Estudio de la heterogeneidad regulatoria en cáncer y sus implicaciones en la medicina personalizada

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Departamento de Biotecnología
Universitat Politècnica de València

Matías Marín Falco

DIRECTOR:
Dr. Joaquín Dopazo Blázquez

Octubre 2020

Agradecimientos

Esta aventura que empezó en Valencia, siguió en Sevilla y no se donde continuará, me ha llevado a conocer a gente maravillosa por el camino y a compartir buenos momentos por los que me gustaría agradecer.

En primer lugar, quería agradecer a Ximo por confiar en mi, apoyarme y motivarme para seguir la carrera investigadora que tanto cuesta lamentablemente en nuestro país. Además, has conseguido formar 2 grupos enormes, no solo en lo profesional sino también en lo humano, lo cual es de admirar.

Gracias a todos los compañeros de Valencia por esos desayunos, la mesa de las comidas, las charlas y viajes que hicieron que día tras día fuese con ganas a trabajar. Muchas gracias especialmente a Jose, por tenerme paciencia y ser mi mentor en mis primeros pasos en la bioinformática

Me fui a Sevilla con bastante pena de dejar al gran grupo de personas que había en Valencia. Mi gran sorpresa ha sido volver a encontrarme con otro grupo igual de bueno con el que también he disfrutado de las comidas y viajes que me han hecho disfrutar esta aventura a pesar de estar lejos de casa. La parte mala es que me voy con la misma pena de Sevilla con la que llegue, lo bueno es que me vuelvo con nuevos buenos amigos.

Muchas gracias a toda mi familia, por su apoyo y amor que me hacen llegar desde la distancia. A mi madre, una de las personas más buenas y justas que conozco, gracias por el apoyo, por haberme inculcado desde pequeño la curiosidad de un buen investigador y los valores de una buena persona. Siempre tendrás mi amor y admiración.

Por último y más importante, a Bea, te estaré eternamente agradecido por haberte cruzado en mi camino, por llenar mis días con gatos, amor y felicidad y por haberme acompañado en esta aventura sin dudar. Estoy feliz porque se que aún nos quedan por delante muchas experiencias, viajes, comidas y amor por compartir.

Resumen

El cáncer es la segunda causa de muerte en el mundo y se caracteriza principalmente por la proliferación descontrolada de las células que forman el tumor. Aunque el desarrollo de un tumor es posible debido a ciertos procesos comunes desencadenados por la desregulación del equilibrio existente entre los componentes moleculares de una célula y sus elementos de control, existe una gran heterogeneidad en los mecanismos a través de los cuales ocurre dicha desregulación. Gracias al desarrollo de nuevas tecnologías de secuenciación ha sido posible observar como esta heterogeneidad no solo se observa entre los distintos tipos de tumores sino entre las propias células de un mismo tumor.

La caracterización de la heterogeneidad tumoral ha tenido un gran impacto en la comprensión de la enfermedad y el desarrollo de nuevas terapias dirigidas. Por este motivo, con el fin de mejorar la caracterización de alteraciones en los distintos mecanismos regulatorios, en esta tesis se han desarrollado dos metodologías con gran potencial para su aplicación en la medicina personalizada y que permiten estudiar la heterogeneidad inter e intratumoral de los estados de activación de elementos reguladores.

En primer lugar, se desarrolló una metodología que permite determinar en una muestra el estado de activación de los factores de transcripción (FTs) a partir de la expresión de los genes a los que regula. Se aplicó la

metodología para realizar un análisis sistemático de varios cánceres (conocido como estudios pan-cáncer) en el que se caracterizó por primera vez el escenario regulatorio de 52 FTs en 11 tipos de cáncer distintos. Además, al poder obtener valores de activación individuales para cada muestra, fue posible observar correlaciones entre la activación de algunos FTs con la supervivencia, sugiriendo así su uso como marcadores pronósticos.

En segundo lugar, se desarrolló otra metodología en la que se emplea un modelo mecanístico para determinar el estado de activación de alrededor de 1000 circuitos de señalización a partir de datos de experimentos transcriptómicos de células únicas (scRNAseq). El uso de este modelo mecanístico en datos de scRNAseq de 4 pacientes de glioblastoma, además de mostrar la heterogeneidad intratumoral presente en las muestras, ha permitido realizar intervenciones *in silico* para simular el efecto de distintas drogas sobre las células. De esta manera, ha sido posible describir posibles mecanismos mediante los cuales un grupo de células pueden evitar el efecto de una terapia dirigida.

Las metodologías desarrolladas en esta tesis, así como los resultados obtenidos tras su aplicación supone una valiosa fuente de información para el desarrollo de marcadores de diagnóstico, pronóstico y respuesta que ayuden a entender mejor los distintos niveles de heterogeneidad presentes en cáncer, y así, poder aumentar la eficacia de las terapias dirigidas.

Abstract

Cancer is the second leading cause of death in the world and is characterized mainly by the uncontrolled proliferation of the cells that make up the tumor. Although the development of a tumor is possible due to certain common processes triggered by the dysregulation of the existing balance between the molecular components of a cell and its control elements, there is great heterogeneity in the mechanisms through which this dysregulation is achieved. Thanks to the development of new sequencing technologies, it has been possible to observe how this heterogeneity is not only observed between the different types of tumors but also between the cells of the same tumor.

The characterization of tumor heterogeneity has had a great impact on the understanding of the disease and the development of new targeted therapies. For this reason, in order to improve the characterization of alterations in the different regulatory mechanisms, in this thesis two methodologies have been developed that allow studying the inter- and intratumoral heterogeneity of the activation states of regulatory elements and with great potential for their application in personalized medicine.

In the first place, a methodology that allows determining in a sample the activation state of the transcription factors (FTs) from the expression of the genes that it regulates was developed. The methodology was applied

to perform a pan-cancer analysis in which the regulatory scenario of 52 FTs was characterized for the first time in 11 different types of cancer. Furthermore, by being able to obtain individual activation values for each sample, it was possible to observe correlations between the activation of some FTs with survival, thus suggesting their use as prognostic markers.

Second, another methodology was developed using a mechanistic model to determine the activation state of around 1000 signaling circuits in single cell transcriptomic experiments (scRNAseq). The use of this mechanistic model in scRNAseq data from 4 glioblastoma patients, in addition to showing the intratumoral heterogeneity present in the samples, has allowed *in silico* interventions to simulate the effect of different drugs on cells. In this way, it has been possible to describe possible mechanisms by which a group of cells can avoid the effect of a targeted therapy.

The methodologies developed in this thesis, as well as the results obtained after its application, is a valuable source of information for the development of diagnostic, prognostic and response markers that help to better understand the different levels of heterogeneity present in cancer, and thus, be able increase the effectiveness of targeted therapies.

Resum

El càncer és la segona causa de mort al món i es caracteritza principalment per la proliferació descontrolada de les cèl·lules que formen el tumor. Encara que el desenvolupament d'un tumor és possible a causa de certs processos comuns desencadenats per la desregulació de l'equilibri existent entre els components moleculars d'una cèl·lula i els seus elements de control, hi ha una gran heterogeneïtat en els mecanismes a través dels quals s'aconseguix aquesta desregulació. Gràcies a el desenvolupament de noves tecnologies de seqüenciació ha sigut possible observar com aquesta heterogeneïtat no només s'observa entre els diferents tipus de tumors sinó entre les pròpies cèl·lules d'un mateix tumor.

La caracterització de l'heterogeneïtat tumoral ha tingut un gran impacte en la comprensió de la malaltia i el desenvolupament de noves teràpies dirigides. Per aquest motiu, per tal de millorar la caracterització d'alteracions en els diferents mecanismes reguladors, en aquesta tesi s'han desenvolupat dues metodologies amb gran potencial per a la seua aplicació en la medicina personalitzada i que permeten estudiar l'heterogeneïtat inter i intratumoral dels estats de activació d'elements reguladors.

En primer lloc es va desenvolupar una metodologia que permet determinar en una mostra l'estat d'activació dels factors de transcripció

(FTs) a partir de l'expressió dels gens als que regula. Es va aplicar la metodologia per a realitzar una anàlisi de pan-cancer en el qual es va caracteritzar per primera vegada l'escenari regulatori de 52 FTs a 11 tipus de càncer diferents. A més, al poder obtenir valors d'activació individuals per a cada mostra, va ser possible observar correlacions entre l'activació d'alguns FTs amb la supervivència, suggerint així el seu ús com a marcadors pronòstics.

En segon lloc, es va desenvolupar una altra metodologia en la qual s'empra un model mecanístic per determinar l'estat d'activació d'al voltant de 1000 circuits de senyalització a partir d'experiments transcriptòmics de cèl·lules úniques (scRNAseq). L'ús d'aquest model mecanístic en dades de scRNAseq de 4 pacients de glioblastoma, a més de mostrar l'heterogeneïtat intratumoral present en les mostres, ha permès realitzar intervencions *in silico* per simular l'efecte de diferents drogues sobre les cèl·lules. D'aquesta manera, ha estat possible descriure possibles mecanismes mitjançant els quals un grup de cèl·lules poden evitar l'efecte d'una teràpia dirigida.

Les metodologies desenvolupades en aquesta tesi, així com els resultats obtinguts després de la seva aplicació suposa una valuosa font d'informació per al desenvolupament de marcadors de diagnòstic, pronòstic i resposta que ajudin a entendre millor els diferents nivells d'heterogeneïtat presents en càncer, i així, poder augmentar l'eficàcia de les teràpies dirigides.

Indice

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1. INTRODUCCIÓN

1.1 CÁNCER

Cáncer es el término que comúnmente se emplea para denominar a un conjunto de enfermedades, que pueden afectar a distintos órganos del cuerpo, en las que se observa un patrón de división celular descontrolado. A pesar de tratarse de distintas enfermedades con distintas etiologías, se ha establecido una serie de clasificaciones comunes para facilitar su estudio y tratamiento clínico. En general, los distintos tipos de tumores se clasifican en función del órgano o tejido en el cual se desarrollan inicialmente. Por otro lado, el estadio de un cáncer sirve para describir el avance del tumor. El sistema de estadificación de cáncer más empleado en la práctica clínica es el determinado por el TNM. Las siglas TNM indican los parámetros que se tienen en cuenta en este sistema, donde T se refiere al tamaño y extensión del tumor principal, la N indica si el tumor se ha infiltrado en los ganglios (o nódulos) linfáticos cercanos y la M si ha habido metástasis, es decir, cuando las células tumorales se han diseminado a otras partes del cuerpo. Se otorga un valor para cada uno de los parámetros mencionados anteriormente, lo que ayuda a describir el avance del tumor de forma detallada, aunque las combinaciones TNM se suelen agrupar en cinco estadios. Se clasifican en el estadio 0 aquellos tumores que no se han diseminado al tejido adyacente, también conocido como carcinoma in situ. En los estadios I, II y III se engloban aquellos tumores que han invadido el tejido cercano, siendo mayor el estadio cuanto mayor sea la extensión que abarcan. Una vez se observa metástasis en órganos distales del tumor, se clasifican en el estadio IV.

Actualmente el cáncer es la segunda causa de muerte en el mundo, por detrás solo de enfermedades cardiovasculares, con alrededor de 17 millones de nuevos casos y 9.5 millones de muertes al año (WHO, 2018).

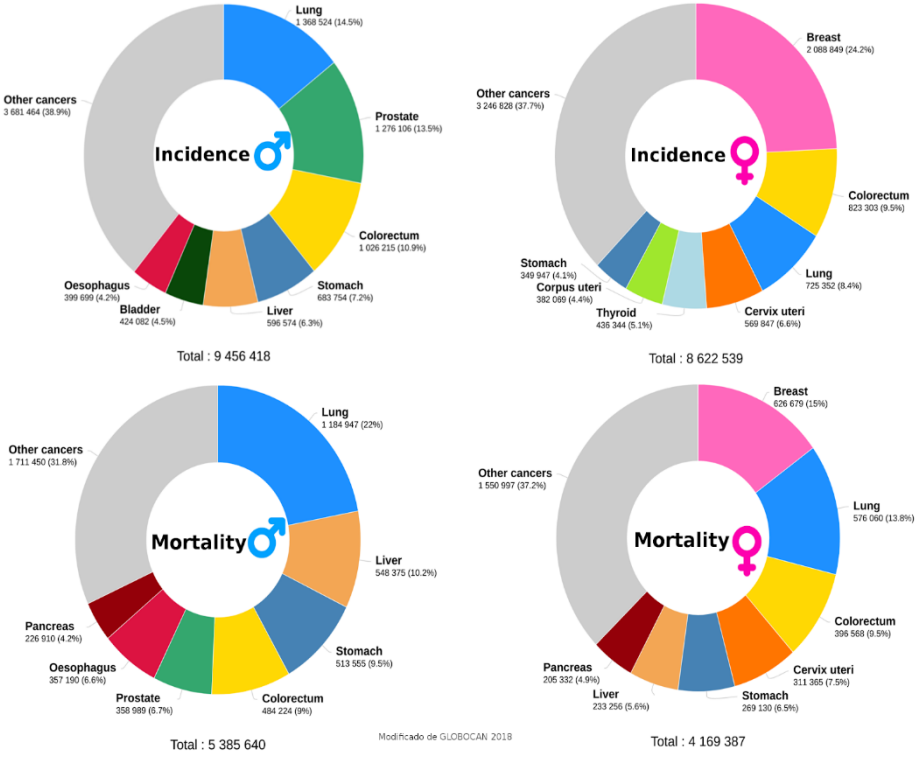


Figura 1.1. Incidencia y mortalidad de cáncer por sexo y órgano. Datos mundiales de nuevos casos diagnosticados (parte superior) y fallecimientos (parte inferior) anualmente, separado por zona de aparición del tumor, en hombres (parte izquierda) y mujeres (parte derecha).

Globalmente, el cáncer de pulmón, de mama (en mujeres) y el colorrectal explican en conjunto un tercio de la incidencia y la mortalidad por cáncer (Bray et al., 2018; Ferlay et al., 2018). En hombres, el cáncer de pulmón es el tipo de tumor más diagnosticado (14.5%) y con mayor letalidad (22%), seguido en incidencia por el cáncer de próstata (13.5%),

colorrectal (10.9%) y estómago (7.2%) y en tasa de mortalidad el cáncer de hígado (10.2%), de estómago (9.5%) y colorrectal (9%). En mujeres, en cambio, es el cáncer de mama el tipo de tumor con más incidencia (24.2%), seguido por el colorrectal (9.5%), de pulmón (8.4%) y de útero (6.6%). Entre los tipos de cáncer con mayor mortalidad en mujeres se encuentran el de mama (15%), seguido por el cáncer de pulmón (13.8%), el colorrectal (9.5%) y el de útero (7.5%) (Figura 1.1).

El número total de muertes por cáncer, así como la prevalencia de los diferentes tipos de cáncer, varían según países, e incluso regiones, debido al importante impacto que tienen sobre la incidencia factores sociodemográficos, como el desarrollo económico y acceso a una sanidad de calidad, factores ambientales y del entorno, el estilo de vida y por último, factores de predisposición poblacional. Si bien el desencadenante de un tumor a nivel molecular, en última instancia, suele ser una o varias alteraciones genéticas, existen varios factores de riesgo que influyen en el riesgo a padecer cáncer. Muchos de esos factores de riesgo son intrínsecos al individuo, como es el caso de la edad, uno de los factores más importantes del cáncer esporádico; los antecedentes familiares, que determinan la probabilidad de padecer cáncer hereditario, como en el caso de familias portadoras de mutaciones en los genes *BRCA1* y *BRCA2*, asociadas con el síndrome hereditario de cáncer de mama y de ovario (HBOC, del inglés hereditary breast and ovary cancer) (Levy-Lahad y Friedman, 2007); o los niveles hormonales, cuyo papel es importante también en el cáncer de mama (Stefanick et al., 2006), entre otros, aunque estos niveles pueden ser modificados por factores externos. La exposición a otros factores de riesgo puede ser prevenible o atenuada, ya que son más dependientes de factores ambientales, relacionados con la cultura o el estilo de vida, como la exposición a radiaciones, la dieta, consumo de

alcohol y tabaco, obesidad o algunas infecciones. Algunos estudios indican que entre un tercio y dos quintos de los nuevos casos de cáncer pueden ser evitados si se eliminan o reducen ciertos factores de riesgos ambientales o ligados al estilo de vida (Brown et al., 2018; Wilson et al., 2018)

1.1.1 REGULACIÓN EN CÁNCER

La diversidad de tipos celulares en un organismo pluricelular, depende en gran medida de los niveles de expresión de genes, ya que prácticamente todas las células del organismo contienen la misma información genética. El perfecto funcionamiento celular depende de un equilibrio entre la expresión de ciertos genes, varios mecanismos moleculares y elementos regulatorios, que a su vez le permite adaptarse a cambios, no solo intracelulares sino también en el ambiente. Cuando este equilibrio (homeostasis) se altera o la célula no es capaz de adaptarse correctamente también existen mecanismos para destruir o invalidar a la célula (p.e. apoptosis o senescencia) y que estos errores no proliferen. El cáncer representa una pérdida de este equilibrio regulatorio y, en consecuencia, la actividad proliferativa se encuentra descontrolada. Esta desregulación suele originarse por la alteración de genes con una función importante en el ciclo celular y que se suelen clasificar en dos categorías: protooncogenes y genes supresores de tumor. Los protooncogenes, mediante mutaciones que producen una ganancia de función, sirven de aceleradores para activar el ciclo celular, mientras que los genes supresores de tumor, mediante mutaciones que producen una pérdida de función, pierden su capacidad de actuar normalmente como frenos para ralentizar el crecimiento celular y prevenir la proliferación indefinida (Chial, 2008). La primera mutación somática descrita en un protooncogen,

fue descubierta en el gen *HRAS* (Reddy et al., 1982). Se han descrito varios mecanismos regulatorios, como factores de transcripción, RNA no codificante, modificación de la cromatina, rutas de señalización, entre otros, que actúan en distintos niveles de la maquinaria celular y cuya modificación podría desencadenar ciertas patologías como el cáncer.

1.1.1.1 Factores de transcripción

Los factores de transcripción juegan un papel muy importante en la maquinaria regulatoria. Los factores de transcripción son proteínas que regulan la expresión de ciertos genes (genes diana) mediante la unión a regiones promotoras o potenciadoras (enhancers) y reclutando la maquinaria necesaria (cofactores y RNA Pol II) para la transcripción de dichos genes (Spitz y Furlong, 2012). Estas proteínas emplean dos superficies de interacción distintas: un dominio de unión a una secuencia específica de ADN (p.e. dedos de zinc, homeodominio o estructuras hélice-bucle-hélice) y un dominio de activación/represión que interactúa con varios cofactores (Bhagwat y Vakoc, 2015).

Los enhancers son unas regiones cortas (50-1500 bp) del DNA a las que se pueden unir ciertas proteínas (factores de transcripción) para aumentar la probabilidad de que la transcripción de un determinado gen ocurra. Pueden estar localizados a varias kilobases del lugar de inicio de la transcripción, aunque luego espacialmente dentro del núcleo de la célula, debido a la conformación espacial de la cromatina, se encuentren más cerca del gen al que regulan. Se pueden encontrar estas regiones tanto en células eucariotas como procariotas, y son regiones ciertamente conservadas. A lo largo de nuestro genoma hay cientos de miles de enhancers (Pennacchio et al., 2013). La inestabilidad en el genoma es una

característica del cáncer, o hallmark (ver apartado 1.1.2), que contribuye a la acumulacion de alteraciones en la secuencia de estas regiones reguladoras, entre otras, y que puede promover la progresión del tumor. Varios estudios han detectado asociaciones entre mutaciones en regiones regulatorias del DNA con distintos tipos de cáncer: cáncer de mama (Jiang et al., 2011), cáncer de próstata (Demichelis et al., 2012; Wasserman et al., 2010), cáncer colorrectal (Lubbe et al., 2012; Pomerantz et al., 2009), cáncer de riñón (Schödel et al., 2012), cáncer de pulmón (Liu et al., 2011), cáncer nasofaríngeo (Yew et al., 2012) y melanoma (Huang et al., 2013; Horn et al., 2013).

Según su papel en la regulación de la transcripción, los factores de transcripción se pueden separar en dos clases: los de control de la iniciación y los de control de la elongación (Adelman y Lis, 2012; Zhou et al., 2012). Aunque esta distinción no es absoluta ya que algunos factores de transcripción pueden contribuir tanto en la etapa de iniciación de la transcripción como en la de elongación. Principalmente su contribución a la transcripción se debe a la atracción de los cofactores al núcleo de elementos de la maquinaria de transcripción (core) (Fuda et al., 2009). Los cofactores son complejos proteicos que contribuyen a la activación y represión de la transcripción pero que no tienen propiedades de unión al DNA por ellos mismos. Es por esto que dependen de los factores de transcripción para que los acerquen al lugar indicado del core transcripcional para que puedan ejercer su función. El genoma humano codifica cerca de 2000 factores de transcripción distintos, muchos de los cuales son expresados en un tipo celular específico para coordinar el programa de expresión génica del que dependen varios procesos celulares vitales (Bhagwat y Vakoc, 2015). A esto hay que sumarle que un factor de transcripción puede llegar a regular la transcripción de hasta cientos de genes (Shalgi et al., 2007). Es por esto que los factores de transcripción

forman una parte muy importante en la maquinaria regulatoria y que una alteración en los mismos puede tener amplios efectos en la homeóstasis celular. Se ha estudiado ampliamente la contribución de ciertos factores de transcripción a la tumorigénesis, y como la sobreexpresión de factores de transcripción oncogénicos pueden alterar el circuito autorregulatorio de las células (Lee y Young, 2013). Por ejemplo, *c-Myc* o *p53* son factores de transcripción de los que dependen la mayoría de los tumores para mantener su crecimiento y proliferación (Littlewood et al., 2012). *MYC* es el oncogen amplificado con mayor frecuencia, y su mayor expresión está asociado a un peor pronóstico y mayor agresividad del tumor (Lee y Muller, 2010).

En resumen, la gran influencia que tienen los factores de transcripción en la homeostasis celular, así como su implicación demostrada en varios procesos necesarios para la tumorogénesis y el desarrollo del tumor, los convierte en unos candidatos más que interesantes para su estudio como dianas terapéuticas y herramientas de diagnóstico. Ya hay varios estudios encaminados en este sentido (Bhagwat y Vakoc, 2015), aunque hace falta un estudio holístico que aproveche la gran cantidad de datos generados y describa el estado de activación de los factores de transcripción en los distintos tipos de cáncer, y en ello consistirá una parte de esta tesis.

1.1.1.2 Rutas de señalización

Otro grupo de genes importantes en la regulación de los procesos celulares son aquellos que participan en rutas de señalización claves para el mantenimiento de la homeostasis.

La transición de células únicas a complejos organismos pluricelulares ha sido posible gracias al desarrollo de mecanismos de comunicación entre

sus células para actuar orquestadamente durante procesos como la adquisición de nutrientes, motilidad y defensa (Cantley, 2012). Estos procesos que permiten a las células percibir y responder correctamente a cambios en su microambiente se conocen como rutas de señalización. Las rutas de señalización están formadas por una serie de moléculas, en su mayoría proteínas o metabolitos, que forman parte de la cascada de cambios secuenciales ocasionados por la activación de un receptor de señal, por ejemplo, la unión de una hormona a una proteína de membrana, y que desencadenará diversas respuestas según el estímulo y la ruta. De esta manera, tras recibir la señal o estímulo, el receptor producirá un cambio en el siguiente eslabón de la cadena de activaciones en la ruta, activando o inhibiéndolo, y así consecutivamente hasta llegar al último elemento efector, encargado de realizar la función celular adecuada en respuesta a la señal recibida. El paso de la señal entre los distintos elementos de las rutas se puede llevar a cabo de distintas maneras, como por ejemplo la adición de una molécula química en el siguiente eslabón de la ruta (p. e. fosforilación), una modificación alostérica, la formación de dímeros o la escisión de una subunidad de una proteína, entre otros. Las respuestas desencadenadas tras la activación de una ruta de señalización suelen ser cambios en la transcripción y traducción de genes, aunque también pueden activar la secreción extracelular de moléculas o modificaciones post-traduccionales en proteínas, como alterar su forma o incluso su localización intracelular.

