC	FOSL2	1,12E-02	2,54	2,10
140	UCEC	IRF4	9,66E-03	-2,59	-2,54
141	UCEC	ATF3	7,28E-03	-2,68	-4,87
142	UCEC	SMARCA4	2,76E-03	-2,99	-3,25
143	UCEC	JUNB	2,73E-03	3,00	13,58
144	UCEC	POU2F2	7,88E-03	2,66	3,77
145	UCEC	PBX3	1,03E-02	2,57	6,66
146	UCEC	SP2	4,31E-02	-2,02	-17,21
147	UCEC	ZBTB7A	3,55E-02	-2,10	-4,97
148	UCEC	GATA1	2,45E-02	2,25	5,94
149	UCEC	NFE2	2,39E-04	-3,67	-7,20

Que estos factores de transcripción hayan sido significativos en cada modelo, indica que hay una relación entre la alteración del conjunto de ellos y la supervivencia en cada cáncer.

Para observar la influencia directa de un factor de transcripción determinado en la supervivencia se realizaron curvas de Kaplan-Meier con un test estadístico. La tabla 4.2 resume aquellos factores de transcripción que resultaron significativos tanto para el modelo de regresión como individualmente en las curvas de Kaplan-Meier.

Tabla 4.2. Tabla con los factores de transcripción en los dos análisis de supervivencia realizados.

Cáncer	Factores de transcripción significativos en el modelo de regresión	Total	Factores de transcripción significativos en Kaplan-Meier	Total
BLCA	CTCF, RXRA, EP300, SRF, TCF12, GATA2, PAX5, TAF7, FOS, IRF4, SMARCA4, SMARCB1, JUNB, POU2F2, POU5F1, ZEB1, GATA1, NR2C2	18		0
BRCA	JUND, EP300, REST, BCLAF1, PAX5, IRF4, SIX5, FOSL1, POU5F1, BCL11A, ZBTB33, HDAC8	12	FOSL1, HDAC8, JUND, PAX5, POU5F1	5

Cáncer	Factores de transcripción significativos en el modelo de regresión	Total	Factores de transcripción significativos en Kaplan-Meier	Total
COAD	CTCF, HNF4A, JUN, JUND, ELF1, REST, USF1, SRF, EBF1, EGR1, TCF12, BCLAF1, E2F4, GATA2, HEY1, SIX5,SMARCB1, POU2F2, ETS1, PBX3, SP2, BCL3, ZBTB7A, ZEB1, BCL11A, HNF4G, ZNF263, GATA1, ZBTB33, NR2C2, XRCC4	31	ERG1 ,ETS1, GATA2, JUN, NR2C2, SIX5, SMARCB1, SP2, SRF, USF1, ZBTB33, ZNF263	12
HNSC	TAF1, SRF, HEY1, BCL3	4		0
KIRC	HNF4A, RXRA, FOXA2, REST, SRF, EBF1, BCLAF1, HEY1, SP1, FOSL2, GTF2B, ATF3, JUNB, POU5F1, ETS1, ZEB1, ZNF274, HNF4G, NANOG, MEF2C, NFE2, BHLHE40	23	ATF3, BHLHE40, EBF1, ETS1, FOXA2, GTF2B, HNF4A, HNF4G, NANOG, REST, SRF	11
LIHC	CTCF, JUN, EP300, RAD21, ELF1, E2F4, SPI1, FOSL2, IRF4, SMARCB1, POU2F2, ETS1, SP2, BLC3, ZNF274, NANOG, MEF2C, NFE2, ZBTB33	19	ELF1, POU2F2	2
LUAD	EGR1, BCLAF1, MYC, SMARCA4, FOSL1, SMARCB1, ZEB1, ZNF274	8	ZNF274	1
LUSC	CTCF, EP300, RAD21, FOXA2, REST, BATF, EGR1, TCF12, GATA2, PAX5, TAF7, FOSL2, ETS1, ZNF263	14	TAF7	1
UCEC	CTCF, HNF4A, MAX, JUND, FOXA2, SRF, EBF1, BCLAF1, E2F4, SPI1, FOSL2, IRF4, ATF3, SMARCA4, JUNB, POU2F2, PBX3, SP2, ZBTB7A, GATA1, NFE2	21	IRF4, NFE2, POU2F2, SPI1	4

En todos los cánceres hay varios factores de transcripción que, en conjunto, condicionan la supervivencia. El cáncer de cabeza y cuello (HNSC) es el que menos factores de transcripción significativos tiene, probablemente debido a que en este cáncer se agrupan tumores que se producen en diferentes localizaciones (cavidad nasal, nasofaringe, cavidad oral, orofaringe, laringe e hipofaringe), lo que incrementa la heterogeneidad de las muestras y dificulta el análisis, no permitiendo asociar más factores de transcripción con la supervivencia.