Estas rutas son las encargadas de controlar y gestionar múltiples funciones celulares esenciales como son el crecimiento celular, la proliferación o el metabolismo entre otros. Es por esto que alteraciones en las rutas de señalización pueden causar diversas patologías como diabetes (Solinas et al., 2007), enfermedades autoinmunes o cáncer (Wang et al., 2013). Existen varias rutas de señalización cuya implicación en la tumorogénesis

y el desarrollo del tumor ha sido estudiada, como por ejemplo las rutas de *mTOR* y *PI3K* (Willems et al., 2012), *JAK/STAT*, *Notch*, *MAPK/ERK*, *NF-kB, Wnt* (Dreesen y Brivanlou, 2007), *TGF-β/BMP* (Prunier et al., 2019), *ErbB* (Hynes y MacDonald, 2009). La mayoría de estas rutas contribuyen en el desarrollo de tumores mediante la síntesis de componentes celulares y la regulación de hallmarks esenciales como el crecimiento celular, proliferación, migración, supervivencia y angiogénesis.

Estos estudios ponen en evidencia cómo las rutas de señalización forman un delicado entramado de elementos que asegura la correcta adaptación celular al medio en el que se encuentra. Si algún elemento falla o se rompe el equilibrio de la red, puede derivar en alguna patología. Su estudio es esencial para entender mejor la etiología y posibles tratamientos a enfermedades como el cáncer. Para entender los mecanismos complejos de las rutas de señalización se requiere de una combinación de aproximaciones tanto experimentales como teóricas, que incluye el desarrollo y análisis de simulaciones y modelos (Kitano, 2002). La biología de sistemas es la rama de la biología que estudia, entre otras cosas, la estructura de las rutas de señalización y cómo cambios en las mismas pueden afectar a la transducción de señal. Este enfoque holístico requiere la integración de distintas fuentes de conocimiento sobre el funcionamiento de los procesos biológicos y sus interacciones internas. Durante los últimos años ha habido un constante desarrollo de múltiples recursos y bases de datos para el estudio de la señalización celular, como la Enciclopedia de Kioto de Genes y Genomas (KEGG por sus siglas en inglés) (Kanehisa et al., 2002) o Reactome (Joshi-Tope et al., 2005).

1.1.2 CARACTERÍSTICAS COMUNES EN CÁNCER

Existen una serie de características biológicas en común (hallmarks) que la mayoría de tumores adquieren durante su desarrollo. Estos hallmarks describen características muy importantes en un tumor para sobrevivir, proliferar y diseminarse, y pueden ser adquiridas por distintos tumores a través de diversos mecanismos y en distintos momentos del desarrollo tumoral. Tras una revisión de su artículo original (Hanahan y Weinberg, 2000), Hanahan y Weinberg (2011) describieron e identificaron 10 hallmarks que comparten la mayoría de tumores (Figura 1.2).

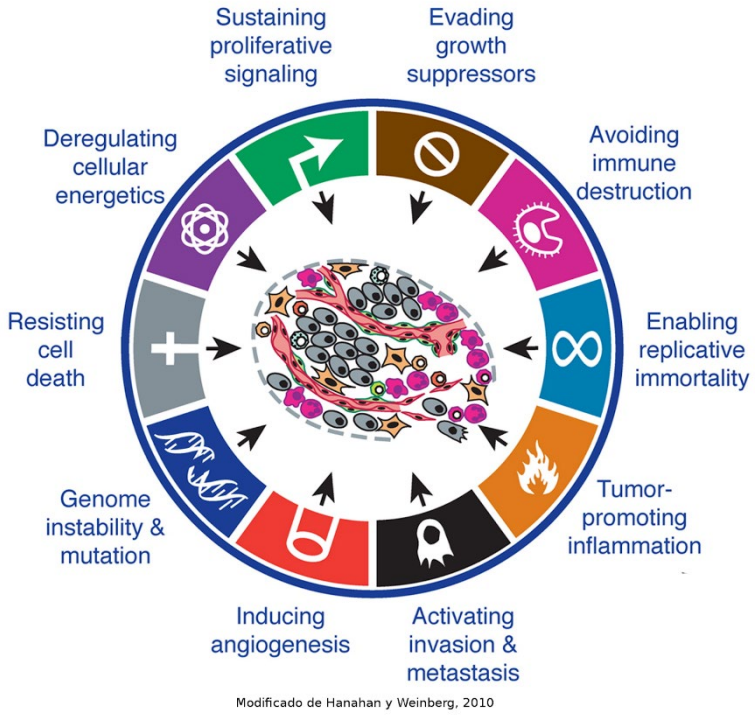


Figura 1.2. ***Hallmarks of cancer.*** Esquema ilustrativo de las características comunes compartidas en la mayoría de tumores.

El mantenimiento de la señalización proliferativa es uno de los rasgos más fundamentales en un tumor. En tejidos normales es necesario un cuidadoso control de la producción y liberación de señales promotoras de crecimiento que instruyen en la entrada y progresión del ciclo celular. Estas señales permiten el mantenimiento del número de células que aseguran una correcta arquitectura y funcionamiento del tejido. Es complicado estudiar de manera precisa las fuentes y mecanismos de señalización proliferativa, debido a la dificultad de acceder experimentalmente a mecanismos que dependen de la regulación temporal y espacial como ocurre en la señalización paracrina. Sin embargo, la señalización mitogénica en células tumorales está mejor descrita (Witsch et al., 2010), así como los mecanismos mediante los cuales son capaces de alterar su regulación. En los tumores estas señales pueden ser desreguladas de distintas maneras, como aumentar el número de receptores de membrana que regulan las rutas de proliferación o la activación constitutiva de alguno de los elementos de la ruta, de manera que no haga falta la activación a través de la unión de un ligando. Por ejemplo, cerca de un 40% de los melanomas contienen mutaciones activadoras de la proteína *BRAF*, lo que induce una activación constitutiva de la ruta MAPK (Davies y Samuels, 2010). También se ha observado cómo las células tumorales son capaces de estimular a las células normales de su microambiente para que produzcan más factores de crecimiento (Cheng et al., 2008). Además de la sustentación positiva de la proliferación, las células cancerosas también deben ser capaces de eludir programas de regulación negativa de la proliferación. Muchos de estos programas dependen de proteínas supresoras de tumores, como *RB* y *TP53*, que son alteradas en un gran número de tumores (Sherr y McCormick, 2002). Un ejemplo de este hallmark podría ser la capacidad que tienen las células neoplásicas de evitar la inhibición por contacto. Cuando las células proliferan mucho también aumentan los contactos

entre ellas, lo cual manda una señal para parar la proliferación. Se ha observado como algunos tumores son capaces de esquivar este mecanismo mediante la supresión de algunos genes encargados de desencadenar la inhibición por contacto, como *LKB1* (Shaw, 2009).

La muerte celular programada por apoptosis sirve de barrera natural al desarrollo del cáncer, es por eso que las células neoplásicas tienen que desarrollar mecanismos para conseguir evitar la muerte celular. En distintos tumores avanzados se ha observado una atenuación de la apoptosis (Adams y Cory, 2007). La apoptosis se desencadena en respuesta a varios estreses fisiológicos que las células experimentan durante la tumorogénesis. La maquinaria apoptótica puede ser activada por una vía extrínseca, como la recepción de señales extracelulares, o intrínseca, cuya activación está más implicada como barrera contra el cáncer y se puede desarrollar tras la detección de daño en el DNA o de niveles elevados de oncogenes. Una vez activada la apoptosis se inicia una cascada de proteolisis mediante la cual se va desmontando la maquinaria celular para ser consumida por células fagocíticas. A pesar de evitar la apoptosis, hay evidencias de la capacidad del sistema inmune de reconocer a los tumores, como el aumento de incidencia de cáncer en pacientes inmunodeprimidos (Vajdic y van Leeuwen, 2009). Sin embargo, la mayoría de tumores que proliferan son capaces de evitar esta inmunidad antitumoral. Aunque aún quedan bastantes estudios para comprender este proceso, se ha propuesto que en el proceso inicial de proliferación tumoral podría ocurrir una especie de selección natural en el que aquellas células más reconocibles por el sistema inmune se destruyen y las que van quedando y proliferando son aquellas que consiguen escapar a su reconocimiento (Kim, 2007). Esta capacidad para evitar la destrucción inmune es otra característica común en la mayoría de los tumores.

Otra barrera fisiológica en contra de la proliferación celular indefinida característica de las células neoplásicas a parte de la apoptosis es la senescencia. La mayoría de células diferenciadas en nuestro organismo son capaces de realizar el ciclo de crecimiento y división celular un número limitado de veces. El número de veces que una célula puede replicarse está determinado por la longitud de los telómeros (Shay y Wright, 2000). Los telómeros son una repetición en tándem de 6 nucleótidos ubicados al final de los cromosomas para evitar su degradación. Por cada ciclo celular en el que entra una célula, estas estructuras se van acortando hasta que llega un punto en el que no son capaces de proteger a los cromosomas y la célula, aunque no muere, no es capaz de replicarse más, entra en senescencia. Existe una enzima capaz de añadir estos segmentos repetitivos al final del ADN telomérico para evitar su erosión, la telomerasa. Prácticamente no se observa actividad de telomerasa en la mayoría de células somáticas, sin embargo, se ha observado como su sobreexpresión es uno de los métodos que emplea el cáncer para conseguir evitar la senescencia (Blasco, 2005).

Al crecer constantemente, los tumores necesitan una reprogramación del metabolismo energético de la célula para satisfacer los requerimientos energéticos necesarios para abastecer el crecimiento y proliferación indefinidos. En tejidos normales, con una disponibilidad de oxígeno, las células obtienen energía principalmente mediante la degradación de glucosa a piruvato, a través de un proceso denominado glucólisis, y éste posteriormente será degradado a dióxido de carbono, obteniendo bastante más energía que en el proceso anterior, con la ayuda de las mitocondrias. Incluso en presencia de oxígeno, las células neoplásicas son capaces de modificar su metabolismo energético para obtener su energía de manera prácticamente exclusiva mediante la glucólisis, entrando en un estado conocido como glicólisis aeróbica (DeBerardinis et al., 2008). Para

atender esta alta actividad metabólica los tumores requieren un mayor abastecimiento de nutrientes, así como la capacidad de evacuar los desechos metabólicos que se generan, y para ello tienen que ser capaces de crear vasos sanguíneos, es decir, de inducir angiogénesis. La angiogénesis es un proceso que en condiciones normales se activa únicamente en determinadas condiciones como el proceso de curación de alguna herida o en el ciclo reproductivo femenino. Sin embargo, en cáncer, mediante la alteración de algunos reguladores angiogénicos, como *VEGFA* o *TSP1*, los tumores son capaces de activar y mantener este proceso para sostener el crecimiento neoplásico (Baeriswyl y Christofori, 2009).

Otra característica importante de los tumores es la capacidad que tienen algunas de sus células de abandonar el tumor original y transportarse a otras partes del cuerpo para formar otros tumores. Este proceso de invasión y metástasis involucra a una serie de cambios celulares conocidos como cascada metastásica. Comienza por una invasión local de la matrix extracelular, seguido por la infiltración de células neoplásicas al torrente circulatorio a través de vasos sanguíneos y linfáticos (intravasación). Una vez circulan por el cuerpo las células cancerosas son capaces de escapar del espacio luminal de los vasos (extravasación) y formar pequeños nódulos de células neoplásicas (micrometástasis) que posteriormente formarán nuevos tumores (colonización). Hay algunas alteraciones que pueden promover la cascada metastásica, como la modificación de la forma o de las uniones de células entre ellas y con la matriz extracelular. En este sentido, se ha observado como la inhibición de las cadherinas E o la activación de las cadherinas N juegan un papel importante en la migración de células de tumores sólidos (Cavallaro y Christofori, 2004). También se ha descrito la importancia de unos programas regulatorios, conocidos como la transición EMT-MET (de sus

siglas en inglés “epithelial-mesenchymal transition” y “mesenchymal-epithelial transition”), implicados en las transformaciones fisiológicas requeridas por las células para abandonar el tumor original y posteriormente, una vez se ha depositado en otra parte del organismo, formar un nuevo tumor (Micalizzi et al., 2010).

La adquisición de los hallmarks nombrados hasta ahora es posible gracias a dos características capacitantes. En primer lugar, es necesario un aumento de la inestabilidad del genoma y del número de mutaciones para la adquisición del listado de genes mutados indispensables para promover la tumorogénesis. En el proceso de creación del tumor, ciertas poblaciones clonales de células van dominando según van adquiriendo mutaciones que le confieren una ventaja selectiva. Esta sucesión de expansiones clonales se ha observado que es acelerada por el incremento de la tasa de mutaciones (Negrini et al., 2010). Esta mutabilidad se consigue por un aumento en la sensibilidad de agentes mutagénicos, el funcionamiento incorrecto de componentes de la maquinaria de mantenimiento del genoma o también de alguno de los sistemas de vigilancia que controlan la integridad del genoma (Jackson y Bartek, 2009). En segundo lugar, se ha observado como la inflamación es capaz de fomentar la tumorogénesis y la adquisición de algunas características tumorales en sus primeras etapas (Qian y Pollard, 2010). La contribución de la inflamación al desarrollo del tumor tiene que ver con el reclutamiento de algunas células inmunes y su efecto promotor del tumor, y también por el suministro de ciertas moléculas bioactivas, como factores de crecimiento, factores que promueven la angiogénesis, invasión y metástasis, e incluso especies reactivas del oxígeno promotoras de mutagénesis (Grivennikov et al., 2010).

1.1.3 HETEROGENEIDAD EN CÁNCER

Se han identificado varias características funcionales en común para la mayoría de tumores (ver sección 1.1.2), sin embargo, los mecanismos causantes de estos fenotipos moleculares pueden ser muy diversos y los tumores pueden ser muy diferentes de un paciente a otro incluso aunque sean en el mismo tejido. Este tipo de variabilidad, conocida como heterogeneidad intertumoral, tiene un profundo impacto clínico a la hora de diagnosticar y elegir la terapia adecuada.

La heterogeneidad intertumoral suele radicar en factores genéticos, aunque recientemente se ha observado como el microambiente del tumor también puede influir (Miething et al., 2014). Un claro ejemplo de la heterogeneidad intertumoral se puede encontrar en la clasificación por subtipos de numerosos tipos de cáncer. Estas clasificaciones suelen atender a características moleculares o fenotípicas del tumor, como la sobreexpresión de cierto receptor o la mutación de un determinado gen. Por ejemplo, el cáncer de mama se suele clasificar en varios subtipos: subtipo luminal, si las células cancerígenas expresan receptores endocrinos (progesterona y/o estrógenos); subtipo *HER2*, si sobreexpresan el receptor *HER2* y no suelen tener receptores de estrógenos; subtipo triple negativo, cuando no expresa receptores de estrógenos, ni progesterona, ni *HER2*.

Un enfoque muy interesante para estudiar la heterogeneidad intertumoral es agrupar el mayor número de tumores, tanto de la misma procedencia como de distintos tejidos, para comparar los mecanismos, comunes y/o distintos, que pueden causar el fenotipo de la enfermedad. Estos estudios, acuñado por la red de investigadores integrados en el proyecto The Cancer Genome Atlas (TCGA) como análisis de pan-cáncer, aportan una visión holística a la investigación de cáncer y permiten caracterizar mejor la

heterogeneidad a nivel molecular, de expresión y alteraciones genéticas existente en distintos tipos de tumor (ICGC y TCGA, 2020). Este enfoque también puede ser de utilidad para detectar distintos tipos de tumor con perfiles fenotípicos similares (Subbiah et al., 2020). Además, estos estudios aportan una fuente de datos muy valiosa, amplia e interesante para toda la comunidad científica, que puede ser reutilizada para diversos estudios y análisis. Una de las mayores aportaciones de datos para el estudio de cáncer es la realizada por el consorcio internacional de genomas del cáncer, ICGC por sus siglas en inglés.

Existe otro nivel de heterogeneidad en los tumores conocido como heterogeneidad intratumoral, que hace referencia a las diferencias fenotípicas que se encuentran entre las células dentro del propio tumor. A pesar de que generalmente los tumores proceden de una única célula (Weinberg, 2013), la mayoría ~~de~~ tumores presentan una amplia heterogeneidad de células con distintas propiedades morfológicas y fisiológicas, como los receptores de membrana que presentan o su capacidad proliferativa. En el proceso de formación del tumor se producen mutaciones aleatorias debido a la inestabilidad genética y la constante proliferación. Esta heterogeneidad genética se traduce en heterogeneidad fenotípica, que será sometida a diversas presiones selectivas, de manera que aquellas mutaciones que confieren una ventaja adaptativa a la célula pueden hacer que esta prospere y prolifere más que otras con mutaciones neutras, aunque estas también pueden fijarse por simple deriva genética (Marusyk y Polyak, 2010). Las poblaciones clonales de células que se forman debido a este proceso selectivo dentro del tumor han sido observadas en la mayoría de tipos de cáncer, por ejemplo, cáncer colorrectal (Testa et al., 2018), de mama (Martelotto et al., 2014), de hígado (Losic et al., 2020), de pulmón (de Briun et al., 2015) o glioblastoma (Sottoriva et al, 2013). También se han descrito otros

factores no genéticos que contribuyen a la heterogeneidad intratumoral, como puede ser la epigenética (Flavahan et al., 2017) o el microambiente del tumor (Fane y Weeraratna, 2019). El estudio de toda esta heterogeneidad a nivel celular no sería posible sin el desarrollo de nuevas tecnologías, como la secuenciación de célula única (scRNAseq), que permiten el análisis a nivel genómico y transcriptómico de células aisladas (ver sección 1.2.2). Aun así, existen varias dificultades técnicas que afrontar a la hora de estudiar la heterogeneidad intratumoral, por ejemplo, problemas de cantidad y calidad de la muestra, ya que una biopsia no tiene necesariamente que representar toda la heterogeneidad del tumor. Además, a esto hay que sumarle que el tumor es un sistema en constante evolución, por lo que el tiempo es un factor importante en la composición del tumor.

En conclusión, la compleja heterogeneidad, tanto intertumoral como de fenotipos celulares dentro de un mismo tumor, depende de múltiples variables, como la genética, epigenética y factores ambientales. La heterogeneidad intertumoral refleja la necesidad de caracterizar bien el tumor para favorecer el desarrollo de terapias personalizadas. Sin embargo, la heterogeneidad intratumoral es la responsable de muchas de las resistencias o recidivas que se observan en este tipo de terapias dirigidas, especialmente en tumores avanzados más heterogéneos, a pesar de que en algunos se observe una respuesta inicial robusta (Marusyk et al., 2020). La gran variabilidad fenotípica observada en distintos tumores, y también dentro de un mismo tumor, pone de manifiesto la necesidad de profundizar en su estudio para comprender mecanismos de resistencia y poder desarrollar nuevas estrategias terapéuticas.

1.1.4 TRATAMIENTOS EN CÁNCER

Actualmente se aplican diversos tipos de tratamiento para el cáncer. Cada paciente recibirá un tratamiento u otro dependiendo del tipo de cáncer, su estadio y el subtipo molecular diagnosticado, aunque también es posible que se aplique una combinación de tratamientos para aumentar la eficacia. Los tratamientos más comunes son:

- **Cirugía**. Consiste en la extirpación quirúrgica del tumor y, en ocasiones, del tejido adyacente para asegurar la eliminación de cualquier extensión del tumor. Suele emplearse en los tumores sólidos en combinación con otros tratamientos. Es la técnica más útil, cuando es posible, ya que extirpar parte o la totalidad del tumor y disminuye drásticamente la probabilidad de recidiva, permitiendo observar *in situ* si el tumor se ha diseminado. Además, una vez extirpado el tumor es posible realizar un estudio para determinar sus características moleculares para así ser más preciso en su diagnóstico y tratamiento posterior. Sin embargo, este tratamiento presenta riesgos derivados de la intervención, y no es posible realizarlo en todos los casos.

- **Radioterapia**. La terapia funciona mediante radiación ionizante focalizada que induce la muerte de las células de rápida división, incluidas las cancerosas. Para ello emplea partículas de alta energía que provocan daños irreparables en el DNA. Existen dos tipos principales de radioterapia, de haz externo y radioterapia interna.

- **Quimioterapia**. Se administran al paciente uno o varios medicamentos citostáticos para la desaparición, detención o reducción del tumor y para aliviar los síntomas derivados del mismo. Es común la combinación de varios fármacos que actúan

sinérgicamente mediante diferentes mecanismos de acción. De esta manera, al controlar la combinación, dosis y tiempo en el que se administran los fármacos, se consigue aumentar la potencia terapéutica y disminuir la toxicidad. Aunque efectivo en muchos casos, estos tratamientos presentan una gran cantidad de efectos secundarios.

- **Inmunoterapia**. Mediante el uso de inmunomoduladores, como citoquinas o anticuerpos monoclonales, se estimula al sistema inmune para que actúe contra las células tumorales. Actualmente existe una nueva técnica de inmunoterapia en desarrollo que se conoce como transferencia adoptiva de linfocitos T. Esta técnica se basa en la expansión *in vitro* de linfocitos T, obtenidos previamente del paciente o de un donante, para posteriormente ser transferidos al paciente. Estos linfocitos pueden ser modificados genéticamente para que reconozcan antígenos específicos de la superficie de las células tumorales.

- **Terapia hormonal**. Se emplea en tejidos hormonodependientes, principalmente del aparato reproductor, como el cáncer de próstata, ovarios, endometrio, o el cáncer de mama. Se basa en la utilización de fármacos que interfieren en la síntesis y metabolismo de hormonas y tiene una gran eficacia para tumores que presentan receptores hormonales en superficie

- **Trasplante de células madre**. En neoplasias hematológicas, como leucemias o linfomas, se realiza un trasplante de médula ósea para añadir al paciente precursores hematopoyéticos capaces de producir células sanas, consiguiendo mitigar los efectos del cáncer.

- **Terapia dirigida**. Tratamiento personalizado atendiendo a las características moleculares específicas del tumor de cada paciente. Algunos de los tratamientos anteriores podrían considerarse dirigidos ya que esta categoría engloba a aquellos tratamientos dirigidos hacia alguna diana molecular concreta.

1.1.4.1 Tratamientos dirigidos y medicina personalizada de precisión

Existe una amplia variedad de tratamientos y terapias disponibles para combatir el cáncer, sin embargo, como hemos visto, los tumores se caracterizan por una enorme heterogeneidad intra e intertumoral, lo que se traduce en un amplio espectro de respuesta al tratamiento entre los pacientes. El auge en las últimas décadas de las tecnologías de secuenciación y otras técnicas moleculares, ha permitido ampliar nuestro conocimiento del cáncer, además, el mayor acceso de los hospitales a estas técnicas moleculares ha tenido como resultado una mejor caracterización de los subtipos moleculares del cáncer, permitiendo estratificar a los pacientes en base a las características intrínsecas e individuales de su tumor.

La integración de este conocimiento molecular y genético junto con los datos clínicos del paciente ha permitido el desarrollo de una práctica clínica adaptada a nivel individual, lo que se conoce como medicina personalizada o de precisión (Fröhlich et al., 2018). Esta aproximación a la salud del individuo permite a la comunidad clínica y científica predecir qué estrategia abordar a la hora de tratar una determinada patología en una persona o grupo específico de personas que comparten características moleculares. Este abordaje, a medida a las características del paciente,

posibilita ajustar mejor las acciones preventivas, diagnósticas y pronósticas, permitiendo una mayor eficiencia y eficacia en el manejo de la enfermedad (Bugarín-González y Carracedo, 2018). Este enfoque es particularmente efectivo en el caso de cáncer, debido a la gran cantidad de recursos que se dedican al tratamiento de esta enfermedad, a la naturaleza molecular del mismo y a la alta tasa de resistencias y recidivas observadas.

Tradicionalmente las estrategias de tratamiento y prevención para una enfermedad solían tener un enfoque único centrado en un paciente promedio, sin tener en cuenta la heterogeneidad que existe entre individuos. De hecho, millones de personas al día toman medicaciones que no les ayudan a superar o mejorar su patología (Schork, 2015). Por ejemplo, en Estados Unidos, para los diez medicamentos con mayor prescripción, entre 1 de cada 3 y 1 de cada 24 pacientes obtienen beneficios al tomar el medicamento prescrito para su patología, el resto no obtienen el efecto deseado (Figura 1.3). Incluso en algunos casos, como ocurre con las estatinas que se prescriben para tratar el colesterol alto, únicamente 1 de cada 50 pacientes que lo toman son beneficiados (Mukherjee y Topol, 2002). Y en ocasiones, no solo no benefician, sino que hay tratamientos con efectos perjudiciales para algunos grupos minoritarios debido al sesgo de etnia, sexo y edad de los participantes que suelen acaparar los ensayos clínicos (Currie et al., 2006).