Además se observó que el número de factores de transcripción que, por sí solos, son capaces de influir en la supervivencia es mucho menor al de factores de transcripción que influyen en conjunto, tal y como se esperaba.

A continuación, se exponen las curvas más representativas, de cada tipo de cáncer con curvas de Kaplan-Meier significativas. En ellas se observa claramente la diferencia entre la curva de aquellos pacientes que tienen el factor de transcripción alterado y los que no.

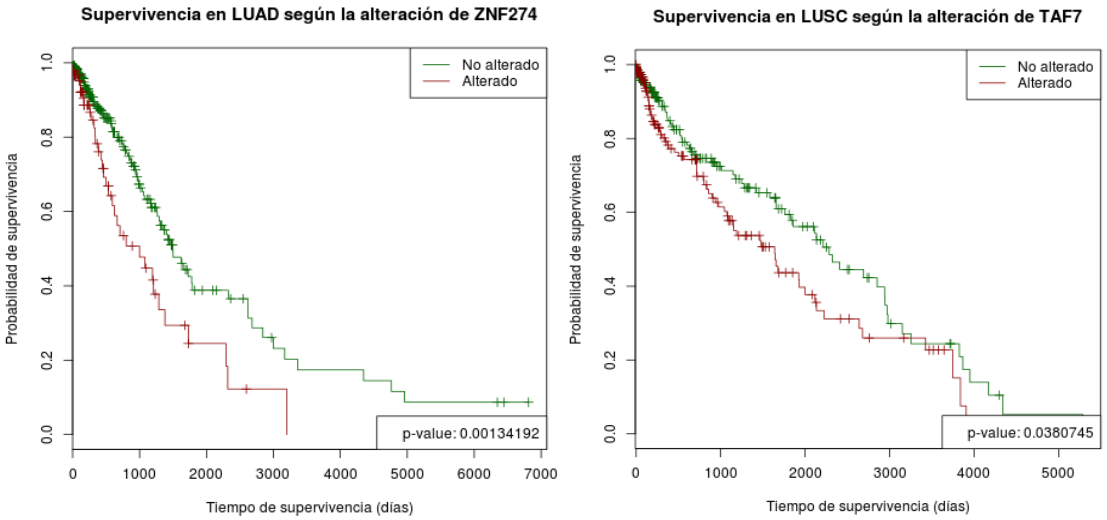

Figura 4.6. Curvas de Kaplan-Meier más significativas para LUAD y LUSC.