Para la implementación de la medicina personalizada hacen falta una gran cantidad de datos y estudios prospectivos que observen la heterogeneidad de la población, tanto genética como en la respuesta a distintos tratamientos. Se han desarrollado algunas iniciativas en algunos países o regiones con la intención de implantar la medicina personalizada en el sistema de salud público. Por ejemplo, en el Reino Unido se ha propuesto obtener los genomas de 100.000 personas afectadas por enfermedades

raras o cáncer, para aumentar el conocimiento sobre las causas, tratamientos y cuidados de las enfermedades (England, 2016).

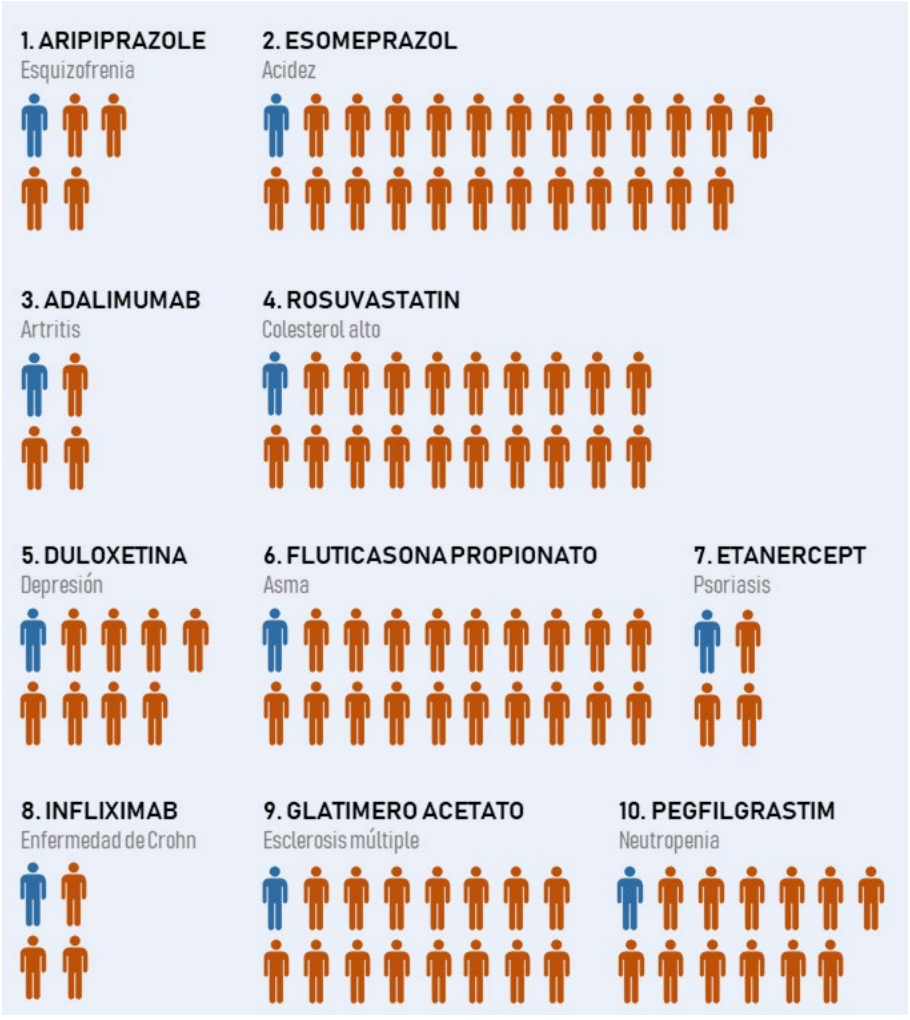


Figura 1.3. Eficacia de los 10 medicamentos más prescritos en EEUU sobre la población general. Se indica el compuesto activo del fármaco y la enfermedad para la que está indicado. Se muestran en azul los individuos en los que el medicamento tiene la eficacia deseada y en rojo aquellos en los que no tiene el efecto deseado. Fuentes: 1) El-Sayeh y Morganti, 2006. 2) Gralnek et al., 2006. 3) Chen et al., 2006. 4) Ridker et al., 2009. 5) Cookson et al., 2006. 6) Suissa, 2015. 7 y 8) Kristensen et al., 2007. 9) Freedman et al., 2008. 10) Dekker et al., 2006

La medicina personalizada aplicada en cáncer busca actuar sobre los cambios moleculares que promueven el crecimiento y la proliferación de las células en un tumor y su migración a otras partes del cuerpo. El cambio de la visión órgano-céntrica para la toma de decisiones en cuanto al tratamiento hacia una aproximación personalizada con un profundo análisis molecular, ha supuesto uno de los mayores avances en la oncología moderna (Hyman et al., 2017). Las nuevas tecnologías de secuenciación han abierto la puerta a la detección de mutaciones en genes, amplificaciones y fusiones de manera masiva, lo que ha permitido caracterizar mejor el tumor de un paciente y entender los mecanismos para poder tratarlo adecuadamente. A pesar de ser una aproximación relativamente reciente, la medicina personalizada ya ha ayudado a mejorar el pronóstico y el tratamiento en numerosos tipos de cáncer (Corey et al., 2020; Matrone et al., 2020; Rambow et al., 2018). Para aplicar terapias dirigidas a estas alteraciones genéticas, es esencial caracterizar bien y distinguir qué grupos de pacientes se beneficiarán o no de este tipo de tratamientos mediante la identificación de distintos tipos de biomarcadores de pronóstico, diagnóstico o respuesta (Gambardella et al., 2020). Por ejemplo, en cáncer de mama, además de la identificación de los receptores de hormonas como dianas terapéuticas, la identificación de otros biomarcadores como mutaciones en los genes *PIK3CA*, *ERBB2* o la amplificación de *HER2* han mejorado ampliamente la aproximación terapéutica (André et al., 2019; Smyth et al., 2020; Ignatiadis y Sotiriou, 2013).

A pesar de los prometedores beneficios de la medicina personalizada, uno de los principales inconvenientes de centrarse en dirigir el tratamiento a una diana molecular es que el tumor puede desarrollar mecanismos de resistencia. Algunos tipos de cáncer son dependientes de algunas rutas de señalización para su crecimiento, y hay terapias dirigidas que se centran

en atacar a dianas de estas rutas. En ocasiones, algunas células del tumor encuentran una manera de evitar el gen al que va dirigido el tratamiento, por ejemplo, mediante la activación de otros genes o rutas (Gallyas et al., 2020). Por esta razón, las terapias dirigidas pueden tener mejor resultado cuando se usan junto con otras terapias dirigidas o con otros tratamientos contra el cáncer, como la quimioterapia y la radiación. Es necesario analizar y describir en profundidad estos mecanismos de resistencia para poder tratarlos posteriormente.

En definitiva, la medicina de precisión tiene el potencial de transformar la medicina general ajustando individualmente los tratamientos. Aún quedan algunos desafíos para su amplia implementación en los sistemas de salud, como la insuficiencia de infraestructuras, de generación de evidencias e intercambio de datos, retos económicos o conseguir una mejor difusión, compromiso y confianza por parte de la población (Dzau y Ginsburg, 2016). En cáncer se ha observado como un tratamiento correcto en el momento adecuado, dota a los médicos de una mejor eficiencia y eficacia, mientras que los pacientes disfrutan de una mejor calidad en su cuidado.

1.2 TÉCNICAS ÓMICAS EN CÁNCER

En las últimas décadas ha habido una explosión en la cantidad de datos biológicos generados por un número, también creciente, de nuevas técnicas, permitiendo la detección de numerosas alteraciones moleculares (figura 1.4). Las ómicas comprenden a todas las nuevas tecnologías que permiten caracterizar o cuantificar grandes cantidades de componentes moleculares y que permite estudiar la estructura, función y/o dinámica de uno o varios organismos (Schneider y Orchard, 2011). Los campos más relevantes dentro de las ómicas son la genómica, transcriptómica, proteómica y metabolómica, y se desarrollaron para estudiar la complejidad biológica a nivel de la secuencia genética, expresión de genes, proteínas o metabolitos presentes en las muestras analizadas, respectivamente.

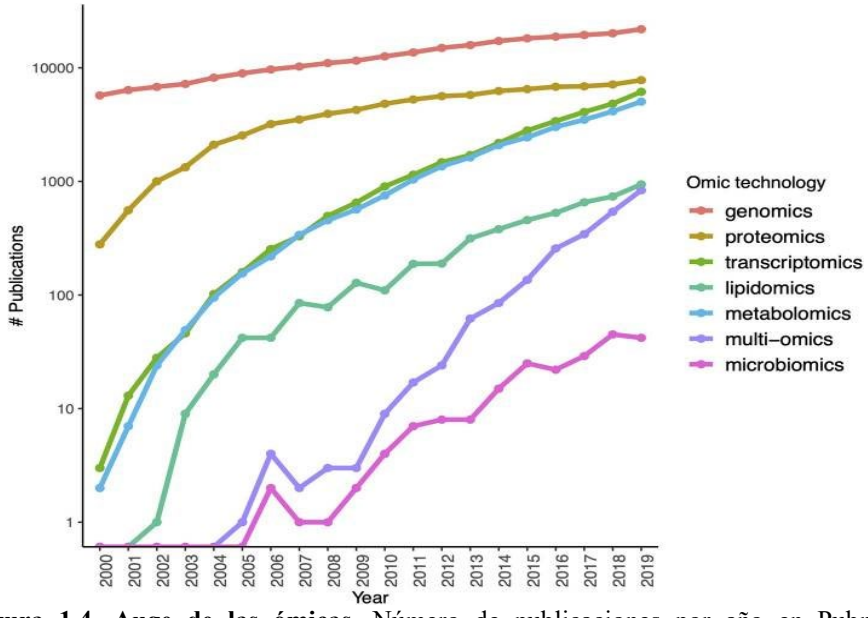


Figura 1.4. Auge de las ómicas. Número de publicaciones por año en Pubmed mencionando cada ómica en su título desde el año 2000. Fuente: de Anda-Jáuregui y Hernández-Lemus, 2020.

Cada ómica contribuye a desvelar cada una las distintas capas de elementos moleculares que contribuyen al desarrollo y mantenimiento de distintas enfermedades. El estudio del cáncer por las distintas ómicas ha hecho posible conocer tanto su diversa heterogeneidad como sus características en común (ver secciones 1.1.3 y 1.1.2). El constante crecimiento de experimentos y bases de datos de cáncer hace posible la integración de distintas fuentes para profundizar en el conocimiento sobre la causa, progresión, caracterización y desarrollo de la enfermedad. Sin duda, una de las mayores contribuciones a la creación y disponibilidad de datos ómicos de cáncer es el llevado a cabo por el consorcio internacional del genoma del cáncer (ICGC/TCGA), que ha conseguido juntar a 76 proyectos o grupos individuales para estandarizar y recabar datos de distintas ómicas de 17000 pacientes de cáncer en 21 tipos de tumor (ICGC).

1.2.1 TRANSCRIPTÓMICA EN CÁNCER

La transcriptómica mide la presencia y abundancia de RNA de una o varias muestras en un determinado contexto fisiológico. Su aplicación más común es la medida de la expresión génica. El perfil de expresión de un fenotipo es un reflejo de su estado biológico, de manera que la comparación de estos perfiles, a través de análisis de expresión diferencial, es útil para observar alteraciones importantes en distintas enfermedades (de Anda-Jáuregui y Hernández-Lemus, 2020). La popularidad de la transcriptómica empezó tras la llegada de los microarrays, que hicieron esta aproximación para medir la expresión más viable económica y técnicamente que la medida de la abundancia de proteínas que se hacía por entonces (Dubitzky et al., 2003). Tras la llegada de las tecnologías de nueva secuenciación surgió otro método para medir

la abundancia de tránscritos, la secuenciación de RNA (RNAseq). Cada metodología tiene sus consideraciones técnicas, pero para el análisis de sus resultados generalmente hay que seguir los mismos pasos: adquisición y preprocesado de datos, eliminar los artefactos técnicos, control de calidad y normalización.

Además de la expresión génica, el análisis completo del transcriptoma también permite identificar RNAs no codificantes (ncRNA) como micro-RNA (miRNA), long non-coding RNAs (lncRNA), small nucleolar (snRNA), Piwi-interacting RNA (piRNA), entre otros. El papel de estos tránscritos y su contribución al programa regulatorio está siendo investigado en cáncer y otras enfermedades (Anastasiadou et al., 2018). Actualmente, tras hacerse más accesible, el RNAseq es el método preferido por parte de los investigadores debido a que ofrece bastante más información específica que los microarrays. A parte de medir la abundancia de los genes y ncRNAs, el RNAseq permite la detección de eventos de fusión de genes, de splicing o incluso variantes clonales en cáncer (Cibulskis et al., 2013). También es posible mediante esta aproximación distinguir patrones de expresión clínicamente relevantes que permitan clasificar mejor a los tumores, por ejemplo, los subtipos en cáncer de mama (luminal, *HER2*+ o triple negativo) o glioblastoma (mesenquimal, clásico y proneural). De hecho, existen varios paneles de genes comerciales que únicamente miran la expresión de ciertos genes de interés. Estos perfiles de expresión son usados por oncólogos para identificar grupos de riesgos en los pacientes en base a la sobreexpresión de ciertos genes involucrados en la regulación o desarrollo del tumor, y así poder mejorar el pronóstico, ajustar su tratamiento e incluso predecir el riesgo de recurrencia. Por ejemplo, en cáncer de mama se emplean varios paneles de genes comerciales, como PAM50, que permite clasificar a los pacientes en uno de los subtipos de la enfermedad, o

MAMMAPRINT, que analiza un panel de 70 genes e indica el riesgo de recurrencia a diez años (Galanina et al., 2011).

1.2.2 SINGLE CELL

Las distintas ómicas analizan una muestra de tejido que puede ser más o menos homogéneo en su contenido celular. Recientemente, en cáncer, la heterogeneidad intratumoral (ver sección 1.1.3) ha adquirido bastante importancia a la hora de estudiar los tumores. El estudio de la composición de los distintos tipos celulares y poblaciones clonales ha sido posible, en gran parte, gracias al desarrollo de nuevas técnicas que permiten analizar la variabilidad del transcriptoma de una muestra célula a célula, la secuenciación de RNA de células únicas (scRNAseq). El primer paso para obtener el transcriptoma de una muestra mediante scRNAseq consiste en la disgregación de las células que la componen, y para ello se suele emplear el ordenamiento de células activado por fluorescencia (FACS) o la microfluídica. Con microfluídica se consiguen aislar las células en pequeñas micro o nanogotas que contendrán todos los reactivos necesarios para realizar la secuenciación, mientras que las FACS consiguen separar las células en una placa con ciertos marcadores que además se pueden emplear para distinguir tipos celulares y seleccionar aquellas que son de interés (Lafzi et al., 2018). Antes de la amplificación por PCR y secuenciación, normalmente se añaden una serie de etiquetas de nucleótidos (UMI) que sirven para identificar de manera exclusiva a cada tránscrito para evitar sesgos posteriormente en la cuantificación.

A priori cabría esperar que los datos obtenidos de experimentos de scRNAseq fueran similares a los de RNAseq, sin embargo, el análisis de células únicas descubre la mayor dispersión, variabilidad y complejidad

en las distribuciones de sus valores. De hecho, uno de los mayores problemas en scRNAseq es que no se consiguen capturar todas las moléculas de mRNA, de manera que en la matriz de expresión se observan muchos más genes sin expresión de los que realmente existen, fenómeno conocido como dropout. Existen varias aproximaciones para bien evitar o intentar solucionar este problema en los análisis. En pocos años se han desarrollado numerosos métodos de imputación de los dropouts empleando distintos métodos matemáticos, estadísticos y de machine learning (van Dijk et al., 2017; Eraslan et al.,2019; Gong et al., 2018).

1.2.3 TÉCNICAS DE ANÁLISIS DE TRANSCRIPTOMAS: GSEA Y MODELOS MECANÍSTICOS

Uno de los principales retos a la hora de analizar la gran cantidad de información procedente de experimentos transcriptómicos consiste en ser capaz de extraer conclusiones biológicas relevantes a partir de los datos generados. Un experimento de RNAseq es capaz de obtener información sobre la expresión de ~20000 genes para numerosas muestras, sin tener en cuenta los ncRNAs. La aproximación inicial al análisis de datos de expresión génica consiste en una comparación entre grupos de interés, generalmente un grupo caso y otro control, mediante un test estadístico como t-test o Wilcoxon. Como resultado se obtiene un p-valor que será empleado para discernir los genes diferencialmente expresados entre dos grupos. Este p-valor suele ajustarse para reducir el número de falsos positivos que se pueden producir al realizar múltiples comparaciones (una por gen). El análisis de expresión diferencial tiene algunos inconvenientes, especialmente a la hora de interpretar sus resultados, por ejemplo, en una comparación de muestras de cáncer contra tejido normal,

se pueden observar fácilmente cientos o miles de genes diferencialmente expresados. A esto hay que sumarle que cerca del 26% de los genes humanos tienen múltiples funciones biológicas (Pritykin et al., 2015). Para mejorar la interpretación de estos resultados se han desarrollado numerosos métodos de análisis funcionales, el GSEA (del inglés, Gene Set Enrichment Analysis), es probablemente uno de los tests más usados. El GSEA es un método de enriquecimiento funcional que observa si un grupo de interés está significativamente más representado en uno de los extremos de una lista ordenada, como la resultante de un análisis de expresión diferencial (Maleki et al., 2020). El grupo de genes de interés suele definirse en base a un conocimiento biológico previo, como por ejemplo una función biológica o los genes pertenecientes a una ruta de señalización (Al-Shahrour et al., 2007). Este método también presenta sus inconvenientes, ya que por ejemplo se ha visto cómo la expresión de los genes que desarrollan una función no necesariamente tienen que estar correlacionados (Montaner et al., 2009). Otro inconveniente, cuando se emplea esta metodología en el análisis de rutas de señalización es que no se está teniendo en cuenta propiedades de las rutas como su estructura o las relaciones activadoras o inhibidoras entre los distintos genes, lo que puede llevar a fallos en la interpretación de su estado de activación. En la última década se han desarrollado varios modelos mecanísticos que tienen en cuenta las distintas propiedades de las rutas de señalización e integran varias fuentes de conocimiento para analizarlas (Nam y Park, 2012; Sebastian-Leon et al., 2014; Hidalgo et al., 2017; Jacob et al., 2012; Li et al., 2016). Una reciente comparación de varios métodos de análisis de rutas de señalización (Amadoz et al., 2019), ha demostrado como uno de estos métodos, Hipathia (Hidalgo et al., 2017), es el que presenta mayor especificidad y sensibilidad entre los métodos desarrollados más usados hasta ahora. Hipathia extrae información de la base de datos KEGG (Kanehisa et al., 2002) y con los mapas de rutas de señalización define los

circuitos de señalización que conectan cualquier posible receptor con una proteína efectora. Una vez definidos los circuitos, se usan valores de expresión de experimentos transcriptómicos para modelar la señal propagada desde los nodos receptores y que pasa a través de los nodos que conforman la red hasta llegar al nodo efector. Se recoge el valor de señal que llega a cada muestra de manera que cada una tiene su propio valor del estado de activación de los distintos circuitos.

Estos métodos de análisis funcional han permitido una mejor comprensión de la compleja biología de distintas enfermedades como el cáncer. Algunas de estas aproximaciones tienen sus limitaciones, y es por eso que existe un constante mantenimiento y mejora de los métodos actuales, así como desarrollo de nuevas formas para estudiar las implicaciones funcionales de los cambios observados en los experimentos de transcriptómica.

Capítulo 2

Objetivos y estructura de la tesis

2. OBJETIVOS Y ESTRUCTURA DE LA TESIS

La pérdida de la homeostasis regulatoria suele ser el desencadenante del desarrollo y crecimiento de la mayoría de los tumores. A pesar de que se han conseguido identificar varias características comunes en cáncer, existe una gran heterogeneidad en las alteraciones de los mecanismos regulatorios promotores de la tumorogénesis. Esta heterogeneidad se puede observar no solo entre los distintos tipos de tumores sino entre las propias células de un mismo tumor. El desarrollo de nuevas tecnologías ha permitido el estudio de la heterogeneidad tanto a nivel intertumoral como intratumoral. Mediante este tipo de estudios es importante la detección de elementos en común, que permitan desarrollar estrategias terapéuticas que puedan servir a un amplio número de tumores, así como de elementos diferenciadores, que sirvan para caracterizar mejor y detectar aquellos casos que se puedan beneficiar de terapias específicas.

En este contexto, se ha planteado como objetivo principal de esta tesis el desarrollo de metodologías que permitan el estudio de la heterogeneidad regulatoria en cáncer a distintos niveles, tanto entre los tumores de distintos tipos de cáncer, como entre las células de un mismo tumor. Para ello se propusieron los siguientes objetivos específicos:

- Desarrollar una metodología que permita detectar el estado de activación de los factores de transcripción en un estudio transcriptómico.
- Desarrollar una metodología que permita emplear modelos mecanísticos en experimentos scRNAseq para inferir el estado de activación de las rutas de señalización.

- Realizar un estudio holístico que describa la heterogeneidad intertumoral regulatoria en varios tipos de tumores.
- Estudiar la heterogeneidad intratumoral en un experimento scRNAseq.
- Realizar intervenciones en el modelo mecanístico simulando el efecto de fármacos para observar la heterogeneidad de respuesta.
- Proponer distintos marcadores para su aplicación en la medicina personalizada.

Para alcanzar estos objetivos, la presente tesis se ha realizado mediante un compendio de artículos que recogen el trabajo realizado. La longitud, el formato y la estructura de los artículos científicos presentados varían según los requerimientos específicos de cada revista. Las referencias citadas en los artículos se encuentran, con el formato demandado por la revista, al final de sus respectivos capítulos, mientras que las referencias completas citadas en el resto de la tesis se encuentran al final del trabajo. El trabajo realizado queda estructurado de la siguiente manera:

- El **capítulo 3** recoge el análisis realizado en el artículo "The pan-cancer pathological regulatory landscape", mediante el cual se busca realizar un análisis holístico que describa la heterogeneidad en la activación de los factores de transcripción a nivel intertumoral y su correlación con variables clínicas, como son el estadio tumoral o la supervivencia. Para ello, se desarrolló una metodología que permita observar el estado de activación de los factores de transcripción no solo en un conjunto de muestras sino en cada muestra individual.
- En el **capítulo 4** donde se presenta el artículo "Mechanistic models of signaling pathways deconvolute the glioblastoma single-cell

functional landscape", en el que se desarrolla una metodología para aplicar modelos mecanísticos en experimentos de scRNAseq permitiendo analizar la heterogeneidad intratumoral de los estados de activación de las rutas de señalización. También se proponen diversos mecanismos de resistencia para ciertos tratamientos dentro de un mismo tumor.

- Los **capítulos 5 y 6** ponen en común los resultados obtenidos en los dos artículos y se exponen las conclusiones obtenidas de los mismos.

Capítulo 3

The pan-cancer pathological regularory landscape

Falco, M. M., Bleda, M., Carbonell-Caballero, J., & Dopazo, J. (2016). The pan-cancer pathological regulatory landscape. Scientific reports, 6, 39709.

3. THE PAN-CANCER PATHOLOGICAL REGULATORY LANDSCAPE

3.1 ABSTRACT

Dysregulation of the normal gene expression program is the cause of a broad range of diseases, including cancer. Detecting the specific perturbed regulators that have an effect on the generation and the development of the disease is crucial for understanding the disease mechanism and for taking decisions on efficient preventive and curative therapies. Moreover, detecting such perturbations at the patient level is even more important from the perspective of personalized medicine. We applied the Transcription Factor Target Enrichment Analysis, a method that detects the activity of transcription factors based on the quantification of the collective transcriptional activation of their targets, to a large collection of 5607 cancer samples covering eleven cancer types. We produced for the first time a comprehensive catalogue of altered transcription factor activities in cancer, a considerable number of them significantly associated to patient's survival. Moreover, we described several interesting TFs whose activity do not change substantially in the cancer with respect to the normal tissue but ultimately play an important role in patient prognostic determination, which suggest they might be promising therapeutic targets. An additional advantage of this method is that it allows obtaining personalized TF activity estimations for individual patients.

3.2 INTRODUCTION

Transcription factors (TFs) play a crucial role in the dynamic regulation of the gene expression program[1]. The knowledge cumulated in the last years on diverse cellular gene expression programs has drastically increased our understanding of the effects of dysregulation of gene expression in disease. In fact, a broad range of diseases and syndromes, including cancer[2], are caused by mutations that affect TFs either directly or indirectly, by affecting cofactors, regulatory sequences, chromatin regulators, and noncoding RNAs that interact with these regions[3]. Specifically, dysregulations or changes in the activation status of distinct TFs are known to be linked to a number of cancers[4,5,6]. Actually, many oncogenes and tumour suppressor genes, including the well-known *P53* gene[7], are in fact[8] TFs. Moreover, many cancer treatments are essentially transcriptional interventions[9]. Thus, hormonal therapies in breast and prostate cancers to block tumour progression are classical examples. More sophisticated interventions are the inhibition of global epigenomic regulators like *BRD4*[10]. Consequently, understanding the determinants of the transcriptional changes leading to disease states in patients is a prerequisite to restore the normal functions of a cell or a tissue.

Alterations in the transcriptional regulatory network due to perturbed TF activity cause the dysregulation of gene expression observed during cancer progression. Different reverse engineering methods[11,12,13,14,15,16,17] have been proposed to infer the specific TF activity that accounts for the observed differential expression across conditions. Reverse engineering methods use the transcription level of a TF to estimate its activity by calculating different types of correlation to its corresponding target genes. However, using TF expression levels as proxies of their activities can be misleading by several reasons. Firstly, the mRNA expression levels of

many TFs are often relatively low compared to other genes, which increase the uncertainty of the corresponding measurements. Secondly, the regulation of TFs at the protein level has shown to be more relevant than changes at the mRNA level, as demonstrated for example in hypoxia-inducible factors[18] and p53[19]. Moreover, the binding of a TF to the corresponding TFBS does not necessarily imply a transcriptional activity because post-transcriptional modifications and some extra co-factors may be required to promote gene expression[20,21].

As a consequence of this, TF expression levels cannot be considered good descriptors of their activity. Contrarily, the expression levels of the TF's targets, in which all the above mentioned effects are integrated, seem to be a more reasonable readout of TF activity. Despite the simplicity of this idea and its enormous potential, only a few algorithmic proposals have been made that exploit TF's target expression levels to infer the corresponding TF activities, such as BASE[22], RENATO[23], REACTIN[24], RABIT[25] or others[26]. These methods have been applied to the study of survival in breast cancer[27] or to obtain signatures of tumour stage in kidney renal clear cell carcinoma[28].

Here we use a simple but efficient method to systematically detect TFs with altered activity by studying the activity of their corresponding target genes across a total of 5607 samples covering eleven cancer types. This study allowed us to produce the first comprehensive catalogue of TFs with activity altered across a broad spectrum of cancer types. Since the method used can also return personalized values of TF activity for each patient, we could also identify a number of TFs whose altered activity was significantly associated to patient's survival, demonstrating their relevance in cancer progression and their potential as therapeutic targets.

3.3 RESULTS AND DISCUSSION

3.3.1 Changes in TF activity across the different cancers

Raw RNA-seq counts for all the eleven cancers studied (Table 3.1) were normalized as described in Methods and tumour samples were compared to their normal tissue counterparts to obtain lists of genes differentially expressed. TF Target Enrichment Analysis (TFTEA) was applied to these lists ranked by value of the statistic. Figure 3.1 show changes in the activity of the different TFs when cancers are compared to their corresponding normal tissues. The predominant observed behaviour is the increase in TF activity. Actually, a set of TFs (*E2F6, E2F4, MYC, MYC:MAX* and *NRF1*) are always significantly more active in cancers than in normal tissues, and others (*EGR1, ELF1, SP1, YY1, USF1, SP2, ZBTB33, MAX, CTCFL* and *NR2C2*) are significantly active in almost all the cancers with a few exceptions, which suggest for them an important role in cancer development and progression. Actually, all of them appear in the COSMIC database[29] and some of them are well-known oncogenes such as *MYC*[30,31,32], *MAX* and *MYC:MAX*[33], or proteins of the *E2F* family[34], whose over-expression induces uncontrolled cell proliferation because they are TFs located upstream in pathways that control cell cycle[35], being also considered prognostic factors[36]. The *YY1* TF is a multifunctional protein that regulates various processes of development and differentiation and have a clear involvement in tumorigenesis, having been proposed as potential prognostic marker of diverse cancers[37]. *SP1* and *SP2* regulate many of the genes involved in the Warburg effect[38], a well-known cancer hallmark[39]. Actually, high levels of *SP1* protein are considered a negative prognostic factor for several cancers[40,41].

There are also a few TFs that show simultaneously significant, though opposite, behaviours across the studied cancers. This is the case, for

example, of *JUN:FOS*, which induces anchorage-independent growth[42] and *SPI1*, a known oncogene that increases the speed of replication[43], which are deactivated in colon (COAD), uterine (UCEC), bladder (BLCA), lung (LUAD and LUSC) and prostate adenocarcinoma (PRAD) cancers, while are activated in the rest of cancers, suggesting the existence of different growing strategies in these two groups of cancers.

On the other hand, a few TFs systematically display a significant decrease in their activities. For example, two TFs with a largely unexplored role in human tumorigenesis, *MEF2A* and *MEF2C*, significantly reduce their activity in uterine (UCED), bladder (BLCA) and lung (LUSC) cancers. Supporting this observation, a significant down-regulation of *MEF2A* and *MEF2C* TFs was recently described in glioblastoma multiforme[44]. Actually, studies suggested that *MEF2C* is as target of *miR-223*[45], a miRNA known to promote the invasion of breast cancer cells[46].

Finally, other TFs display activations or deactivations shared by a few cancers and some of them present cancer-specific activities (See Fig. 3.1). Thus, *FOS* is activated in LIHC and THCA, or *FOSL1* and *FOSL2* are activated in KIRP, KIRC and THCA. Genes of the *FOS* family have been implicated as regulators of cell proliferation, differentiation, and transformation and are involved in many tumorigenic processes. Also *REST* gene, a transcriptional repressor that represses neuronal genes in non-neuronal tissues, is significantly activated in LIHC but significantly deactivated in COAD, maybe due to its dual role as a tumour suppressor and oncogene[47].

Regarding TFs specific of cancers, *JUNB*, with a known role in liver regeneration[48], but previously associated to different lymphomas such as Hodgkin[49] or cutaneous T-cell[50], seems to be also relevant in LIHC

Table 3.1. Cancer samples available for any cancer type selected.

Cancer type	Tumour	Normal	Stage I
Bladder Urothelial Cancer **[BLCA]**	294	17	1
Breast Cancer **[BRCA]**	1039	113	177
Colon Adenocarcinoma **[COAD]**	428	41	73
Head and Neck Squamous Cell Carcinoma **[HNSC]**	480	42	26
Kidney Renal Clear Cell Carcinoma **[KIRC]**	517	72	256
Kidney Renal Papillary Cell Carcinoma **[KIRP]**	222	32	138
Liver Hepatocellular carcinoma **[LIHC]**	294	48	132
Lung Adenocarcinoma **[LUAD]**	473	55	255
Lung Squamous Cell Carcinoma **[LUSC]**	426	45	217
Head and Neck Thyroid Carcinoma **[THCA]**	500	58	282
Uterine Corpus Endometrial Carcinoma **[UCEC]**	508	23	318

Continued on next page

tumorigenesis. Thyroid carcinoma (THCA) presents a quite atypical pattern of TF activation. While it lacks some ubiquitous TFs, such as *YY1, SP2, ZBTB33* or *NR2C2*, it presents significant activations in *HNF4A, RXRA* and *RXR::RAR_DR5*. Although HFN4A has traditionally been linked to diabetes, it has recently been suggested that this TF could be the link between ulcerative colitis and colorectal cancer[51] and it has even be proposed as a biomarker of this cancer[52] (colorectal cancer is not among the cancers included in this study). *RXR* and *RAR* are retinoid receptors that regulate cell growth and survival[53], which have been proposed as cancer therapeutic targets[54].

Table 3.1 – continued from previous page

Cancer type	Stage III	Stage IV	Alive	Deceased
Bladder Urothelial Cancer **[BLCA]**	99	98	221	73
Breast Cancer **[BRCA]**	237	17	937	98
Colon Adenocarcinoma **[COAD]**	120	58	374	53
Head and Neck Squamous Cell Carcinoma **[HNSC]**	72	245	320	158
Kidney Renal Clear Cell Carcinoma **[KIRC]**	125	81	358	159
Kidney Renal Papillary Cell Carcinoma **[KIRP]**	43	13	199	23
Liver Hepatocellular carcinoma **[LIHC]**	71	5	222	72
Lung Adenocarcinoma **[LUAD]**	81	24	355	118
Lung Squamous Cell Carcinoma **[LUSC]**	75	6	290	136
Head and Neck Thyroid Carcinoma **[THCA]**	110	52	481	14
Uterine Corpus Endometrial Carcinoma **[UCEC]**	114	27	464	43

Cancers can be grouped in three main clusters according to their TF activity patterns. One of them is composed of uterine (UCEC), bladder (BLCA), lung (LUAD and LUSC) and prostate adenocarcinoma (PRAD) cancers. Another, more dispersed cluster is composed of breast (BRCA), kidney papillary cell (KIRP) head and neck squamous cell (HNSC) and liver (LIHC) cancers. Although showing a regulatory behaviour quite different among them, kidney clear cell (KIRC) and head and neck thyroid (THCA) carcinomas cluster together. Colon adenocarcinoma (COAD) maps closer to the first cluster but seems to be an outlier in terms of TF activity pattern.

Some cancers, however, display atypical activity patterns of activity for several TFs. For example, COAD shows a specific significant activity decrease of *REST*, *CTCF* (a known chromatin insulator protein that may play a central role in mediating long-range chromatin interactions, whose deregulation has an increasingly important role in the epigenetic imbalance in cancer[55]), *EBF1* (identified as a tumour suppressor[56]) and *TCF12*. These regulatory differences might account, at least partially, for the different clinical behaviours of the distinct cancers analysed.

With respect to tissue of origin, both lung cancers, LUAD and LUSC, present quite similar TF activity profiles. Contrarily, kidney cancers KIRC and KIRP display remarkably different TF activity profiles. Interestingly, *FOSL1* and *FOSL2* TFs are specifically active almost uniquely in both cancers, while *FOXA1* is significantly inactive. In particular, *FOXA1*, a TF involved in the differentiation of the pancreas and liver, is known to be expressed in breast cancer[57] and others. Its remarkable down-activation in the two cancers originated in kidney could be part of the tumorigenesis in this organ.

It is worth noticing that, as previously mentioned, the expression level of TFs in the tissues according to The Protein Human Atlas database[58] is uncorrelated with the corresponding activity detected from the expression of the corresponding targets (Supplementary Figure 3.1). This reinforces the usefulness of this approach, given that the direct observation of TF expression would have not rendered detectable changes in their behaviours.

Supplementary Table 3.1 contains the complete list of p-values obtained for all the TFs in all the cancers studied.

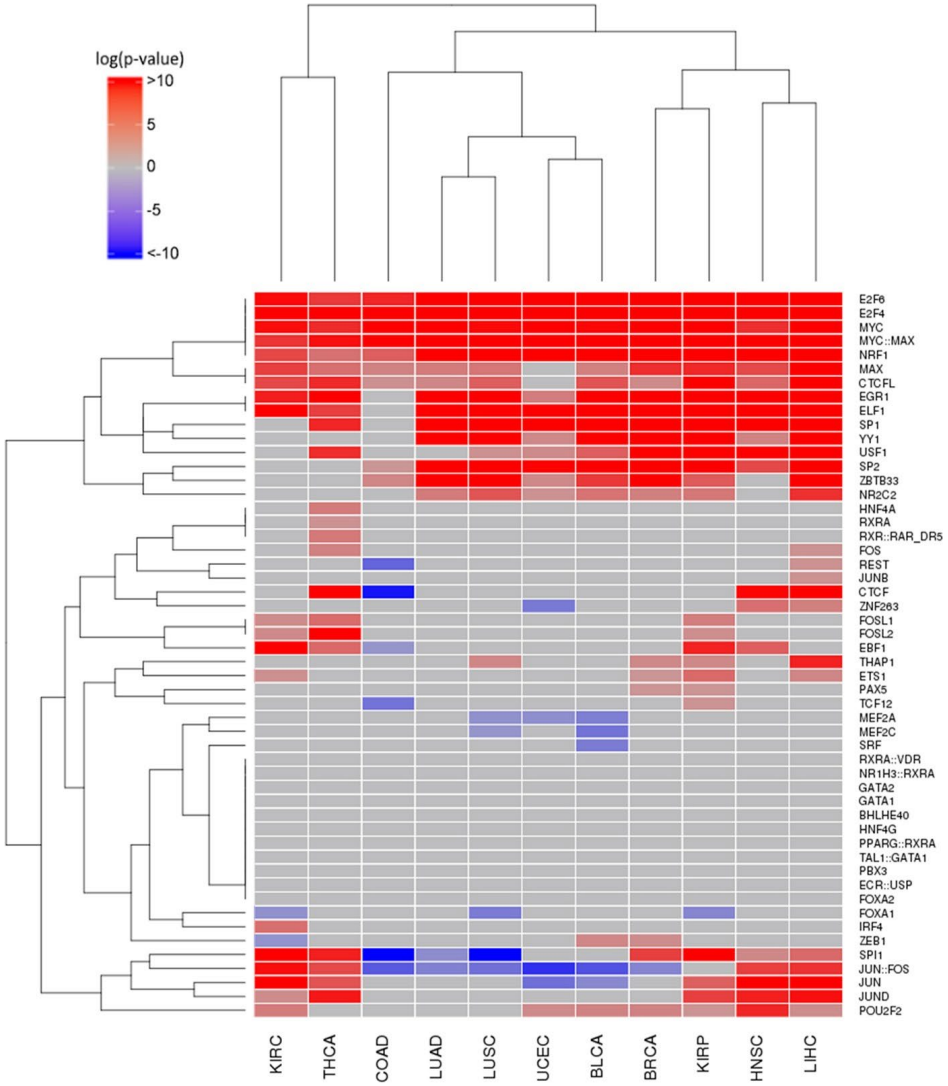


Figure 3.1. Change of TF activity in the different cancers studied. Cells in red indicate a significant increased activity of the TF in the cancer with respect to the corresponding normal tissue, according to the TFTEA, cells in blue indicate a significant decreased activity and cells in grey indicate that no significant change in activity was detected. Columns correspond to cancers and rows to TFs.

3.3.2 Changes in TF activity across cancer stages

The availability of clinical information, such as cancer stage allowed the stratification of cancer samples into their different stages. In any cancer type, the samples in any stage were compared to the corresponding normal samples. Figure 3.2 summarizes the changes in the activity status of TFs with respect to the normal situation in all the stages of the cancers analysed. Although the profiles of TF activity observed in this analysis are overall similar to the results produced by the comparison of cancer versus normal gene expressions, this analysis renders a more detailed picture of the changes in TF activity across stages in the different cancers. In fact, all the TFs present some activity change in some stages, even if this activity was not detected in the general cancer-control comparison. Thus, for example, *RXRA*, that was significantly active only in THCA in the cancer – control comparison, here presents a complex activation pattern across stages in BLCA as well. Other TFs, for which no significant change in the activity was previously found comparing cancer versus normal tissues, present however significant stage-specific activations, such as *PPARG::RXRA* activated in THCA, *TAL::GATA1*, down-activated in LUSC, *ECR::USP*, down-activated in several stages of LIHC and COAD.

Supplementary Table 3.1 contains the complete list of p-values obtained for all the comparisons of TF activities across stages in all the cancers studied.

3.3.3 TF activity and survival

The availability of survival data for the cancers analysed (Table 3.1) allows testing hypotheses on the contribution of distinct TF activities in the cancers to the disease outcome by validating their association with

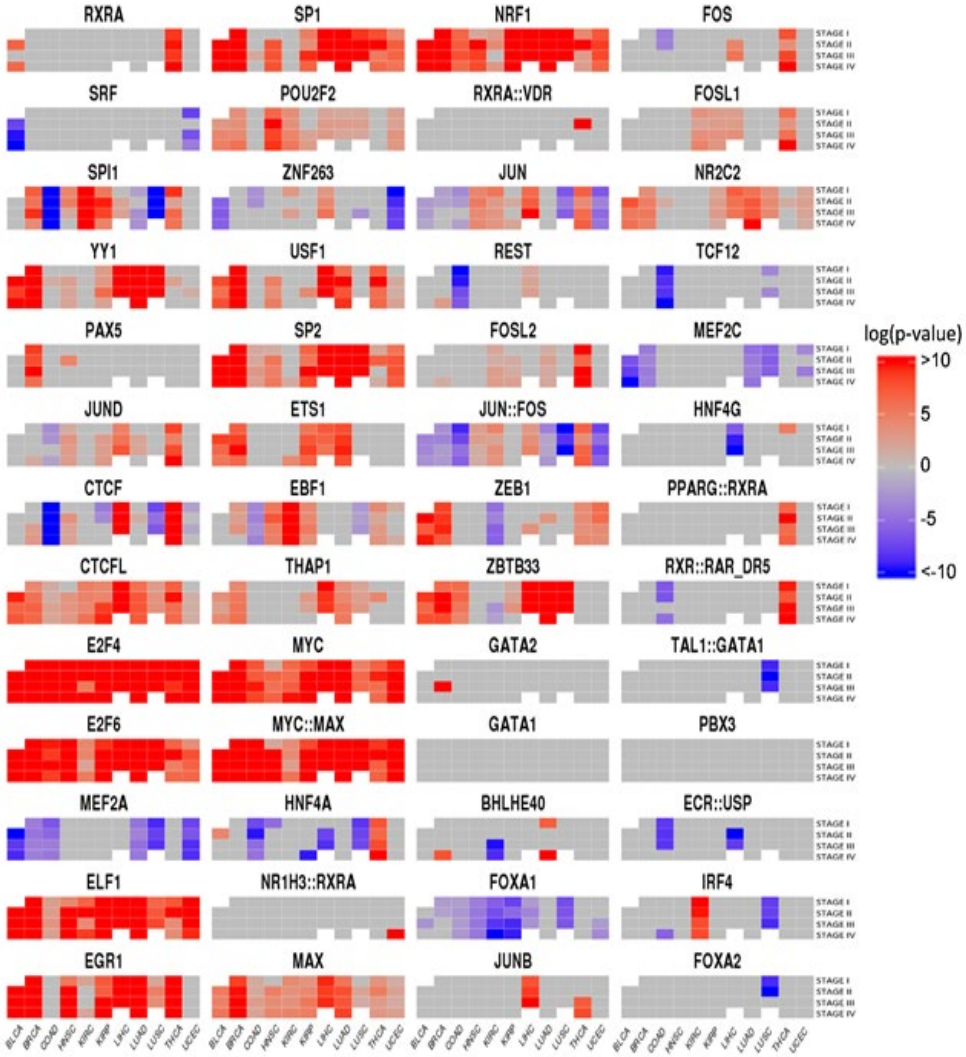


Figure 3.2. Change of activity in all TF included in this study across cancer stages in the different cancers studied. Each panel corresponds to a single TF, with stages in rows and cancers in columns. The colour scale in the figure ranges from red, indicating a significant increased activity of the TF in the stage of the cancer with respect to the corresponding normal tissue, according to the TFTEA, to blue, indicating a significantly decreased activity. The colour scale represents −log10 (adjusted p-value). Cells in grey indicate that no significant change in activity was detected. Cells in white correspond to stages in cancers with very few individuals (see Table 3.1) in which the analysis could not be carried out.

patient's survival. Since TFTEA can be applied in a personalized way to individual samples (see Methods) it is possible to know what TFs are active in any particular sample. Therefore, it is straightforward to test the relationship between TF activity and patient's survival using Kaplan-Meier (K-M) curves[59]. Figure 3.3 summarises the K-M plots representing TF activities significantly associated to patient survival (See detailed plots in Supplementary Figure 3.2 and Supplementary Table 3.2). As expected, more significant results were found in the cancers with more detailed data on survival, which are KIRC, BRCA and HNSC (See Table 3.1). A total of 19 TFs presented a strong significant (adjusted p-values < 0.05) association between its activity and patient survival in BRCA, HNSC and KIRC. The number of TFs in the figure increases to 92 if we consider significant nominal p-values, and cover all the cancers (Fig. 3.3, Table 3.2 and Supplementary Table 3.3). Some of the TFs highly associated to survival have been detected in the study of TF activity across cancers. Examples in KIRC are: *JUN:FOS,* known to be correlated to KIRK survival[28] and probably related to metastatic proliferation[42], *SPI1*, whose activation has been linked to survival in gastric cancer[41] in agreement with our observation (Fig. 3.4A), *JUND*, whose upregulation is significantly related to bad prognostic (Fig. 3.4B) and it has been described that can collaborate with NF-κB to increase antiapoptotic gene expression[60] and also *NRF1, EGR1, ETS1, ZEB1, MAX* and *FOSL1*. Previous studies of TF activity in KIRK reveal a number of them significantly correlated to survival[28]. Among the TFs that overlap with this study, *FOS, JUN::FOS, REST* and *TCF12* are found to be significantly related to survival, while *GATA1* did not reached the significance threshold. In HNSC, *JUND* and *ELF1* are differentially activated between the cancer and the normal tissue and also significantly associates to survival. In agreement with our results (Fig. 3.3, Table 3.2 and Supp. Table 3.3), it has already been described that *TAL1* was significantly correlated with breast cancer survival[27].

Table 3.2 TFs significantly associated to survival. The first column denoted the cancer type analysed. The second column contains the variables included in the Cox multiple regression model, which can be TFs and tumour purity. The third column contains the total number of TFs included in the Cox model. The fourth column shows TFs that show a significant association to survival by themselves. The fifth column contains the number of TFs significant in the K-M analysis. TFs in bold are significant with an adjusted p-value < 0.05. TFs in grey and in italic are significant with a nominal p-value < 0.05.

Cancer type	Variables (TFs and PURITY) selected by the Cox model	Total	TFs with individual effect on survival (K-M)	Total
BLCA	**SRF, YY1, CTCFL, POU2F2, ZNF263, USF1, EBF1, THAP1, MYC, MYC::MAX, NR1H3::RXRA, JUN, REST, JUN::FOS, GATA1, PPARG::RXRA,** *E2F6, ZEB1, BHLHE40, FOS, RXR::RAR_DR5, ECR::USP*	22	*EBF1, MEF2C, PAX5, SRF*	4
BRCA	**RXRA, E2F4, USF1, MYC::MAX, MAX, ZBTB33, FOXA1, PPARG::RXRA, RXR::RAR_DR5, TAL1::GATA1, IRF4**, *SPI1, YY1, JUND, SP2, ZEB1, GATA1, PURITY*	18	*EBF1, FOSL1, GATA2, TAL1::GATA1, ZBTB33*	5
COAD	**PAX5, CTCFL, E2F6, EGR1, THAP1, MYC, HNF4A, NRF1, JUN, JUN::FOS, ZEB1, ZBTB33, FOXA1, TCF12, HNF4G, PBX3**, *SPI1, JUND, CTCF, E2F4, PURITY*	21	*MYC, E2F6, EBF1, FOXA1, FOXA2, HNF4G, MAX*	7
HNSC	**SPI1, ELF1, EGR1, EBF1, RXRA::VDR, BHLHE40, RXR::RAR_DR5**, *JUN::FOS, GATA2*	9	**ELF1, JUND,** *FOS, JUN, MYC, CTCF, CTCFL, ETS1, FOSL1, FOSL2, JUN::FOS, NRF1, RXR::RAR_DR5, SP1, SP2, NR2C2, YY1*	17
KIRC	**RXRA, E2F6, MEF2A, ELF1, POU2F2, USF1, MYC, MYC::MAX, GATA1, MEF2C, HNF4G, RXR::RAR_DR5,**	17	**EGR1, JUN::FOS, JUN, MEF2A, NRF1, SPI1, SRF, USF1, YY1, ZEB1**, *FOS, MYC, CTCF, CTCFL, E2F4, E2F6, ELF1,*	38

Continued on next page

Table 3.2 – continued from previous page

Cancer type	Variables (TFs and PURITY) selected by the Cox model	Total	TFs with individual effect on survival (K-M)	Total
	FOXA2, *CTCFL, FOXA1, TCF12, PURITY*		*ETS1, FOSL1, FOXA2, HNF4G, IRF4, JUNB, JUND, MAX, MYC::MAX, REST, PAX5, PBX3, RXR::RAR_DR5, RXRA::VDR, SP1, SP2, TCF12, NR2C2, ZBTB33, ZNF263*	
KIRP	—	0	*RXRA*	1
LIHC	YY1, JUND, ELF1, USF1, THAP1, HNF4A, MAX, NRF1, ZBTB33, JUNB, MEF2C, HNF4G, TAL1::GATA1, RXRA, SP2, ETS1, NR1H3::RXRA, REST, ECR::USP, IRF4	20	*EGR1*	1
LUAD	PURITY, ELF1, SP2, MAX, REST, FOSL2, GATA1, NR2C2, JUND, E2F6, BHLHE40, JUNB	12	*JUND, MEF2C, TAL1::GATA1, TCF12*	4
LUSC	PURITY, E2F4, MEF2A, EGR1, SP1, POU2F2, ETS1, NR1H3::RXRA, NRF1, RXRA::VDR, JUN, JUN::FOS, FOSL1, NR2C2, ELF1, ZEB1, BHLHE40, HNF4G, PBX3	19	*FOS, ELF1, NRF1, REST*	4
THCA	—	0	*EBF1, ECR::USP, HNF4A, PAX5, RXRA, RXRA::VDR*	6
UCEC	PURITY, SRF, E2F6, ELF1, EGR1, SP1, USF1, SP2, MYC::MAX, HNF4A, FOSL2, ZEB1, MEF2C, HNF4G, TAL1::GATA1, ECR::USP, FOXA2, CTCF, E2F4, MAX, JUN::FOS, FOSL1	22	*E2F4, E2F6, ELF1, GATA2, HNF4G, PAX5*	6

Interestingly, there are TFs whose activity does not change significantly between the cancer and the normal tissue (see Fig. 3.1), but play, however, an unquestionable role in survival. These are the cases of *EBF1* in BRCA, which is a tumour suppressor[56] and its lower activity is associated to higher mortality or *CTCF* in KIRC[61] (see Fig. 3.4C). The case of *MEF2A* and *MEF2C* is similar: lower activity is significantly associated to worst prognostic (Fig. 3.4D), which is supported by the fact that its inhibition by miR-223 promotes the invasion of breast cancer cells[45,46]. These observations suggest that TFs whose activity is not especially relevant in the cancer tissue are however important in the determination of the prognostic of the patients and might be interesting therapeutic targets.