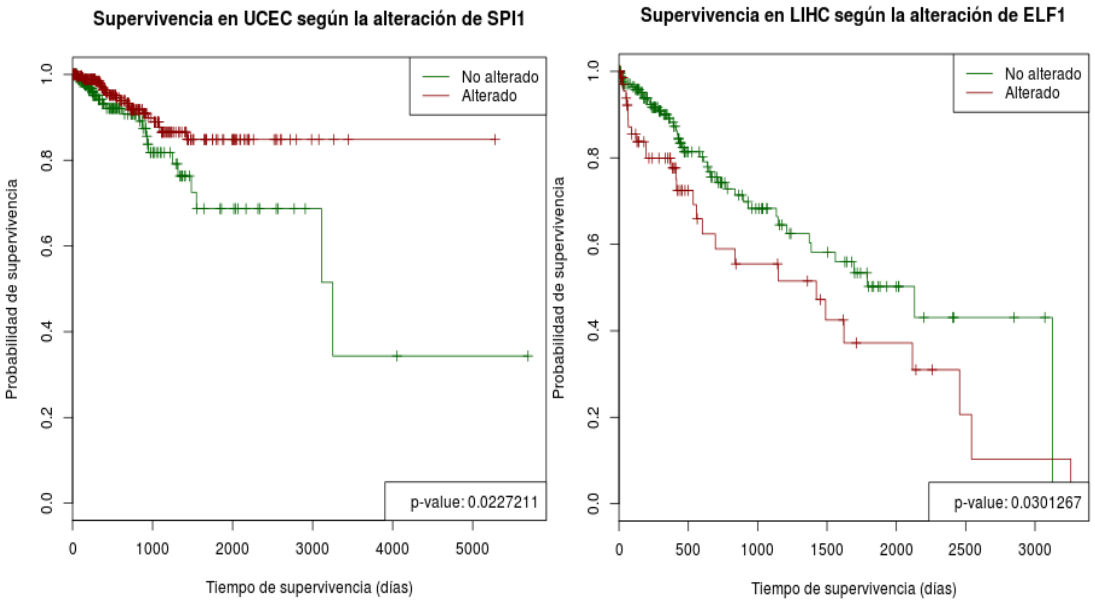

Figura 4.7. Curvas de Kaplan-Meier más significativas para UCEC y LIHC.

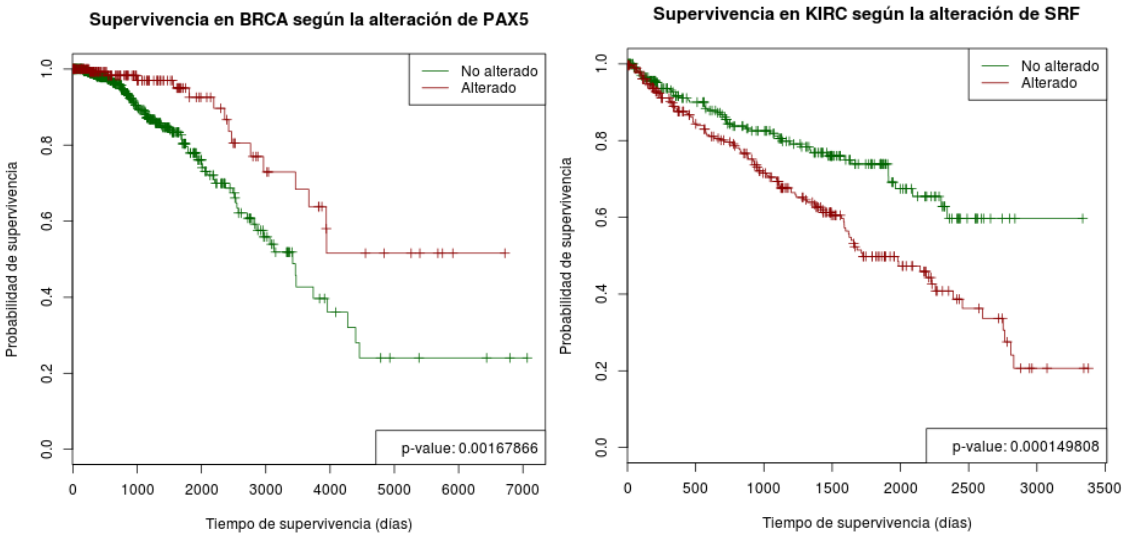

Figura 4.8. Curvas de Kaplan-Meier más significativas para BRCA y KIRC.

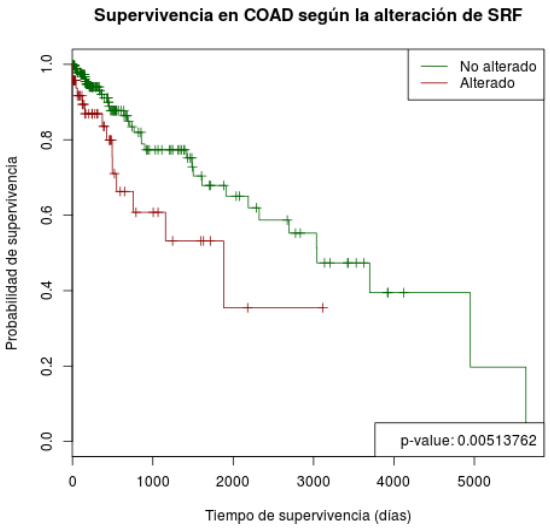

Figura 4.9. Curva de Kaplan-Meier más significativas para COAD.