A complete list of p-values obtained for all the relationships of TF activities with survival in all the cancers studied can be found in Supplementary Table 3.3.

3.3.4 Combined contribution of TF activity to survival and the impact of tumour purity

Despite the obvious impact of individual TF activities in patient survival, it is clear that such a complex phenotype cannot be the effect of unique TF activities but rather will require of the interplay of several TFs. In order to capture at least part of the complexity of this interplay of TF activities that will ultimately affect patient survival we used a multivariate procedure. Conceptually, increasing levels of TF activity, as reported by TFTEA, accounts for higher expressions of increasingly larger number of targets of the TF. This continuous variable is modelled for multiple TFs with respect to the event of death in the patients by applying Cox multiple

regression models and using a stepwise algorithm (see details in Methods).

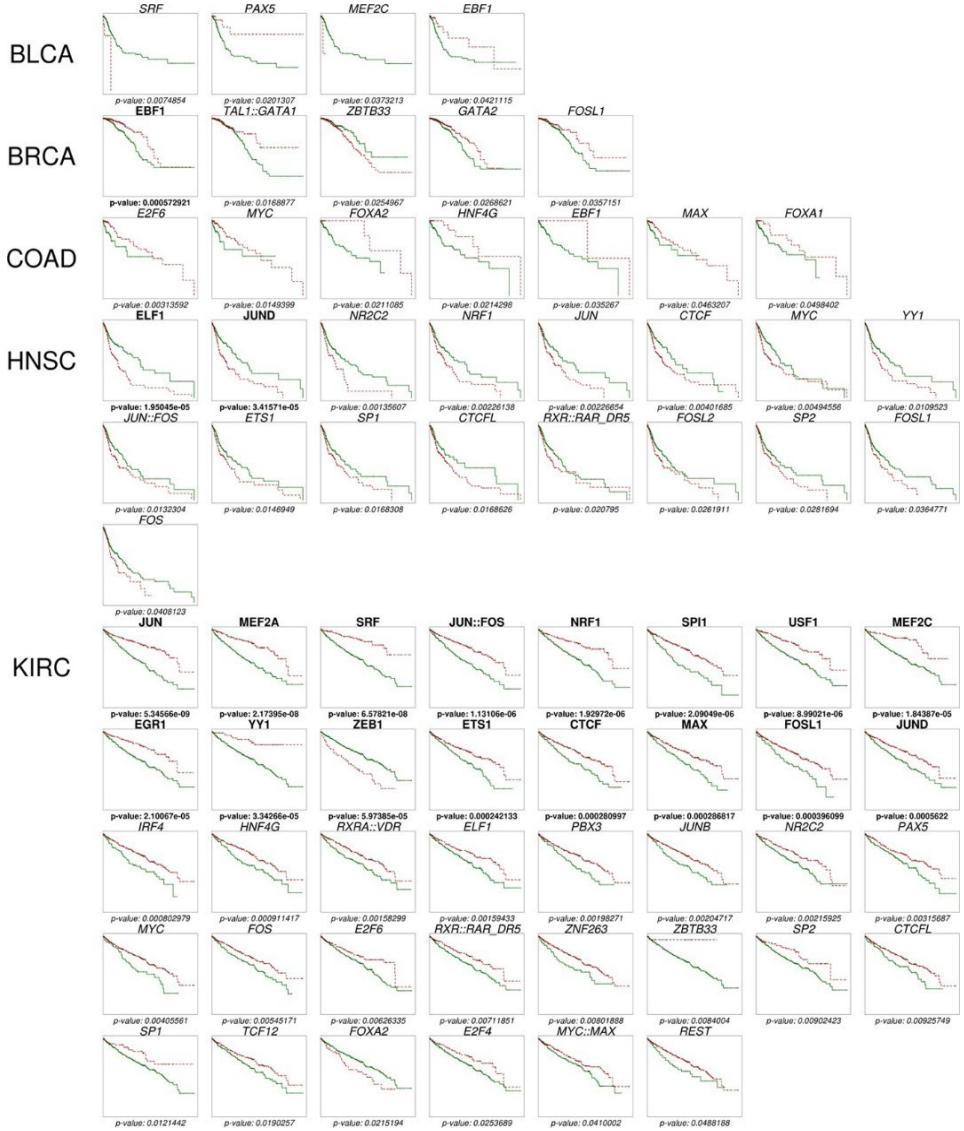
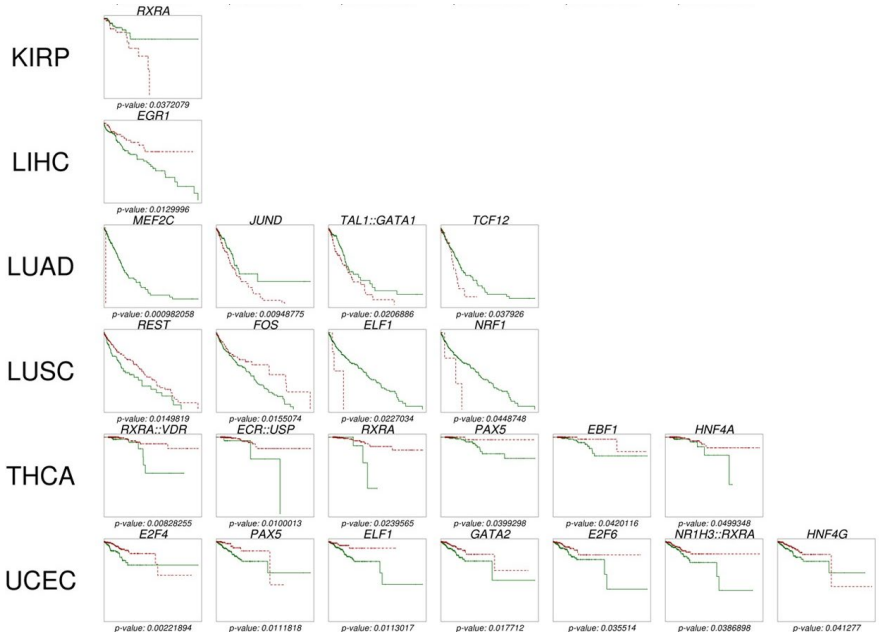


Figure 3.3. K-M plots representing TF activities significantly associated to patient survival in all the cancers analysed. TFs in bold present a significant association with adjusted p-values < 0.05 and TFs in italics have nominal p-values < 0.05. (*continued on next page*)

Figure 3.3 – continued from previous page



Recently, the importance that the non-cancerous components of the tumour (that include immune cells, fibroblasts, endothelial cells and normal epithelial cells) may have in cancer biology has been described[62]. Actually, it has been shown in some circumstances, the presence of these cells may alter the results of genomic analyses, including survival[62]. In order to check potential alterations in the TF activities inferred from the datasets studied here, we have compared the mean tumour purities with the outcome of the application of the method to see if there was any relationship between the mean purity of the cancer and the potential sensibility of the method in detecting TF activations (measured as the number of significant TF activity changes detected). Supplementary Figure 3.3 clearly shows that there is no observable trend between both variables, which strongly suggests that the application of the method to the analysed datasets was not significantly affected by the mean cancer

purity. However, the fact that TF activity estimations are not affected by tumour purity does not discard a possible confounding effect of the non-cancerous component of the cancer in this measurement. To study this potential confounding effect, the value of tumour purity was introduced in the Cox model as another variable.

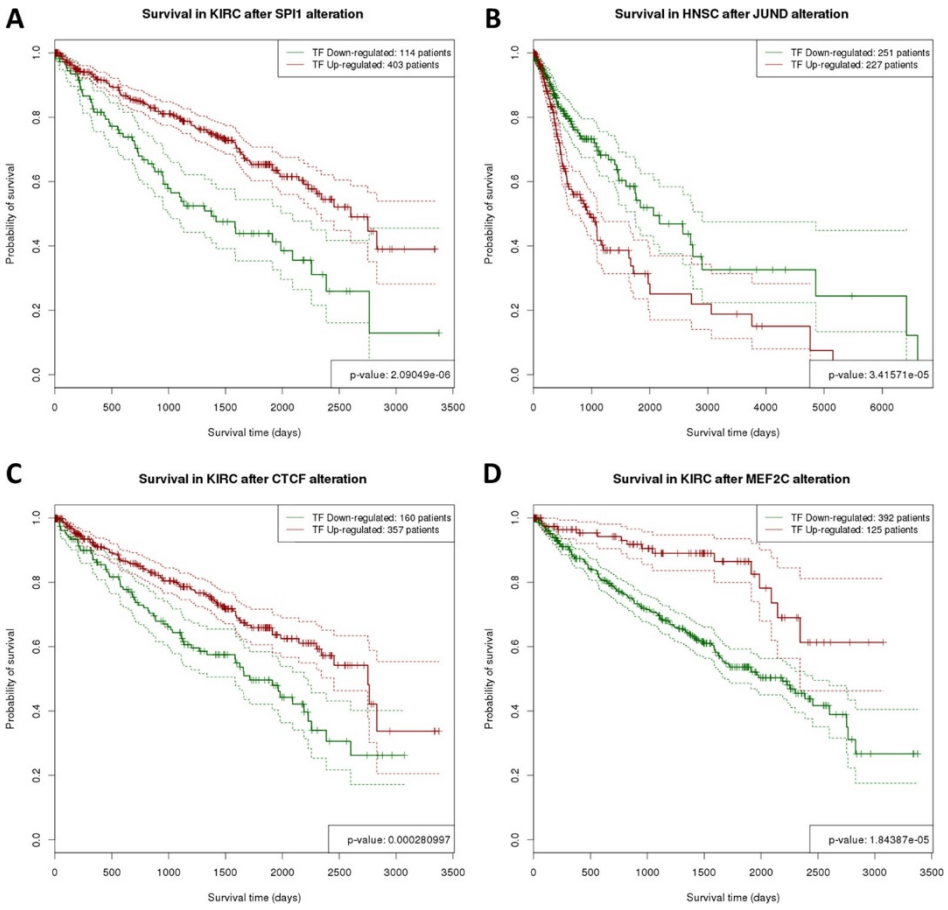


Figure 3.4. K-M plots representing TF activities significantly associated to patient survival. Survival curves are represented as solid lines and their corresponding confidence intervals as dotted lines. (A) High activity of *SPI1* in KIRC is significantly associated to patient survival; (B) High activity of *JUND* in HNSC is significantly associated to bad prognostic; (C) Low activity of *CTCF* in KIRC is significantly associated to bad prognostic; (D) Low activity of *MEF2C* in KIRC is significantly associated to bad prognostic.

The results obtained, listed in Table 3.2 and summarized in Fig. 3.5, clearly demonstrate a significant connection between multiple TF activity and patient survival for all the cancers analysed. The influence of TF activity in bad prognostic of the tumour seems to be a complex process in which different TF act cooperatively to activate (or deactivate) a large number of cell programmes that initiate and/or progress distinct cancer hallmarks[39] in the tumour cells. The results depict a relevant contribution of tumour purity to patient survival in three out of the eleven cancers analysed (both lung cancers LUAD and LUSC, and the endometrial carcinoma, UCEC). Although non-significant, tumour purity is still selected by the Cox model in another three cancers (BRCA, COAD and KIRK), where probably plays a more marginal role.

3.3.4 Potential limitations of the method

It must be taken into account that the information on TFBSs might contain a non-negligible number of false positives along with the true TFBSs in the real TF targets. This reduces the power of detection of the method given that, if a TF is activating their real targets, and a number of genes with random activity are considered to be part of the gene set of the TF, the complete gene set will show an activity lower than the actual activity. In addition, only a relatively low number of TFs are well characterized in terms of target genes they activate.

Another potential problem that can reduce the sensitivity of the analysis is the fact that many TFs need of a combination of factors to properly carry out transcription.

Finally, only KIRC (and to a lesser extent HNSC) had enough data on deceased patients to carry out robust survival analysis. Supplementary

Figure 3.4 clearly depicts this trend. The survival is best detected in KIRC and HNSC because a higher number of deceased patients is present in the dataset analysed (which also explains the high values of these cancers, unrelated with cancer purity, observed in Supplementary Figure 3.3).

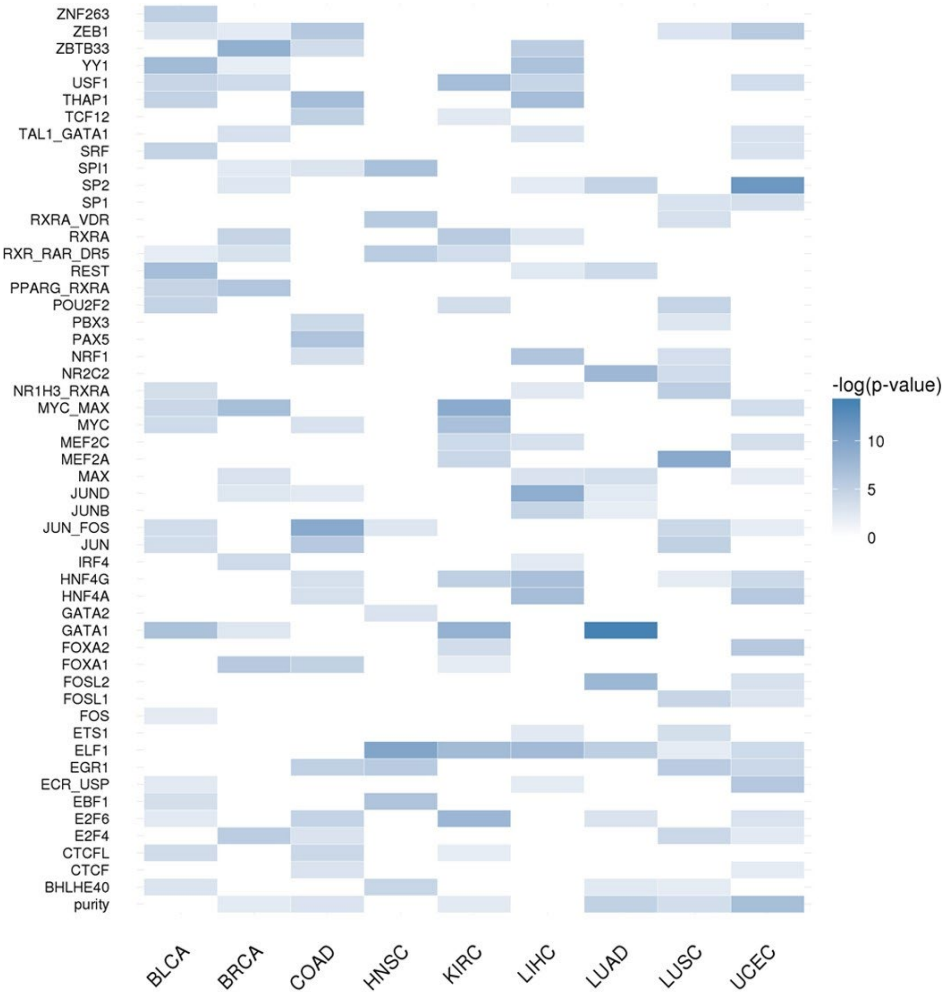


Figure 3.5. Combinations of TFs significantly associated to patient survival in the different cancers when a Cox model is applied. Cancers are represented in columns and TFs in rows. For each cancer, several TFs and sometimes tumour purity were included in a cox model. The colour intensity is related to the significance of this association (p-value).

In spite of these problems that reduce the potential of discovery of the proposed methodology in the current datasets, we have discovered a reasonable amount of significant associations of TF activity with cancer progression and with survival. Despite limited, the results obtained, which can be considered the "tip of the iceberg", and are quite encouraging.

3.4 CONCLUSIONS

The availability of survival and other relevant clinical data makes the analysis of pan-cancer Big Data repositories (ICGC and others) especially compelling, given that new unexpected associations of genomic data to relevant clinical outcomes can be found. Despite the relevance of regulation in cancer this seems to be the first pan-cancer analysis carried out to date. We have applied the TFTEA, a simple but robust methodology, to detect significant changes in the TF activity status when two groups of individuals are compared. In addition, the methodology also provides TF activity values per individual. This interesting property allows detecting TF-mediated deregulations specific for individuals, thus opening the door to possible personalized therapeutic interventions.

Regardless of the expectable reduction in the detection power that the current definitions of TF target gene sets could produce in methods that rely on this knowledge, the TFTEA still discovered a considerable number of significant associations between TF activity and the acquisition of cancer, the progression of cancer across stages or the survival of patients. Actually, many of the altered activities in TFs found were described in the literature either directly as causal alterations or, at least, linked to cancer, providing an extra support to the validity of the proposed methodology. Moreover, statistical modelling allowed detecting an important role of

tumour purity in survival. This suggests that, in some cases, the TF activity related to survival detected by the test could be due in part to other non-cancerous components of the tumour (probably immune cells, but also fibroblasts, endothelial cells and normal epithelial cells).

Actually, the findings of this work constitute most probably an underestimation of the total number of TFs linked to bad prognostic, due to the lack of enough survival data among the samples that precluded obtaining significant results for more TFs. This suggests that more detailed results would be obtained by the application of TFTEA to patient cohorts with richer clinical annotations.

3.5 METHODS

3.5.1 Cancer samples used

Eleven cancer types amounting to 5607 samples (Table 3.1) were selected on the basis of the simultaneous availability of paired samples (transcriptome analysis from both tumour sample and adjacent healthy tissue) and clinical data (tumour stage and survival). Raw read count data files were downloaded from the ICGC data portal[63] and clinical data were downloaded from the TCGA data portal[64] using sample IDs to cross-reference patient's data.

3.5.2 Gene expression data processing

The trimmed mean of M-values normalization (TMM)[65] was the method of choice and was applied using the edgeR package[66], using the default parameters. The differential expression analysis between cases and

controls was carried out using the limma package[67,68]. Firstly, the voom function[69] is applied to weight and transform TMM normalized values to make them suitable for lineal model analysis. Then, the lmFit function is used to adjust a lineal model and an empirical Bayes method is used to estimate differential expression values.

The Human Protein Atlas[58,70] was used as a reference for the gene expression levels of TFs in normal tissues.

3.5.3 Transcription factors used in the study

We have used a total of 52 TF available in ENSEMBL (GRCh38.p3), which are: *RXRA, SRF, SPI1, YY1, PAX5, JUND, CTCF, CTCFL, E2F4, E2F6, MEF2A, ELF1, EGR1, SP1, POU2F2, ZNF263, USF1, SP2, ETS1, EBF1, THAP1, MYC, MYC::MAX, HNF4A, NR1H3::RXRA, MAX, NRF1, RXRA::VDR, JUN, REST, FOSL2, JUN::FOS, ZEB1, ZBTB33, GATA2, GATA1, BHLHE40, FOXA1, JUNB, FOS, FOSL1, NR2C2, TCF12, MEF2C, HNF4G, PPARG::RXRA, RXR::RAR_DR5, TAL1::GATA1, PBX3, ECR::USP, IRF4* and *FOXA2*.

Any of these TFs activates a set of genes. Here we consider that a gene can potentially be activated by a TF if it includes possible binding sites for it, located between 5000 bp upstream from the most external transcription origin and the first exon. TFBSs have been mapped by Ensembl[71,72] along the genome. Briefly, for any TF which has both a ChIP-seq data and a JASPAR[73] publicly available position weight matrix (PWM), Ensembl annotates the position of putative TFBSs within the ChIP-seq peaks (details can be found in specific Ensembl web pages[74]). This information is accessible in a more efficient way in different publicly

available resources, such as CellBase[75], whose web services[76] were used here. Supplementary Table 3.4 shows the list of target genes for each TF.

3.5.4 Estimation of significant transcription factor activities in a cancer datasets

Since direct inference of TF activity from its own gene expression level is problematic, in this work we indirectly infer its activity from the collective activity of their gene targets. The method used here is an analysis of Gene Set Enrichment (GSE) that we call TF Target Enrichment Analysis (TFTEA). In this approach, each TF has an associated gene sets composed by the all their target genes (those containing a TFBS for the TF located between 5000 bp upstream from the transcription origin and the first exon of the gene).

Like other GSE methods, the TFTEA algorithm detects asymmetrical distributions of targets of TFs in the top (or the bottom) of a list of ranked genes. When two conditions are compared and the genes are ranked by differential expression (or fold change or any other related parameter), the detection of a significant accumulation of targets of a given TF in the upper (or lower) part of the ranked list indicates that such TF has significantly increased (or decreased) its activity in one of the conditions with respect to the other one. Here, differential expression in calculated by means of a limma test67, and the results of the statistic are used to define the ranked list of genes. A logistic regression is the most efficient methodology used for the detection of gene sets with a significant systematic over- or under-expressed77,78. Specifically, the association of a gene set composed by the targets of a specific TF to high or low values of the ranked list of genes is tested by means of the value of the slope of the logistic regression. The null hypothesis, slope = 0, is tested against the

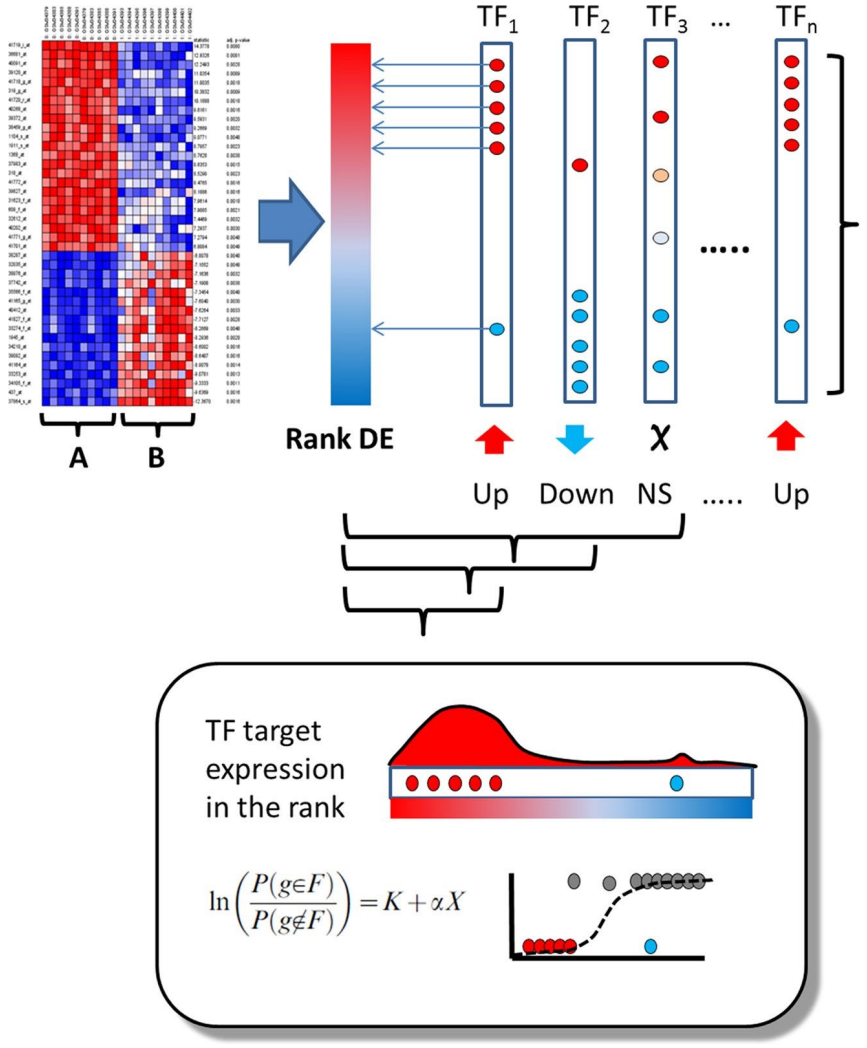


Figure 3.6 . Schema of the TFTEA method to obtain TFs differentially activated between two conditions compared. The method uses gene expression values and compares two conditions (**A** and **B**) by means of any test to obtain a rank of differentially expressed genes (Rank DE) based on the statistic. Then, for each TF, a logistic regression[78] is applied to discover associations of the TF targets to high or low values of the rank (lower panel). Thus, targets of TF_1 show a clear association to high values of the statistic, meaning that have significantly higher expression in condition (**A**) than in condition ( ), which demonstrated the differential activity of TF_1. TF_2 is the opposite case, in which the TF is significantly less active in (**B**) than in (**A**). TF_3 have their targets active or inactive in both conditions, meaning that these activities are not a collective property and consequently are not due to TF_3, but maybe to other regulators.

alternative slope ≠ 0 based on the maximum likelihood parameter estimates and the Wald test. For testing slope = 0, the Wald statistic can be shown to follow a chi-square distribution with one degree of freedom and the p-value is calculated assuming this null distribution[77,78].

Since many TFs were tested with the logistic regression across the eleven cancers, multiple testing effects need to be corrected- Here we have used the popular FDR method[79] for this purpose. Figure 3.6 schematizes the application of the method.

3.5.5 Estimation of personalized transcription factor activities per individual

We also used TFTEA to relate individual survival events to TF activity. Since this method requires of a ranked list of genes, each normalized patient sample needs to be compared to a reference. This reference value was obtained as the average normalized expression value across all the normal samples (see Fig. 3.7). For each gene (g) of any cancer sample (c), its expression value (v_{cg}) is compared with the corresponding average expression value for this gene in all the healthy samples (m_g) and the resulting value is divided by the standard deviation of the gene expression value in the healthy samples. This comparison provides for each cancer sample a value per gene that can be interpreted as a fold change (F_{cg}) with respect to the average healthy expression:

$$F_{cg} = \frac{v_{cg} - m_g}{SD_g}$$

F_{cg} values can thus be used to rank genes in a unique individual according to its relative expression with respect to the average expression values of their counterparts in a normal tissue.

Once a list of genes ranked by decreasing F_{cg} values is obtained for each patient, the TFTEA can be applied in a personalized manner to detect those TFs significantly activated (or deactivated) in each particular individual.

If samples are paired, the F_{cg} rank can be generated by direct comparison or each pair.

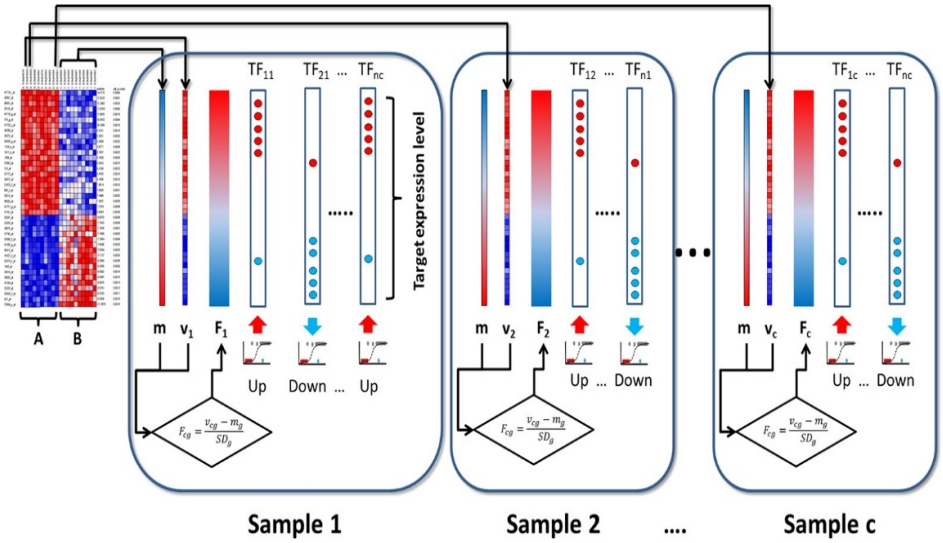


Figure 3.7. Schema of the TFTEA method to obtain personalized values of survival. The method uses gene expression values and compares two conditions (**A** and **B**). However, in this case all the samples in the (**B**) condition are used to produce an average expression value for any of the genes in this condition (m vector). Then, each sample in the (**A**) condition (v_c) can be compared to the average expression in the (**B**) condition and a rank of fold change (F_{cg}) is generated for each sample. Then, this ranking is used in the same way that the rank of differential expression was used in Fig. 3.6 to find differentially activated TFs in samples from the condition (**A**) with respect to the average (**B**) condition. Is samples are paired the F_{cg} value can be derived from the direct comparison between them.

3.5.6 Correlation between transcription factor activity and patient survival

Kaplan-Meier (K-M) curves[59] are used to relate TF activity to survival in the different cancers. The value of the statistic of each TF in each individual was used as a proxy of its activity. Only FT-cancer pairs with 10 events (deaths) or more were taken into account and the multiple testing adjustments were made taking into account only the pairs analysed. Calculations were carried out using the function survdiff from the survival R package[80].

Cox regression analysis[81] is used to relate combined TF activity to survival in the different cancers. Since tumour purity has been involved in survival[62], we used individual tumour purity values as an extra variable in the cox regression. Calculations were carried out using the function coxph from the survival R package[80]. A stepwise algorithm, implemented in the step function from the R package stats[82,83], is used to add or remove TFs or the tumour purity value according to the significance of their contributions to explain survival in the multiple regression model. In this way a final list of variables (TFs and cancer purity) whose combination is significantly related to survival is obtained. The step function uses Akaike Information Criterion (AIC) to select the best model by iteratively adding and removing variables.

3.5.7 Tumour purity estimation

There are different approaches to estimate tumour purity values, such as ESTIMATE, based on gene expression profiles of known immune stromal genes[84]; ABSOLUTE, based on somatic copy-number data[85]; LUMP

(leukocytes unmethylation for purity), based on averages of non-methylated immune-specific CpG sites[62].

Consensus measurement of purity estimations (CPE) is the median purity level after normalizing levels from all methods to give them equal means and s.d.'s (75.3 ± 18.9%).

Here, the per individual purity values provided in Supplementary Data 1 in the Aran's paper[62] are used to study the contribution of tumour purity to survival in the Cox regression.

3.5.8 Code availability

The code is open and available at: https://github.com/babelomics/TFTEA.

3.5.9 Supplementary data

Due to size requirements supplementary data can be found in the online version of this paper: https://www.nature.com/articles/srep39709#Sec19

3.6 REFERENCES

1. Hobert, O. Gene regulation by transcription factors and microRNAs. *Science* **319,** 1785–1786, doi: 10.1126/science.1151651 (2008).
2. Furney, S. J., Higgins, D. G., Ouzounis, C. A. & Lopez-Bigas, N. Structural and functional properties of genes involved in human cancer. *BMC Genomics* **7,** 3, doi: 10.1186/1471-2164-7-3 (2006).

3. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152,** 1237–1251 (2013).

4. Blancafort, P. *et al.* Genetic reprogramming of tumor cells by zinc finger transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 11716–11721 (2005).

5. Sakakura, C. *et al.* Frequent downregulation of the runt domain transcription factors RUNX1, RUNX3 and their cofactor CBFB in gastric cancer. *International journal of cancer* **113,** 221–228 (2005).

6. Gilliland, D. G. The Diverse Role of the ETS Family of Transcription Factors in Cancer Commentary re: B. Davidson, Ets-1 Messenger RNA Expression Is a Novel Marker of Poor Survival in Ovarian Carcinoma. Clin. Cancer Res., 7: 551–557, 2001. *Clinical Cancer Research* **7,** 451–453 (2001).

7. Strano, S. *et al.* Mutant p53: an oncogenic transcription factor. *Oncogene* **26,** 2212–2219 (2007).

8. Introna, M. & Golay, J. How can oncogenic transcription factors cause cancer: a critical review of the myb story. *Leukemia (08876924)* **13** (1999).

9. Darnell, J. E. Transcription factors as targets for cancer therapy. *Nature Reviews Cancer* **2,** 740–749 (2002).

10. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150,** 12–27 (2012).

11. Jang, I. S., Margolin, A. & Califano, A. hARACNe: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests. *Interface focus* **3,** 20130011, doi: 10.1098/rsfs.2013.0011 (2013).

12. Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7** Suppl 1, S7, doi: 10.1186/1471-2105-7-S1-S7 (2006).

13. Gao, F., Foat, B. C. & Bussemaker, H. J. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5,** 1 (2004).

14. Faith, J. J. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5,** e8, doi: 10.1371/journal.pbio.0050008 (2007).

15. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & Di Bernardo, D. How to infer gene networks from expression profiles. *Molecular systems biology* **3,** 78 (2007).

16. Roven, C. & Bussemaker, H. J. REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic acids research* **31,** 3487–3490 (2003).

17. Pournara, I. & Wernisch, L. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics* **8,** 61 (2007).

18. Min, J. H. *et al.* Structure of an HIF-1alpha -pVHL complex: hydroxyproline recognition in signaling. *Science* **296,** 1886–1889, doi: 10.1126/science.1073440 (2002).

19. Harris, M. L., Baxter, L. L., Loftus, S. K. & Pavan, W. J. Sox proteins in melanocyte development and melanoma. *Pigment Cell Melanoma Res* **23,** 496–513, doi: 10.1111/j.1755-148X.2010.00711.x (2010).

20. Filtz, T. M., Vogel, W. K. & Leid, M. Regulation of transcription factor activity by interconnected post-translational modifications. *Trends in pharmacological sciences* **35,** 76–85, doi: 10.1016/j.tips.2013.11.005 (2014).

21. Tootle, T. L. & Rebay, I. Post-translational modifications influence transcription factor activity: a view from the ETS superfamily. *Bioessays* **27,** 285–298, doi: 10.1002/bies.20198 (2005).

22. Cheng, C., Yan, X., Sun, F. & Li, L. M. Inferring activity changes of transcription factors by binding association with sorted expression profiles. *BMC Bioinformatics* **8,** 1 (2007).

23. Bleda, M. *et al.* Inferring the regulatory network behind a gene expression experiment. *Nucleic Acids Res* **40,** W168–172, doi: 10.1093/nar/gks573 (2012).

24. Zhu, M., Liu, C.-C. & Cheng, C. REACTIN: regulatory activity inference of transcription factors underlying human diseases with application to breast cancer. *BMC Genomics* **14,** 1 (2013).

25. Jiang, P., Freedman, M. L., Liu, J. S. & Liu, X. S. Inference of transcriptional regulation in cancers. *Proceedings of the National Academy of Sciences* **112,** 7731–7736 (2015).

26. Schacht, T., Oswald, M., Eils, R., Eichmuller, S. B. & Konig, R. Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics* **30,** i401–407, doi: 10.1093/bioinformatics/btu446 (2014).

27. Cheng, C., Li, L. M., Alves, P. & Gerstein, M. Systematic identification of transcription factors associated with patient survival in cancers. *BMC Genomics* **10,** 1 (2009).

28. Liu, Q., Su, P.-F., Zhao, S. & Shyr, Y. Transcriptome-wide signatures of tumor stage in kidney renal clear cell carcinoma: connecting copy number variation, methylation and transcription factor activity. *Genome medicine* **6,** 1–12 (2014).

29. *The COSMIC database*, http://cancer.sanger.ac.uk/cosmic (2015).

30. Gabay, M., Li, Y. & Felsher, D. W. MYC activation is a hallmark of cancer initiation and maintenance. *Cold Spring Harbor perspectives in medicine* **4,** doi: 10.1101/cshperspect.a014241 (2014).

31. Wolf, E., Lin, C. Y., Eilers, M. & Levens, D. L. Taming of the beast: shaping Myc-dependent amplification. *Trends Cell Biol* **25,** 241–248, doi: 10.1016/j.tcb.2014.10.006 (2015).

32. Kolch, W., Halasz, M., Granovskaya, M. & Kholodenko, B. N. The dynamic control of signal transduction networks in cancer cells. *Nat Rev Cancer* **15,** 515–527, doi: 10.1038/nrc3983 (2015).

33. Amati, B. & Land, H. Myc—Max—Mad: a transcription factor network controlling cell cycle progression, differentiation and death. *Current opinion in genetics & development* **4,** 102–108 (1994).

34. Nevins, J. R. The Rb/E2F pathway and cancer. *Hum Mol Genet* **10,** 699–703 (2001).

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

36. Khaleel, S. S., Andrews, E. H., Ung, M., DiRenzo, J. & Cheng, C. E2F4 regulatory program predicts patient survival prognosis in breast cancer. *Breast Cancer Res* **16,** 486 (2014).

37. Sui, G. The regulation of YY1 in tumorigenesis and its targeting potential in cancer therapy. *Molecular and Cellular Pharmacology* **1,** 157–176 (2009).

38. Archer, M. C. Role of sp transcription factors in the regulation of cancer cell metabolism. *Genes & cancer* **2,** 712–719 (2011).

39. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674, doi: 10.1016/j.cell.2011.02.013 (2011).

40. Vizcaíno, C., Mansilla, S. & Portugal, J. Sp1 transcription factor: A long-standing target in cancer chemotherapy. *Pharmacology & therapeutics* **152,** 111–124 (2015).

41. Wang, L. *et al.* Transcription factor Sp1 expression is a significant predictor of survival in human gastric cancer. *Clinical Cancer Research* **9,** 6371–6380 (2003).

42. Van Dam, H. & Castellazzi, M. Distinct roles of Jun: Fos and Jun: ATF dimers in oncogenesis. *Oncogene* **20** (2001).

43. Rimmelé, P. *et al.* Spi-1/PU. 1 oncogene accelerates DNA replication fork elongation and promotes genetic instability in the absence of DNA breakage. *Cancer research* **70,** 6757–6766 (2010).

44. Laddha, S. V. *et al.* Genome-wide analysis reveals downregulation of miR-379/miR-656 cluster in human cancers. *Biology direct* **8** (2013).

45. Johnnidis, J. B. *et al.* Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. *Nature* **451,** 1125–1129 (2008).

46. Yang, M. *et al.* Microvesicles secreted by macrophages shuttle invasion-potentiating microRNAs into breast cancer cells. *Molecular cancer* **10,** 1 (2011).

47. Weissman, A. M. How much REST is enough? *Cancer Cell* **13,** 381–383 (2008).

48. Hsu, J., Bravo, R. & Taub, R. Interactions among LRF-1, JunB, c-Jun, and c-Fos define a regulatory program in the G1 phase of liver regeneration. *Molecular and cellular biology* **12,** 4654–4665 (1992).

49. Mathas, S. *et al.* Aberrantly expressed c-Jun and JunB are a hallmark of Hodgkin lymphoma cells, stimulate proliferation and synergize with NF-κ B. *The EMBO journal* **21,** 4104–4113 (2002).

50. Mao, X. *et al.* Amplification and overexpression of JUNB is associated with primary cutaneous T-cell lymphomas. *Blood* **101,** 1513–1519 (2003).

51. Barrett, J. C. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nature genetics* **41,** 1330–1334 (2009).

52. Chellappa, K., Robertson, G. R. & Sladek, F. M. HNF4α : a new biomarker in colon cancer? *Biomarkers in medicine* **6,** 297–300 (2012).

53. Altucci, L., Leibowitz, M. D., Ogilvie, K. M., De Lera, A. R. & Gronemeyer, H. RAR and RXR modulation in cancer and metabolic disease. *Nature Reviews Drug Discovery* **6,** 793–810 (2007).

54. Altucci, L. & Gronemeyer, H. The promise of retinoids to fight against cancer. *Nature Reviews Cancer* **1,** 181–193 (2001).

55. Filippova, G. N. Genetics and epigenetics of the multifunctional protein CTCF. *Current topics in developmental biology* **80,** 337–360 (2007).

56. Liao, D. Emerging roles of the EBF family of transcription factors in tumor suppression. *Molecular Cancer Research* **7,** 1893–1901 (2009).

57. Badve, S. *et al.* FOXA1 expression in breast cancer—correlation with luminal subtype A and survival. *Clinical Cancer Research* **13,** 4415–4421 (2007).

58. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 394, doi: 10.1126/science.1260419 (2015).

59. Kaplan, E. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481, doi: 10.1080/01621459.1958.10501452 (1958).

60. Lamb, J. A., Ventura, J.-J., Hess, P., Flavell, R. A. & Davis, R. J. JunD mediates survival signaling by the JNK signal transduction pathway. *Molecular cell* **11,** 1479–1489 (2003).

61. Rasko, J. E. *et al.* Cell growth inhibition by the multifunctional multivalent zinc-finger factor CTCF. *Cancer research* **61,** 6002–6007 (2001).

62. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nature communications* **6** (2015).

63. *ICGC Data Portal*, https://dcc.icgc.org/ (2015).

64. *TCGA Data Portal*, https://tcga-data.nci.nih.gov/tcga/ (2015).

65. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11,** R25, doi: 10.1186/gb-2010-11-3-r25 (2010).

66. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140, doi: 10.1093/bioinformatics/btp616 (2010).

67. Ritchie, M. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* In press (2015).

68. Smyth, G. *Linear Models for Microarray Data*, https://bioconductor.org/packages/release/bioc/html/limma.html (2015).

69. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15,** R29, doi: 10.1186/gb-2014-15-2-r29 (2014).

70. *The Human Protein Atlas* http://www.proteinatlas.org/ (2015).

71. *Ensembl. Datasets and Data Processing, regulation sources*, http://www.ensembl.org/info/genome/funcgen/regulation_sources.html (2015).

72. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res* **43,** D662–669, doi: 10.1093/nar/gku1010 (2015).

73. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* **36,** D102–106, doi: 10.1093/nar/gkm955 (2008).

74. *Ensembl regulatory elements*, http://www.ensembl.org/info/genome/funcgen/regulatory_build.html#tfbs (2016).

75. Bleda, M. *et al.* CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. *Nucleic Acids Res* **40,** W609–614, doi: 10.1093/nar/gks575 (2012).

76. Medina, I. *The CellBase database*, http://wwwdev.ebi.ac.uk/cellbase/webservices/ (2015).

77. Sartor, M. A., Leikauf, G. D. & Medvedovic, M. LRpath: A logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics* **25,** 211–217 (2008).

78. Montaner, D. & Dopazo, J. Multidimensional gene set analysis of genomic data. *PLoS ONE* **5,** e10348, doi: 10.1371/journal. pone.0010348 (2010).

79. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57,** 289–300 (1995).

80. Therneau, T. *Survival Analysis*, https://cran.r-project.org/web/packages/survival/ (2015).

81. Cox, D. Regression models and life-tables. Journal of the Royal Statistical Society. *Series B (Methodological)* **34,** 187–220 (1972).

82. Ripley, B. *Choose a model by AIC in a Stepwise Algorithm*, https://stat.ethz.ch/R-manual/R-devel/library/stats/html/step.html (2015).

83. Venables, W. & Ripley, B. *Modern Applied Statistics with S*. (Springer, 2002).

84. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications* **4** (2013).

85. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30,** 413–421 (2012).

Capítulo 4

Mechanistic models of signaling pathways deconvolute the glioblastoma single-cell functional landscape

Falco, M. M., Peña-Chilet, M., Loucera, C., Hidalgo, M. R., & Dopazo, J. (2020). Mechanistic models of signaling pathways deconvolute the glioblastoma single-cell functional landscape. NAR Cancer, 2(2), zcaa011.

4. MECHANISTIC MODELS OF SIGNALING PATHWAYS DECONVOLUTE THE GLIOBLASTOMA SINGLE-CELL FUNCTIONAL LANDSCAPE

4.1 ABSTRACT

Single-cell RNA sequencing is revealing an unexpectedly large degree of heterogeneity in gene expression levels across cell populations. However, little is known on the functional consequences of this heterogeneity and the contribution of individual cell fate decisions to the collective behavior of the tissues these cells are part of. Here, we use mechanistic modeling of signaling circuits, which reveals a complex functional landscape at single-cell level. Different clusters of neoplastic glioblastoma cells have been defined according to their differences in signaling circuit activity profiles triggering specific cancer hallmarks, which suggest different functional strategies with distinct degrees of aggressiveness. Moreover, mechanistic modeling of effects of targeted drug inhibitions at single-cell level revealed, how in some cells, the substitution of *VEGFA*, the target of bevacizumab, by other expressed proteins, like *PDGFD*, *KITLG* and *FGF2*, keeps the *VEGF* pathway active, insensitive to the *VEGFA* inhibition by the drug. Here, we describe for the first time mechanisms that individual cells use to avoid the effect of a targeted therapy, providing an explanation for the innate resistance to the treatment displayed by some cells. Our results suggest that mechanistic modeling could become an important asset for the definition of personalized therapeutic interventions.

4.2 INTRODUCTION

Since the beginning of the century, transcriptomic technologies, which evolved from microarrays (1) to RNA sequencing (RNA-seq) (2), have provided an increasingly accurate insight into mRNA expression (3). The technological advances of RNA-seq technologies have increased the resolution in the quantification of transcripts until the unprecedented level of the mRNA component of individual single cells. The possibility of studying gene expression at the single-cell level opens the door to novel biological questions that were not possible using current tissue-level RNA-seq approaches. For example, single-cell RNA-seq (scRNA-seq) has allowed a high-resolution analysis of developmental trajectories (4,5), the detailed characterization of tissues (6,7), the identification of rare cell types (8) or the analysis of stochastic gene expression and transcriptional kinetics (9,10), just to cite a few cases.

The continuous publication of scRNA-seq studies is producing an increasingly large wealth of data on cell-level gene activity measurements under countless conditions. However, the functional consequences of such gene activity at single-cell level remains mostly unknown. Among the many methods and applications published for the management of scRNA-seq data (11), only a small proportion of them provide some functional insights on the results. For example, MetaNeighbor (12), SCDE (13) or PAGODA (14) annotates cell types based on conventional gene set enrichment analysis (15,16). Other algorithms, such as SCENIC (17), PIDC (18), SCODE (19) or SINCERITIES (20), offer the possibility of inferring regulatory networks as well. However, functional profiling methods have evolved from the analysis of simple gene sets or inferred regulatory gene networks to more sophisticated computational systems biology approaches that allow a mechanistic understanding on how

molecular cell signaling networks enable cells to make cell fate decisions that ultimately define a healthy tissue or organ, and how deregulation of these signaling networks leads to pathological conditions (21–23). Specifically, mechanistic models have helped to understand the disease mechanisms behind different cancers (24–27), rare diseases (28,29), the mechanisms of action of drugs (29,30) and other physiologically interesting scenarios such as obesity (31) or the postmortem cell behavior of a tissue (32). Although there are several mechanistic modeling algorithms available that model different aspects of signaling pathway activity, Hipathia (24) has been demonstrated to outperform other competing algorithms in terms of sensitivity and specificity (23).

Here, we propose the use of mechanistic models of signaling activities (24,33) that trigger cell functionalities related with cancer hallmarks (34), as well as other cancer-related relevant cellular functions to understand the consequences of gene expression profiles on cell functionality at single-cell level. Such mechanistic models use gene expression data to produce an estimation of activity profiles of signaling circuits defined within pathways (24,33). An additional advantage of mechanistic models is that they can be used not only to understand molecular mechanisms of disease or of drug action but also to predict the potential consequences of gene perturbations over the circuit activity in a given condition (35). Actually, in a recent work, our group has successfully predicted therapeutic targets in cancer cell lines with a precision of over 60% (25).

An interesting model to be studied from the viewpoint of mechanistic models is glioblastoma, the most common and aggressive of gliomas (36). The current treatment for glioblastoma includes maximal safe surgical resection followed by radiotherapy and chemotherapy (37), often combined with other drugs such as bevacizumab in an attempt to overcome resistances (38). Despite this intense treatment, the mean

survival of glioblastoma patients is only 15 months and resistances to the therapy are quite common (39–41). This high rate of treatment failure has been attributed to the lack of specific therapies for individual tumor types (42,43). Moreover, it is well known that glioblastoma tumors with a common morphological diagnosis display a high heterogeneity at the genomic level (44).