Por último, cabe añadir, que no siempre hay una relación directa que implique la alteración de un factor de transcripción con una menor supervivencia de los pacientes. Por ejemplo en la figura 4.8, se puede observar que mientras que en KIRC los pacientes cuyo factor de transcripción no ha sido alterado tienen mayor supervivencia, en BRCA pasa al contrario.

Los factores de transcripción que han sido significativos en el análisis de supervivencia pueden constituir una base para la obtención de marcadores de pronósticos específicos para cada cáncer.

CAPÍTULO 5

Conclusiones

- Los factores de transcripción forman complejas redes de regulación génica que se ven alteradas en la mayoría de los cánceres.

- Hay factores de transcripción que se encuentran alterados en todos los tipos de cáncer, lo cual indica que hay procesos generales en carcinogénesis.

- No se pudo encontrar ningún factor de transcripción asociado específicamente a un estadio tumoral. En cambio, se pudo observar que la alteración de un factor de transcripción suele ser a lo largo de todos los estadios.

- Según el tipo de cáncer, hay un grupo de factores de transcripción cuya alteración influye en la supervivencia de los pacientes. También existen factores de transcripción que por sí solos son capaces de influir en la supervivencia de los pacientes.

Estas aproximaciones al conocimiento de la relación de factores de transcripción con distintas variables clínicas podrían suponer unas bases que permitan posteriormente, como futuras líneas de trabajo, la elección de marcadores de pronóstico. Por lo tanto, la determinación de los niveles de expresión de dichos marcadores podría constituir una herramienta para el diagnóstico y tratamiento personalizado de los pacientes.

Bibliografía

Adler, D., Glaser, C., Nenadic, O., Oehlschlagel, J., Zucchini, W. (2013). ff: memory-¨ efficient storage of large data on disk and fast access functions. R package version 2.2-11. http://CRAN. R-project.org/package=ff

Alonso, R., Salavert, F., Garcia-Garcia, F., Carbonell-Caballero, J., Bleda, M., Garcia-Alonso, L., Sanchis-Juan, A., Perez-Gil, D., Marin-Garcia, P., Sanchez, R., Cubuk, C., Hidalgo, M.R., Amadoz, A., Hernansaiz-Ballesteros, R.D., Alemán, A., Tarraga, J., Montaner, D., Medina, I., Dopazo, J. (2015). Babelomics 5.0: functional interpretation for new generations of genomic data. *Nucleic Acids Res.* gkv384

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Mínguez, P., Montaner, D., Dopazo, J. (2007). From genes to functional classes in the study of biological systems. *BMC Bioinformatics*. 8:114.

Altman, R.B., Miller, K.S. (2011) Translational bioinformatics year in review. *J Am Med Inform Assoc*. 18:358-366.

Balmain, A. (2001). Cancer genetics: from Boveri and Mendel to microarrays. *Nature Reviews Cancer*, 1(1), 77-82.

Byrd, D.R., Compton, C.C., Fritz, A.G., Greene, F.L., Trotti, A.I.I.I. (2010). *AJCC cancer staging manual* (Vol. 649). New York: Springer.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. (2008). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421.

Cancer Genome Atlas Research Network. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67-73.

Casado-Vela, J., Cebrián, A., Gómez Del Pulgar, M.T., Lacal, J.C. (2011). Approaches for the study of cancer: towards the integration of genomics, proteomics and metabolomics. *Clin Transl Oncol*. 13(9):617-28.

Chen, H.Z., Tsai, S.Y., Leone, G. (2009). Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nature Reviews Cancer*, *9*(11), 785-797.

Chial, H. (2008) Genetic regulation of cancer. *Nature Education* 1(1):67

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187-220.

de Jonge, E., Wijffels, J., van der Laan, J. (2015). ffbase: Basic statistical functions for package ff. R package version 0.12.1. http://CRAN. R-project.org/package=ffbase

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, 11(3), 385-388.

Edwards, B.K., Brown, M.L., Wingo, P.A., Howe, H.L., Ward, E., Ries, L.A.G., Schrag, D., Jamison, P.M., Jemal, A., Wu, X.C., Friedman, C., Harlan, L., Warren, J., Anderson, R.N., Pickle, L.W. (2005). Annual Report to the Nation on the Status of Cancer, 1975–2002, Featuring Population-Based Trends in Cancer Treatment. *J Natl Cancer Inst*, 97:1407–27.