The availability of glioblastoma single-cell gene expression data (45) provides a unique opportunity to understand the behavior of a cancer type at the cell level. Here, we show for the first time how mechanistic models applied at single-cell level provide an unprecedentedly detailed dissection of the tumor into functional profiles at the scale of individual cells that throw new light on how cells ultimately determine its behavior. Moreover, since mechanistic models allow simulating interventions on the system studied, we show a comprehensive simulation of the potential effect of drugs at single-cell level that discloses, for the first time, the mechanisms and strategies used by subpopulations of cells to evade the effect of the drug.

4.2 MATERIALS AND METHODS

4.2.1 Data

A large scRNA-seq dataset containing 3589 cells of different types obtained in four patients from a glioblastoma study (45) was downloaded from GEO (GSE84465). Cells corresponded to the tumor, and to the periphery of the tumor.

4.2.2 Data imputation and primary processing

Count values for the scRNA-seq were downloaded from GEO. Since many of these data are affected by dropout events (13), they were subjected to the three imputation methods, MAGIC (46), scImpute (47) and DrImpute (48), as implemented in the corresponding software packages. Each method has its own preprocessing pipeline explained in the corresponding publication. The Rand index (49), which represents the frequency of occurrence of agreements of elements in the same cluster with respect to the random expectation, was used as an objective criterion for clustering comparison.

Once imputed, samples were log transformed and a truncation by quantile 0.99 was applied. Finally, the values were normalized between 0 and 1, as required by the downstream functional analysis with Hipathia.

4.2.3 Hipathia mechanistic model

The Hipathia method uses KEGG pathways (50) to define circuits that connect any possible receptor protein to specific effector proteins. Gene expression values are used in the context of these circuits to model signaling activity, which ultimately triggers specific cell activities, as described in (24). A total of 98 KEGG pathways involving a total of 3057 genes that form part of 4726 nodes were used to define a total of 1287 signaling circuits. The intensity value of signal transduced to the effector is estimated by the following recursive formula:

$$S_n = v_n \cdot (1 - \prod_{s_a \in A}(1 - s_a) \cdot \prod_{s_i \in I}(1 - s_i)$$

where Sn is the signal intensity for the current node n, vn is its normalized gene expression value, A is the set of activation signals (sa) arriving to the current node from activation edges and I is the set of inhibitory signals (si) arriving to the node from inhibition edges (24).

The Hipathia algorithm (27) is implemented as an R package available (https://bioconductor.org/packages/release/bioc/html/hipathia.html) in Bioconductor as well as at a web server (http://hipathia.babelomics.org/) and as a Cytoscape application (http://apps.cytoscape.org/apps/cypathia).

4.2.4 Differential signalling activity

Two groups of circuit activity profiles can be compared and the differences in activity of any circuit can be tested by means of different tests. Although non-parametric tests seem more adequate, and are suitable for small size studies, it has been noted that for larger sizes and, especially, when data display a highly skewed distribution, which is exactly this case, they tend to systematically give smaller P-values and parametric tests are preferable (51). In particular, limma (52), which has been demonstrated to be very efficient for gene expression data analysis, will be used.

4.2.5 Signaling circuits associated with cancer hallmarks

Each effector is known to be associated with one or several cell functions. This information is extracted from both the UniProt (53) and Gene Ontology (54) annotations corresponding to the effector gene (24). However, in some cases, the annotations are too ambiguous or refer to roles of the gene in many different conditions, tissues, developmental stages, etc., thus making it difficult to understand its ultimate functional role. In addition, in this study the activity of signaling circuits relevant in cancer is particularly interesting. Since a number of these effector genes have been related specifically with one or several cancer hallmarks (34)

in the scientific literature, the CHAT tool (55), a text mining-based application to organize and evaluate scientific literature on cancer, allows linking gene names with cancer hallmarks.

4.2.6 Subtyping of cancer cells

The SubtypeME tool from the GlioVis data portal (56) was used to obtain the subtype of cancer (classical, proneural or mesenchymal), based on the signature of 50 genes (57). This tool provides three methods to assign subtype: single-sample gene set enrichment analysis, K-nearest neighbors and support vector machine. Subtype was assigned when at least two of the methods made an identical subtype prediction. The subtyping tools use gene data without imputation.

4.2.7 Supplementary data

Due to size requirements supplementary data can be found in the online version of this paper:
https://academic.oup.com/narcancer/article/2/2/zcaa011/5862620#204971015

4.3 RESULTS

4.3.1 Selection of the optimal imputation method

Since mechanistic models consider the topology of signaling circuits to estimate signal transduction activity in the cell, the discrimination between genes with missing expression values and genes that are not expressed is crucial, given that, depending on the location of the gene within the circuit, it can play the role of a switch. Since dropout events (the observation of a gene at a moderate expression level in one cell that cannot be detected in another cell) are quite common in scRNA-seq

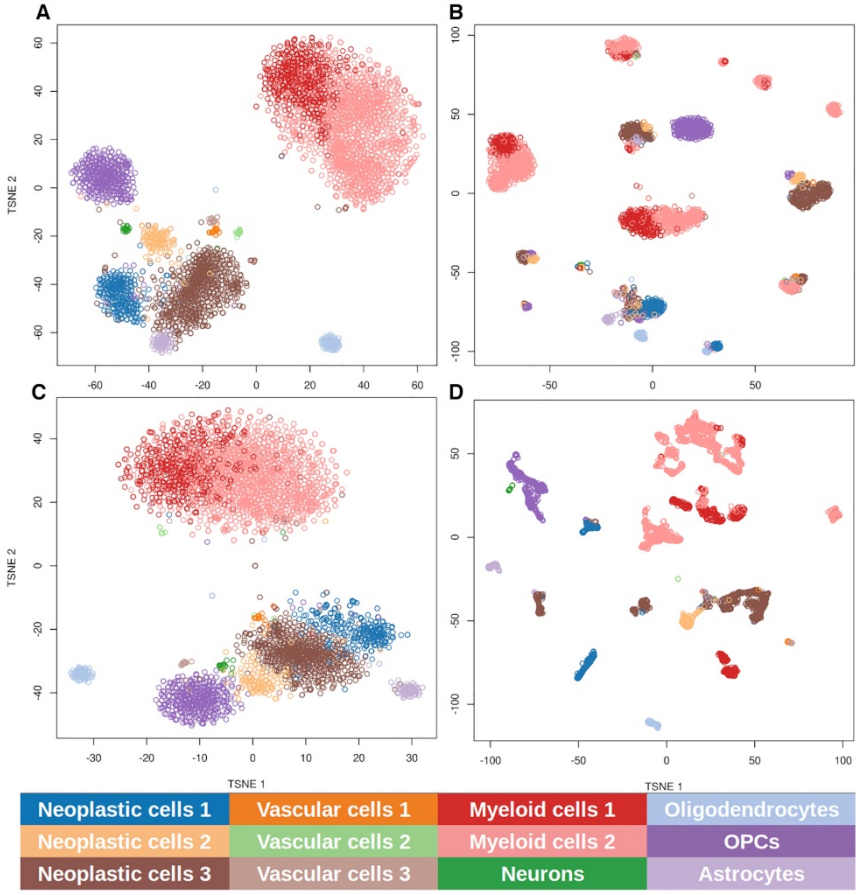


Figure 4.1. Clustering of the samples based on gene expression and signaling circuit activities obtained with different gene imputation methods. Data were subjected to t-SNE dimensionality reduction and the k-means clustering of the two main components is represented. (A) The clustering obtained with the gene expression values following the procedure described in the original glioblastoma study (45). Clustering obtained using all the circuit activities inferred using gene expression values imputed with (B) scImpute, which imputes 48% of the genes, (C) DrImpute, which imputes 85% of the genes, and (D) MAGIC, which makes the imputation over the whole set of genes. Cell types are labeled with colors.

experiments (13), and taking them as zero values can disturb the inferred activity of the circuit, the use of imputation methods is crucial for the application of the mechanistic model. Among the best performer

imputation methods available (58), three of them were checked to decide which one is optimal in the context of signaling pathway activity inference: MAGIC (46), scImpute (47) and DrImpute (48).

In order to decide which imputation method produced the most realistic results, we used the clustering produced by the highly expressed genes in the original single-cell glioblastoma study (45) as ground truth. There, the authors applied t-SNE (59) over the 500 most variable and highly expressed genes and then clustered the resulting data with k-means. They found 12 main clusters with a homogeneous cell composition that was further experimentally validated, which were astrocytes, two myeloid cell clusters, three neoplastic cell clusters, neurons, oligodendrocytes, oligodendrocyte progenitor cells and three vascular cell clusters. Then, gene expression values were imputed using the above-mentioned methods (MAGIC, scImpute and DrImpute). Next, gene expression values were used to infer signaling circuit activities with the Hipathia algorithm (24) as implemented in the Bioconductor application (https://bioconductor.org/packages/release/bioc/html/hipathia.html). The values of circuit activity were subjected to the same procedure (t-SNE dimensionality reduction and k-means clustering) and the resulting clusters were compared to the original ones obtained in the glioblastoma study using the Rand index (49). Figure 4.1A shows the clustering obtained with the genes following the procedure described above [equivalent to Figure 4.2 of the original study (45)], which can be compared with the clustering of the samples using the circuit activities obtained with the gene expression values imputed with scImpute (Figure 4.1B), DrImpute (Figure 4.1C) and MAGIC (Figure 4.1D). The comparison of the clusters obtained with the three imputation methods was follows: scImpute, 0.745; DrImpute, 0.852; and MAGIC, 0.858. Although MAGIC rendered a slightly better Rand index, DrImpute was

chosen as the imputation method because the dispersion of the clusters obtained was very similar to the one observed in the ground truth clustering (Figure 4.1A). The similarity in the clustering, which accounts for cell types, suggests that the imputation method is rendering values that result in coherent signaling circuit estimations.

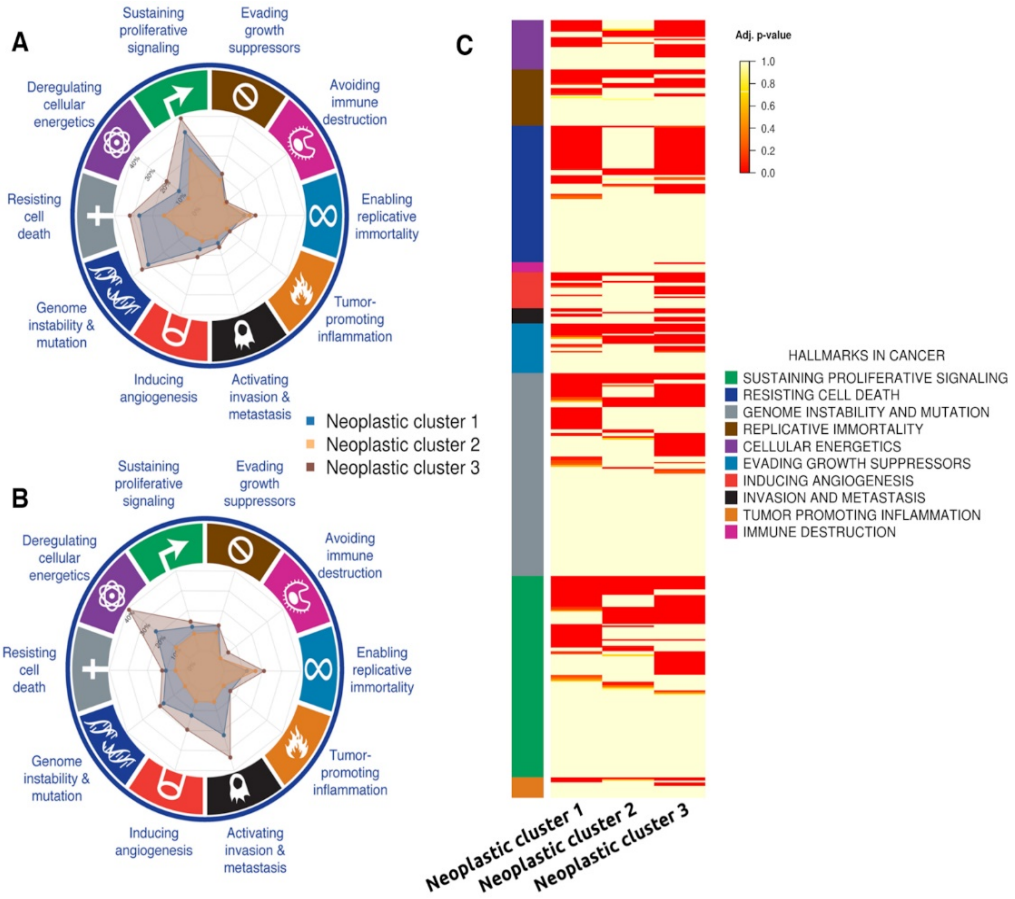


Figure 4.2. Circuits related to cancer hallmarks observed in the three neoplastic cell clusters. (A) Percentage of the total number of circuits with a significant differential activity in the neoplastic cells. The most internal division is 10% and every division increases a 10%. (B) Percentage of circuits with a differential activity with respect to the total number of circuits annotated to any of the cancer hallmarks. (C)Heatmap with the signaling circuits related to the different cancer hallmarks that have been found to be differentially activated in cells of each neoplastic cluster.

4.3.2 Functional characterization of cancer cells

Once verified that cell types defined by gene expression profiles (45) are supported by signaling profiles as well, the obvious comparison is the glioblastoma cell clusters versus the clusters composed by the different brain cells (oligodendrocytes, neurons, astrocytes and oligodendrocyte progenitor cells). It is interesting to note that normal cells, no matter which patient they were sampled from, display a similar functional profile; that is, the patients are intermingled within the clusters corresponding to any cell type. However, in the case of the neoplastic clusters, although some among-cluster overlap exists, their composition is mainly driven by the patient sampling origin (see Supplementary Figure 4.1). Since circuit activity bridges gene expression to signaling activity and ultimately cell functionality, an assessment of the differences between cell types from a functional perspective can be achieved by means of a differential cell activity statistical contrast. The cell functional responses triggered by the circuits differentially activated can be easily retrieved, and among them, those related with cancer hallmarks (34) can be identified using the CHAT tool (55), as explained in the 'Materials and Methods' section.

In order to detect which of the circuits display a significant change in activity, the three neoplastic cell clusters (1, 2 and 3 in Figure 4.1C) are compared to the normal brain cells (oligodendrocytes, oligodendrocyte precursor cells, astrocytes and neurons, labeled as O, OPC, A and N, respectively, in Figure 4.1C).

The comparison between the neoplastic clusters against the brain normal cells resulted in two different patterns of circuit activity: neoplastic clusters 1 and 3 present a higher number of signaling circuits differentially activated (309 and 336, respectively) than neoplastic cluster 2 (only 96 circuits; see Supplementary Table 4.1). Figure 4.2 represents the number

of differentially activated signaling circuits involved in cancer hallmarks observed in the three neoplastic cell clusters. This representation provides a summary of the strategy used by any particular neoplastic cluster in terms of the number of signaling circuits that control cell functionalities identifiable as cancer hallmarks. Figure 4.2A depicts the absolute number of circuits with a significant differential activity in the neoplastic cells and Figure 4.2B depicts the same results but as percentages with respect to the total number of circuits annotated to any of the cancer hallmarks. Table 4.1 summarizes the number of signaling circuits related to cancer hallmarks common to the three clusters (first column) and specific for each cancer type (subsequent columns). The common functional signature of this cancer is clearly driven by circuits related to 'Resisting cell death', 'Sustaining proliferative signaling' and 'Enabling replicative immortality' hallmarks, completed with circuits related to 'Evading growth suppressors', 'Inducing angiogenesis' and 'Tumor promoting inflammation' hallmarks. From Table 4.1 it becomes apparent that neoplastic clusters 1 and 3 are using a functional strategy different from that used by neoplastic cluster 2. The first two display a functional signature compatible with a more aggressive behavior: they have many extra circuits related to 'Resisting cell death' and 'Sustaining proliferative signaling' hallmarks but, in addition, both clusters have circuit activity related to 'Deregulation of cellular energetics', 'Genome instability and mutation' and 'Invasion and metastasis' hallmarks (see Table 4.1 and Figure 4.2C for details). It is also interesting to note from Figure 4.2C that the individual circuits involved in triggering the same functions are not exactly the same across the neoplastic clusters (Supplementary Table 4.2 lists details of the circuits involved in the figure). Conversely, neoplastic cluster 2 does not seem to have much more extra functional activity beyond the common functional signature, which suggests a less

aggressive character, especially because of the absence of circuit activity related to cell energetics or to invasion and metastasis.

Table 4.1. Summary of the different functional strategies followed by the different cells in the three neoplastic clusters in terms of the circuits differentially activated with respect to the normal tissue.

Cancer hallmark	Common circuits	Neoplastic cluster 1	Neoplastic cluster 3	Neoplastic cluster 2
Resisting cell death	4	13	14	2
Sustaining proliferative signaling	8	14	14	2
Deregulating cellular energetics		4	8	
Genome instability and mutation		5	6	
Inducing angiogenesis	2	2	5	
Enabling replicative immortality	5	1	3	
Activating invasion and metastasis		2	3	
Evading growth suppressors	3	3	2	
Tumor promoting inflammation	1	1	1	
Avoiding immune destruction			1	

4.3.3 Function-based stratification of glioblastoma cells

Neoplastic clusters have been defined according to the individual profiles of signaling circuit activities observed for each cell. The advantage of this way of cell stratification is that the functional profiles of each group are well defined. Current glioblastoma classification stratifies tumors into three subtypes, classical, proneural, and mesenchymal, from less to more aggressive, based on the signature of 50 genes (58). The SubtypeME tool from the GlioVis data portal (56) was used to assign subtype to each individual cell using this signature. Interestingly, when cells of the three neoplastic clusters are typed, the distribution of markers is very coherent with their functional activity profiles. Thus, neoplastic cluster 2 is mainly composed by cells belonging to the classical subtype (see Table 4.2), in coincidence with its functional profile being less aggressive. On the other hand, neoplastic cluster 1 has an important component of proneural cells, as well as a smaller proportion of mesenchymal cells, which is coherent with its more aggressive functionality triggered by its signaling activity, which includes modifications in circuits related to cell metabolism, genomic instability and metastasis. Moreover, the functional profile of neoplastic cluster 3 seems to be even more aggressive than that of neoplastic cluster 1. This group of glioblastoma cells not only has more circuits related to the same hallmarks as neoplastic cluster 1 but also has circuits that trigger functionalities for 'Avoiding immune destruction' (Table 4.2 and Figure 4.1). It is interesting to note that, although the conventional stratification in classical, mesenchymal and proneural classes is illustrative of the behavior of the cells, it does not completely fit with the stratification based on whole cell functional profiles.

Table 4.2. Distribution of the different glioblastoma subtypes across the three neoplastic cell clusters.

	Classical	Mesenchymal	Proneural	Total
Neoplastic cell cluster 1	92	44	135	271
Neoplastic cell cluster 2	107	3	15	125
Neoplastic cell cluster 3	540	141	14	697

4.3.4 Effect of a drug at single-cell level

Mechanistic models can be used to simulate the effect of an intervention over the system studied (25,35). Specifically, single-cell transcriptomic data offer, for the first time, the possibility of modeling the effects of a targeted drug at the level of individual cells.

The current indication for the treatment of glioblastoma patients is temozolomide, which induces DNA damage, that can be combined with other drugs such as bevacizumab to overcome resistances (38). Moreover, bevacizumab, which is indicated for several advanced cancer types, has recently been suggested for glioblastoma targeted treatment (60–62). Actually, the effect of bevacizumab, a humanized murine monoclonal antibody targeting the vascular endothelial growth factor ligand (*VEGFA*), can easily be simulated in the mechanistic model. *VEGFA* gene participates in six pathways ('*VEGF* signaling pathway', 'Ras signaling pathway', '*Rap1* signaling pathway', '*HIF-1* signaling pathway', '*PI3K–Akt* signaling pathway' and 'Focal adhesion pathway') and is part of 81 circuits, 39 of them directly related to cancer hallmarks (18 to 'Resisting cell death', 9 to 'Sustaining proliferative signaling', 4 to 'Genome

instability and mutation', 3 to 'Evading growth suppressors', 2 to 'Enabling replicative immortality', 2 to 'Inducing angiogenesis' and 1 to 'Deregulation of cellular energetics'). As described in the 'Materials and Methods' section, the inhibition of *VEGFA* can be simulated by taking the gene expression profile of a single cell, creating a simulated profile by setting the inhibited gene to a low value and comparing two profiles (24,35).

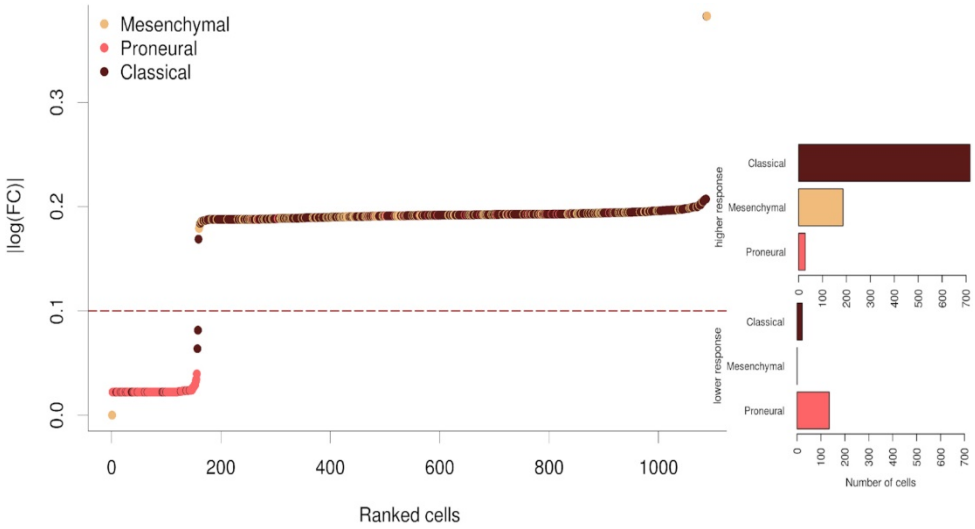


Figure 4.3. Impact of the inhibition of *VEGFA* by bevacizumab over the different neoplastic cells in terms of changes in the activities of signaling circuits in which this protein participates. The Y-axis depicts the magnitude of this change in the activities of signaling circuits. In the right part, two bar plots represent the proportion of the different cell types in the responder and non-responder groups.

Figure 4.3 shows the impact of the inhibition of *VEGFA* on the different cells in terms of changes in the activities of signaling circuits in which this protein participates. The Y-axis depicts the magnitude of this change in the activities of signaling circuits. There are clearly two different behaviors in the response: most of the cells present a drastic change in many signaling circuit activities (responders), while a smaller number of

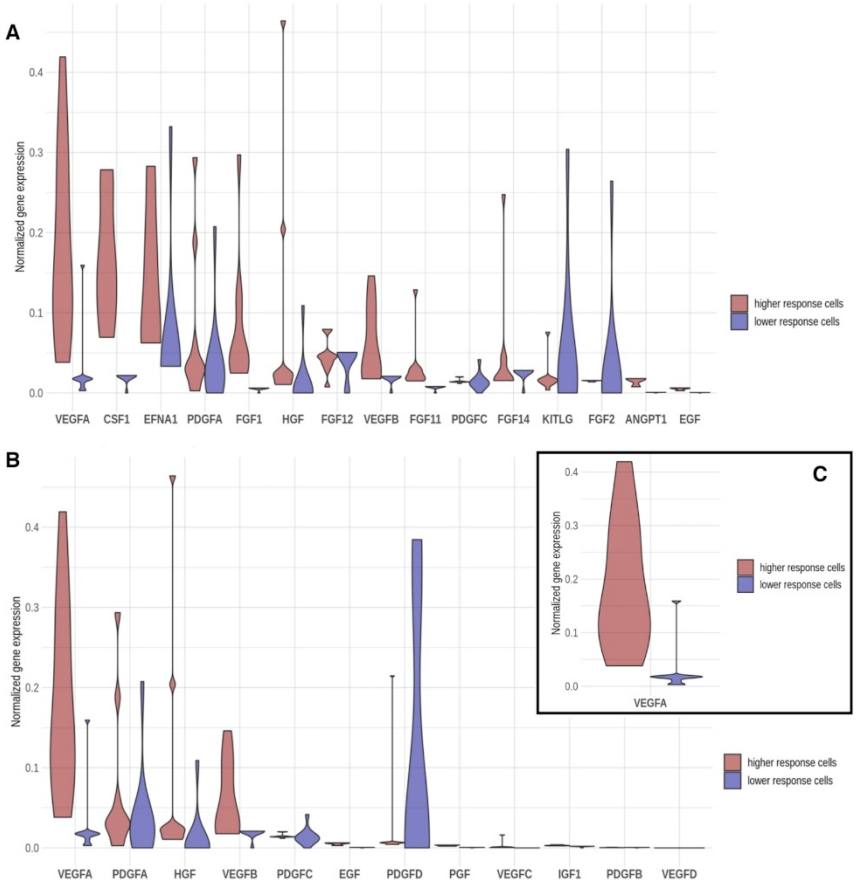


Figure 4.4. Distribution of the values of imputed and normalized gene expression values of the genes located within the effector node of the different signaling circuits affected by the bevacizumab inhibition. The distribution of observed expression levels in responder cells appears in red and in the low-responder cells in blue. (**A**) In the receptor node of the circuits within 'Ras signaling pathway', 'Rap1 signaling pathway' and 'PI3K–Akt signaling pathway', the *VEGFA* protein potentially shares the role of signal transducer with other 40 proteins (*CSF1, EFNA1, PDGFA, FGF1, HGF, FGF12, VEGFB, FGF11, PDGFC, FGF14, KITLG, FGF2, ANGPT1, EGF, PDGFD, EFNA5, ANGPT2, PGF, VEGFC, FGF18, EFNA3, FGF5, EFNA4, IGF1, EFNA2, FGF9, FGF13, FGF17, PDGFB, NGF, ANGPT4, FGF7, FGF22, FGF16, FGF23, FGF19, FGF20, FGF8* and *VEGFD*). (**B**) In the receptor node of the circuits within 'Focal adhesion pathway', the *VEGFA* potentially shares the signal transduction role with other 12 proteins (*PDGFA, HGF, VEGFB, PDGFC, EGF, PDGFD, PGF, VEGFC, IGF1, PDGFB* and *VEGFD*). (**C**) In the six signaling circuits belonging to the 'HIF signaling pathway' and '*VEGF* signaling pathway,' the protein *VEGFA* is the only signal transducer in the node.