Ensembl. Datasets and Data Processing, regulation sources [Internet] Disponible en: http://www.ensembl.org/info/genome/funcgen/regulation_sources.html. Consulta: [24-07-2015]

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F. (2012).GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet].Lyon, France: International Agency for Research on Cancer. Disponible en: http://globocan.iarc.fr. Consulta: [23-06-2015]

Filtz, T.M., Vogel, W.K., Leid, M. (2014). Regulation of transcription factor activity by interconnected post-translational modifications. *Trends in pharmacological sciences*, *35*(2), 76-85.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., García Girón, C., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri A.K., Keenan, S., Kulesha, E., Martin, F.J., Maurel, T., McLaren, W.M., Murphy, D.N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S.J., Vullo, A., Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T.J.P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D.R., Searle, S.M.J. (2013). Ensembl 2014. *Nucleic acids research*, gkt1196.

Flynn, R. (2012). Survival analysis. *Journal of clinical nursing*, *21*(19pt20), 2789-2797.

Gabay, M., Li, Y., Felsher, D.W. (2014). MYC activation is a hallmark of cancer initiation and maintenance. *Cold Spring Harbor perspectives in medicine*, *4*(6), a014241.

Gentzsch, W. (2001). Sun grid engine: Towards creating a compute power grid. In *Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symposium on* (pp. 35-36). IEEE.

Hanahan, D., Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646-674.

He, L., Wennerberg, K., Aittokallio, T., Tang, J. (2015). TIMMA-R: an R package for predicting synergistic multi-targeted drug combinations in cancer cell lines or patient-derived samples. *Bioinformatics*, 31(11), 1866-1868.

Ihaka, R., Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statics*, 5 (3): 299-314.

Instituto de Biología Molecular y Celular del Cáncer (IBMCC) del Centro de Investigación del Cáncer (CIC). Factores de transcripción [Internet] Disponible en: http://www.cicancer.org/es/factores-de-transcripcion . Consulta: [17-07-2015]

International Cancer Genome Consortium (ICGC). Consortium goals [Internet] Disponible en: https://icgc.org/icgc/goals-structure-policies-guidelines/b-consortium-goals . Consulta: [18-07-2015]

Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457-481.

Klug, W. S., Cummings, M. R., Spencer, C. A. (2006) Conceptos de genética.

Kolch, W., Halasz, M., Granovskaya, M., Kholodenko, B.N. (2015). The dynamic control of signal transduction networks in cancer cells. *Nature Reviews Cancer*, *15*(9), 515-527.

Law, C.W., Chen, Y., Shi, W., Smyth, G.K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, *15*(2), R29.

Luscombe, N.M., Greenbaum, D., Gerstein, M. (2001). What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics*, 1, 83-99.

Marioni, J. C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, *18*(9), 1509-1517.

McDermott, U., Settleman, J. (2009). Personalized cancer therapy with selective kinase inhibitors: an emerging paradigm in medical oncology. *Journal of Clinical Oncology*, *27*(33), 5650-5659.

Montaner, D., Dopazo, J. (2010). Multidimensional gene set analysis of genomic data. *PLoS One*, 5(4), e10348.

NCBI Resource Coordinators. (2013). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, *41*(Database issue), D8.

Nevins, J.R. (2001). The Rb/E2F pathway and cancer. *Human molecular genetics*, *10*(7), 699-703.

Quinlan, A.R., Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842.

Racine, J. S. (2012). RStudio: A Platform-Independent IDE for R and Sweave. *Journal of Applied Econometrics*, 27(1), 167-172.

Reddy, E. P., Reynolds, R. K., Santos, E., Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*, 300, 149-152.

Reichhardt, T. (1999). It's sink or swim as a tidal wave of data approaches. *Nature*, 399, 517-520

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

Robinson, M.D., McCarthy, D.J., Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140.

Robinson, M.D., Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, *11*(3), R25.

Simmons, J.P., Nelson, L.D., Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 0956797611417632.

Spedicato, G. A. (2013). The lifecontingencies Package: Performing Financial and Actuarial Mathematics Calculations in R. *Journal of Statistical Software*, 55(10), 1-36.

Stewart, A.J., Hannenhalli, S., Plotkin, J.B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics*, *192*(3), 973-985.

Stewart, B.W., Wild, C.P., editors (2014).World Cancer Report 2014. Lyon, France: International Agency for Research on Cancer.

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A., Shefler, E., Ramos, A.H., Stojanov, P., Carter, S.L., Voet, D., Cortés, M.L., Auclair, D., Berger, M.F., Saksena, G., Guiducci, C., Onofrio, R.C., Parkin, M., Romkes, M., Weissfeld, J.L., Seethala, R.R., Wang, L., Rangel-Escareño, C., Fernandez-Lopez, J.C., Hidalgo-Miranda, A., Melendez-Zajgla, J., Winckler, W., Ardlie, K., Gabriel, S.B., Meyerson, M., Lander, E.S., Getz, G., Golub, T.R., Garraway, L.A., Grandis, J.R. (2011). The mutational landscape of head and neck squamous cell carcinoma. *Science*, *333*(6046), 1157-1160.

Schacht, T., Oswald, M., Eils, R., Eichmüller, S.B., König, R. (2014). Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics*, *30*(17), 401-407.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545-15550.

Tang, Y., He, W., Wei, Y., Qu, Z., Zeng, J., Quin, C. (2013). Screening key genes and pathways in glioma based on gene set enrichment analysis and meta-analysis. *Journal of Molecular Neuroscience*, 50 (2): 324-332.

The Cancer Genome Atlas (TCGA) del National Cancer Institute (NCI). Home page [Internet] Disponible en: http://cancergenome.nih.gov/ . Consulta: [16-07-2015]

Therneau, T. (2014): A Package for Survival Analysis in S, R package version 2.37-7.

http://cran.r-project.org/web/packages/survival

Tootle, T.L., Rebay, I. (2005). Post-translational modifications influence transcription factor activity: A view from the ETS superfamily. *Bioessays*, *27*(3), 285-298.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65: 87–108. doi: 10.3322/caac.21262

van Rossum, G.(2003) The Python Language Reference Manual. *Network Theory Ltd*.

Wall, L., Christiansen, T., Orwant, J. (2004). Programming perl. *O'Reilly Media, Inc.*

Wang, N.J., Sanborn, Z., Arnett, K.L., Bayston, L.J., Liao, W., Proby, C.M., Leighe, I.M., Collisson, E.A., Gordong, P.B., Jakkula, L., Pennypacker, S., Zou, Y., Sharma, M., North, J.P., Vemula, S.S., Mauro, T.M., Neuhaus, I.M., LeBoit, P.E., Hur, J.S., Park, K., Huh, N., Kwok, P., Arron, S.T., Massion, P.P., Bale, A.E., Haussler, D., Cleaver, J.E., Gray, J.W., Spellman, P.T., South, A.P., Aster, J.C., Blacklow, S.C., Cho, R.J. (2011). Loss-of-function mutations in Notch receptors in cutaneous and lung squamous cell carcinoma. *Proceedings of the National Academy of Sciences*, *108*(43), 17761-17766.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Cancer Genome Atlas Research Network. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10), 1113-1120.

Weng, A.P., Ferrando, A.A., Lee, W., Morris, J.P., Silverman, L.B., Sanchez-Irizarry, C., Blacklow, S.C., Look A.T., Aster, J.C. (2004). Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science*, *306*(5694), 269-271.

WHO. World Health Organization. [Internet] Disponible en: http://www.who.int/mediacentre/factsheets/fs297/en/. Consulta: [22-06-2015]

Wolf, E., Lin, C.Y., Eilers, M., Levens, D.L. (2015). Taming of the beast: shaping Myc-dependent amplification. *Trends in cell biology*, *25*(4), 241-248.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C. L., Lin, F., Sung, W. K. (2010). A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, *26*(9), 1199-1204.

Zagouras, P., Stifani, S., Blaumueller, C.M., Carcangiu, M.L., Artavanis-Tsakonas, S. (1995). Alterations in Notch signaling in neoplastic lesions of the human cervix. *Proceedings of the National Academy of Sciences*, *92*(14), 6414-6418.