Table 4.3. The eight drugs whose effect on the neoplastic population cell has been simulated

Drug name	Gene name	Action	Circuits	Invasion and metastasis	Immune destruction	Cellular energetics	Replicative immortality	Evading growth suppressors	Genome instability and mutation	Inducing angiogenesis	Resisting cell death	Sustaining proliferative signaling	Tumor promoting inflammation
Bevacizumab	*VEGFA*	Inhibitor	81			1	2	3	4	2	18	9	
Imiquimod	*TLR7*	Agonist	11		1					1	1		4
Sulfasalazine	*PTGS2*	Inhibitor	5						1				
Sulfasalazine	*PTGS1*	Inhibitor	2						1				
Sulfasalazine	*PPARG*	Agonist	21			2				1			
Pentoxifylline	*PDE4B*	Inhibitor	1										
Pentoxifylline	*ADORA1*	Antagonist	59						7		6	4	
Pentoxifylline	*PDE4A*	Inhibitor	1										
Pentoxifylline	*ADORA2A*	Antagonist	50						6	1	3	5	
Fenofibrate	*PPARA*	Agonist	32			1			1	1			
Doxepin	*HRH1*	Antagonist	4										
Doxepin	*HRH2*	Antagonist	0										
Quetiapine	*HTR2A*	Antagonist	13										
Quetiapine	*DRD2*	Antagonist	52						5		3	4	
Nilotinib	*ABL1*	Inhibitor	14								2		

them present a much lower affectation on them (low responders). It is interesting to note that the distribution of cell types between both groups is also asymmetric: the responder group is mainly composed of cells that have been typed as classical or mesenchymal, while the non-responder group is predominantly composed of proneural cells.

A close look at the consequences of the inhibition of *VEGFA* in different cells provides an interesting explanation for the observed differences. *VEGFA* is upstream in the chain of signal transduction in several circuits of different pathways. In the circuits within 'Ras signaling pathway', 'Rap1 signaling pathway' and 'PI3K–Akt signaling pathway', the *VEGFA* protein potentially shares the role of signal transducer with other 40 proteins. Figure 4.4A clearly depicts how the balance between the expression level of *VEGFA* in the responsive cells and *KITLG* and *FGF2* proteins, which can take a similar signaling role, changes. Supplementary Figure 4.2 shows the impact of the simulation of *VEGFA* inhibition in the 'PI3K–Akt signaling pathway', where the differences in the impact of this inhibition, measured as the log fold change in signaling activity, are remarkable between responder and low-responder cells. The inhibition of *VEGFA* in the responsive cells will radically inhibit the signal. However, the low-responder cells have already *VEGFA* at low expression levels and the signal is transmitted by *KITLG* and *FGF2* instead, which ultimately compromises the success of the drug. A similar scenario occurs with the 'Focal adhesion pathway', in which *VEGFA* shares the signal transduction role with other 12 proteins. In this case, the low-responder cells are characterized by a low level of *VEGFA* compensated with a high level of *PDGFD*, which makes these signaling circuits in the low-responder cells virtually insensitive to the inhibition of *VEGFA* (see Figure 4.4B). Only in the case of six signaling circuits belonging to the '*HIF* signaling pathway' and '*VEGF* signaling pathway', the protein *VEGFA* is the only

signal transducer in the node. In this case, low-responder cells have this circuit constitutively down and, consequently, are not affected by the inhibition (Figure 4.4C). Actually, except for *FGF2*, the genes potentially responsible for this switch presented a significant differential expression when responders were compared to low-responders [applying the limma (52) test, the FDR-adjusted P-values for *VEGFA*, *PGDFD*, *KITLG* and *FGF2* were, respectively, 6.4 × 10−4, 9.2 × 10−2, 1.39 × 10−2 and 5.59 × 10−1, using the gene expression values prior to the imputation). Supplementary Figure 4.3 represents the expression values of the same genes as in Figure 4.4 but with no imputation, showing no significant differences.

The same modeling strategy used with bevacizumab can be applied to simulate the effect of other drugs. Recently, a drug repurposing in silico experiment that combines human genomic data with mouse phenotypes has suggested the possible utility of a number of drugs with different indications (see Table 4.3) for potential glioblastoma treatment (63). The intensity and the degree of heterogeneity in the response are very variable across the eight drugs tested here. At the scale we tested the drugs, there are no correlations either between the number of genes targeted by the drug and the intensity of the effect or between the number of circuits potentially affected and the intensity of the effect. For example, pentoxifylline targets four proteins (*PDE4B*, *ADORA1*, *PDE4A* and *ADORA2A*) that participate in a total of 111 circuits and the fold change caused in the circuit activity after simulating its effect is comparatively low (log fold change <0.05 for all the cell types; see Supplementary Figure 4.4), while the simulation of the effect of fenofibrate, which targets only one protein *PPARA* that participates in only 32 circuits, renders a comparatively high effect (log fold change >2 for all the cell types; see Supplementary Figure 4.4). It is interesting to note that, depending on the

case, the different drug effects simulated can affect a larger or a smaller number of cells with distinct intensity in their impacts on the activity of the signaling circuits affected, but always, no matter which drug is simulated, there are some cells that manage to escape from the inhibitory effect of the drug.

4.4 DISCUSSION

The goal of most scRNA-seq publications revolves around the characterization of cell populations, which can be accurately achieved using only a subset of the total number of genes (those displaying the highest variability across cells). However, the use of mechanistic models to estimate global signaling circuit activity profiles for individual cells requires reasonably accurate measures of the expression levels of all the genes involved in the signaling circuits. Dropout events, quite common in scRNA-seq experiments (13), are particularly problematic given that taking by mistake a missing value by a real zero value can cause erroneous determinations of the inferred activity of the circuits. Thus, we explored the performance of three different imputation methods in producing cell-specific profiles of signaling circuit activity whose clustering resulted in a grouping similar to that observed and validated in the original glioblastoma study. Here, two machine learning-based methods, DrImpute and MAGIC, produced a clustering compatible with the original validated clustering, and specifically, DrImpute, the method of choice, rendered clusters with a similar shape as well (see Figure 4.1).

Focusing on neoplastic cells, the existence of three different clusters is also apparent at the level of functional profiles, which suggests the existence of different functional behaviors. Several attempts to stratify

glioblastoma patients have been proposed by discriminating different subtypes according to different properties, such as patient survival (64), mutational status of some genes (65) or the tumor microenvironment (66). In the most used classification, glioblastoma tumors were divided into three subtypes (from less to more aggressive: classical, proneural and mesenchymal) based on the signature of 50 genes (57). Although this conventional subtyping provides an approximate descriptor of tumor aggressiveness, subtyping based on functional profiles related to cancer hallmarks provides an interesting alternative for the stratification of glioblastoma that offers, in addition, a mechanistic description on the functional activity of the tumor. Actually, it has been reported in neuroblastoma that signaling pathway models used as biomarkers outperform traditional biomarkers as predictors of patient survival (26). Supplementary Figure 4.5 provides an interactive view of the circuits activated and deactivated within the different pathways.

Among pathways that are commonly altered in all three clusters, we find well-known factors contributing to carcinogenesis, such as those related to hypoxia (*HIF-1*, *SOD2*), cancer stem cells (CSCs), cell cycle proteins, like *CDK* family, signal transduction pathways and hormone signaling (67–69). Moreover, these are mainly related to 'Sustaining proliferative signaling', 'Enabling replicative immortality' and 'Resisting cell death' hallmarks that can be defined as the core cell functions involved in glioblastoma initiation and proliferation.

Each cluster exhibits a characteristic deregulation of pathways; however, cluster 2 barely has four unique sub-pathways, all related to common hallmarks 'Resisting cell death' and 'Sustaining proliferative signaling'. Neoplastic clusters 1 and 3, but not cluster 2, exhibit 'Genome instability', a hallmark observed in almost all sporadic human cancers, including glioblastoma (70,71). Besides, clusters 1 and 3 show deregulated 'Cellular

energetics', a process that has been suggested as a suitable target for tumor cell elimination (72). Furthermore, both clusters show pathways associated with matrix metalloproteins and Snail family that have been linked to cancer invasion and metastasis also in glioblastoma (73–75). Interestingly, only cluster 3 may be avoiding immune destruction due to the deregulation of 'Toll-like receptor signaling pathway'. Glioblastoma is known to have a strongly immunosuppressive microenvironment; thus, blocking these cells by activating downstream *TLR* signaling pathways can reduce tumor growth and disrupt CSC self-renewal (76,77).

We have demonstrated that not all the cells in a tumor are driven by the same cancer processes, and that those alterations can define subpopulations that may confer tumors different aggressiveness and invasion abilities, highlighting the relevance of heterogeneity, beyond the widely accepted stratification of glioblastoma in three/four subtypes (78,79).

The emergence of mechanisms of resistance in targeted therapies has been attributed to either the selection of rare pre-existing genetic alterations upon drug treatment (80) or the transient acquisition of a drug-refractory phenotype by a small proportion of cancer cells through epigenetic modifications (81). In both cases, these alterations would be detectable in the expression of the corresponding genes. An interesting property of mechanistic models is that they can be used to model the effect of an intervention over the system studied (25,35). Thus, the use of mechanistic models on single-cell transcriptomic data offers for the first time the possibility of modeling the effects of a targeted drug in individual cells. From Figure 4.4 it becomes apparent that low-responder cells have a constitutive level of *VEGFA* lower than high-responder cells. However, these cells maintain active the same *VEGFA*-activated pathways by the upregulation of others alternative *TRK* receptor ligands (i.e. *FGF2*,

PDGFD and *KITLG*) that have been implicated in the development and drug resistance in other cancer types (82–84). Interestingly, the switch in the expression levels of *FGF1* and *FGF2* (downregulated and upregulated in low-responder cells, respectively) as well as the upregulation of one member of the *PDGF* family (i.e. *PDGFD*) might be potentially driving the tumor progression in these GBM low-responder cells. Specifically, *FGF2* is the main member of the *FGF* family implicated in cancer development and drug resistance (85), and *PDGFD* and its receptor (*PDGFRB*) have been recently defined as key drivers of tumor progression since a *PDGFRB* downregulation impairs immediately GBM progression (A.C.V.-B. Fuentes-Fayos et al., submitted for publication). Moreover, although the relevance of *KITLG* has not been defined in GBM, the upregulation found in GBM low-responder cells might be linked to the drug resistance of these cells as has been reported in other tumor pathologies (84). In fact, this particular expression phenotype found in low-responder cells could be similar to that previously described in neural stem cell progenitors that are directly associated with the development and drug resistance in GBMs (86,87). Thus, the mechanistic model provides a simple potential interpretation of the molecular mechanisms behind the differential effect of drugs over cells with different signaling profiles that ultimately cause different functional strategies. Obviously, in order to gain insights into the true mechanisms driving cell resistance, further studies are needed in line with these findings. Nevertheless, we have proven that functional single-cell analyses, and the methodology here presented, are a helpful tool for discovering tumor heterogeneity, and the results can be applied by clinical community to forward tailored treatment, therefore improving patient's prognosis.

Supplementary Figure 4.4 depicts the simulated response of individual cells to the treatment with different targeted drugs. Despite the variety of

effects in the cells, it is worth noting that there is always a group of cells that manages to escape from the inhibition of the drug. The heterogeneity observed in the cell population in terms of use of different strategies to activate the essential cancer hallmark through different signaling circuits produces a consequent diversity in the response to drugs. Although the number of drugs simulated is relatively low, given that drug repurposing was beyond the scope of this paper, the results obtained here suggest that the escape of a relatively small number of cells from the effect of the drug could be a relatively frequent event that occurs as a natural consequence of the heterogeneity in the signaling strategies followed by the cell population. When this subpopulation becomes dominant some time later, it becomes resistant to the drug.

4.5 CONCLUSIONS

The use of mechanistic models provides a detailed insight into the functional strategies used by tumors to proliferate and open new avenues for the design of interventions à la carte. The extension of this analytical approach to single-cell transcriptomic data allows an unprecedented detail on how cancer cells display different functional strategies to proliferate that have consequences in their respective vulnerabilities to targeted therapies.

Although the existence of resistant clones in a tumoral cell population is well known, the specific mechanisms used by resistant cells to escape from the inhibitory effects of targeted therapies remain unknown yet. Mechanistic models offer for the first time a plausible and contrastable hypothesis on how and why some cells are insensitive to treatments, illustrated here with bevacizumab in the case of glioblastoma. Mechanistic modeling of effects of bevacizumab inhibitions at single-cell level revealed, how in some cells, with low *VEGFA* expression, the *VEGF*

pathway remains active because the initial signaling was assumed by other proteins like *PDGFD*, *KITLG* and *FGF2*, thus making the signaling circuit insensitive to the *VEGFA* inhibition by the drug.

The use of this modeling strategy offers a systematic way for detecting tumoral cells that may be resistant to specific targeted treatments. Conversely, the same models could be used to find an alternative treatment for resistant drugs. In fact, our results suggest that the search for new, more efficient therapeutic targets would be benefited by the use of mechanistic models that guide to the intervention points with more likelihood of success in inhibiting the proliferation of the largest possible part of the spectrum of functional strategies in the tumor cell ecosystem.

4.6 REFERENCES

1. Hoheisel,J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. Nat. Rev. Genet., 7, 200–210.

2. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., 10, 57–63.

3. Dopazo,J. (2014) Genomics and transcriptomics in drug discovery. Drug Discov. Today, 19, 126–132.

4. Bendall,S.C., Davis,K.L., Amir,E.-A.D., Tadmor,M.D., Simonds,E.F., Chen,T.J., Shenfeld,D.K., Nolan,G.P. and Peér,D. (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell, 157, 714–725.

5. Haghverdi,L., Buettner,M.,Wolf,F.A., Buettner,F. and Theis,F.J. (2016) Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods, 13, 845–848.

6. Zeisel,A., Muñoz-Manchado,A.B., Codeluppi,S., Lönnerberg,P., La Manno,G., Juréus,A., Marques,S., Munguba,H., He,L. and Betsholtz,C. (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science, 347, 1138–1142.

7. Aizarani,N., Saviano,A., Mailly,L., Durand,S., Herman,J.S., Pessaux,P., Baumert,T.F. and Grün,D. (2019) A human liver cell atlas reveals heterogeneity and epithelial progenitors. Nature, 572, 199–204.

8. Grün,D., Lyubimova,A., Kester,L., Wiebrands,K., Basak,O., Sasaki,N., Clevers,H. and van Oudenaarden,A. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature, 525, 251–255.

9. Kim,J.K. and Marioni,J.C. (2013) Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol., 14, R7.

10. Kar,G., Kim,J.K., Kolodziejczyk,A.A., Natarajan,K.N., Torlai Triglia,E., Mifsud,B., Elderkin,S., Marioni,J.C., Pombo,A. and Teichmann,S.A. (2017) Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. Nat. Commun., 8, 36.

11. Zappia,L., Phipson,B. and Oshlack,A. (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput. Biol., 14, e1006245.

12. Crow,M., Paul,A., Ballouz,S., Huang,Z.J. and Gillis,J. (2018) Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. Nat. Commun., 9, 884.

13. Kharchenko,P.V., Silberstein,L. and Scadden,D.T. (2014) Bayesian approach to single-cell differential expression analysis. Nat. Methods, 11, 740–742.

14. Fan,J., Salathia,N., Liu,R., Kaeser,G.E., Yung,Y.C., Herman,J.L., Kaper,F., Fan,J.-B., Zhang,K. and Chun,J. (2016) Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods, 13, 241–244.

15. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics, 20, 578−580.

16. 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. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. U.S.A., 102, 15545−15550.

17. Aibar,S., Gonz´alez-Blas,C.B.,Moerman,T., Huynh-Thu,V.A., Imrichova,H., Hulselmans,G., Rambow,F., Marine,J.-C., Geurts,P., Aerts,J. et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nat. Methods, 14, 1083−1086.

18. Chan,T.E., Stumpf,M.P.H. and Babtie,A.C. (2017) Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst., 5, 251−267.

19. Matsumoto,H., Kiryu,H., Furusawa,C., Ko,M.S.H., Ko,S.B.H., Gouda,N., Hayashi,T. and Nikaido,I. (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics, 33, 2314−2321.

20. Papili Gao,N., Ud-Dean,S.M.M., Gandrillon,O. and Gunawan,R. (2017) SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. Bioinformatics, 34, 258−266.

21. Fisher,J. and Henzinger,T.A. (2007) Executable cell biology. Nat. Biotechnol., 25, 1239−1249. 22. Fryburg,D.A., Song,D.H., Laifenfeld,D. and de Graaf,D. (2014) Systems diagnostics: anticipating the next generation of diagnostic tests based on mechanistic insight into disease. Drug Discov. Today, 19, 108−112.

23. Amadoz,A., Hidalgo,M.R., C¸ ubuk,C., Carbonell-Caballero,J. and Dopazo,J. (2019) A comparison of mechanistic signaling pathway activity analysis methods. Brief. Bioinform., 20, 1655–1668.

24. Hidalgo,M.R., Cubuk,C., Amadoz,A., Salavert,F., Carbonell-Caballero,J. and Dopazo,J. (2017) High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget, 8, 5160–5178.

25. Cubuk,C., Hidalgo,M.R., Amadoz,A., Pujana,M.A.,Mateo,F., Herranz,C., Carbonell-Caballero,J. and Dopazo,J. (2018) Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. Cancer Res., 78, 6059–6072.

26. Fey,D., Halasz,M., Dreidax,D., Kennedy,S.P., Hastings,J.F., Rauch,N., Munoz,A.G., Pilkington,R., Fischer,M., Westermann,F. et al. (2015) Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. Sci. Signal., 8, ra130.

27. Hidalgo,M.R., Amadoz,A., Cubuk,C., Carbonell-Caballero,J. and Dopazo,J. (2018) Models of cell signaling uncover molecular mechanisms of high-risk neuroblastoma and predict disease outcome. Biol. Direct, 13, 16.

28. Chacón-Solano,E., León,C., Díaz,F., García-García,F., García,M., Escámez,M., Guerrero-Aspizua,S., Conti,C., Mencía,A. and Martínez-Santamaría,L. (2019) Fibroblasts activation and abnormal extracellular matrix remodelling as common hallmarks in three cancer-prone genodermatoses. J. Br. J. Dermatol., 181, 512–522.

29. Esteban-Medina,M., Peña-Chilet,M., Loucera,C. and Dopazo,J. (2019) Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models. BMC Bioinformatics, 20, 370.

30. Amadoz,A., Sebastian-Leon,P., Vidal,E., Salavert,F. and Dopazo,J. (2015) Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity. Sci. Rep., 5, 18494.

31. Razzoli,M., Frontini,A., Gurney,A., Mondini,E., Cubuk,C., Katz,L.S., Cero,C., Bolan,P.J., Dopazo,J. and Vidal-Puig,A. (2016) Stress-induced activation of brown adipose tissue prevents obesity in conditions of low adaptive thermogenesis. Mol. Metab., 5, 19–33.

32. Ferreira,P.G., Muñoz-Aguirre,M., Reverter,F., Godinho,C.P.S., Sousa,A., Amadoz,A., Sodaei,R., Hidalgo,M.R., Pervouchine,D. and Carbonell-Caballero,J. (2018) The effects of death and post-mortem cold ischemia on human tissue transcriptomes. Nat. Commun., 9, 490.

33. Cubuk,C., Hidalgo,M.R., Amadoz,A., Rian,K., Salavert,F., Pujana,M.A., Mateo,F., Herranz,C., Caballero,J.C. and Dopazo,J. (2019) Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. npj Syst. Biol. Appl., 5, 7.

34. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. Cell, 144, 646–674.

35. Salavert,F., Hidalgo,M.R., Amadoz,A., Cubuk,C., Medina,I., Crespo,D., Carbonell-Caballero,J. and Dopazo,J. (2016) Actionable pathways: interactive discovery of therapeutic targets using signaling pathway models. Nucleic Acids Res., 44, W212–W216.

36. Consortium, T.G. (2018) Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis Consortium. *Neuro-Oncology*, **20**, 873–884.

37. Khasraw,M. and Lassman,A.B. (2010) Advances in the treatment of malignant gliomas. *Curr. Oncol. Rep.*, **12**, 26–33.

38. Bahadur,S., Sahu,A.K., Baghel,P. and Saha,S. (2019) Current promising treatment strategy for glioblastoma multiform: a review. *Oncol. Rev.*, **13**, 417.

39. Ostrom,Q.T., Gittleman,H., Xu,J., Kromer,C., Wolinsky,Y., Kruchko,C. and Barnholtz-Sloan,J.S. (2016) CBTRUS statistical report: primary brain and

other central nervous system tumors diagnosed in the United States in 2009–2013. *Neuro-Oncology*, **18**, v1–v75.

40. Helseth,R., Helseth,E., Johannesen,T., Langberg,C., Lote,K., Rønning,P., Scheie,D., Vik,A. and Meling,T. (2010) Overall survival, prognostic factors, and repeated surgery in a consecutive series of 516 patients with glioblastoma multiforme. *Acta Neurol. Scand.*, **122**, 159–167.
41. Siegel,R.L., Miller,K.D. and Jemal,A. (2017) Cancer statistics, 2017. *CA Cancer J. Clin.*, **67**, 7–30.
42. Omuro,A. and DeAngelis,L.M. (2013) Glioblastoma and other malignant gliomas: a clinical review. *JAMA*, **310**, 1842–1850.
43. Bai,R.-Y., Staedtke,V. and Riggins,G.J. (2011) Molecular targeting of glioblastoma: drug discovery and therapies. *Trends Mol. Med.*, **17**, 301–312.
44. Soeda,A., Hara,A., Kunisada,T., Yoshimura,S.-i., Iwama,T. and Park,D.M. (2015) The evidence of glioblastoma heterogeneity. *Sci. Rep.*, **5**, 7979.
45. Darmanis,S., Sloan,S.A., Croote,D., Mignardi,M., Chernikova,S., Samghababi,P., Zhang,Y., Neff,N., Kowarsky,M., Caneda,C. *et al.* (2017) Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.*, **21**, 1399–1410.
46. Moon,K.R., Stanley,J.S. III, Burkhardt,D., van Dijk,D., Wolf,G. and Krishnaswamy,S. (2018) Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.*, **7**, 36–46.
47. Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.
48. Gong,W., Kwak,I.-Y., Pota,P., Koyano-Nakagawa,N. and Garry,D.J. (2018) DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, **19**, 220.

49. Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

50. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.

51. Fagerland,M.W. (2012) *t*-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMCMed. Res.Methodol.*, **12**, 78.

52. Ritchie,M., Phipson,B., Wu,D., Hu,Y., Law,C., Shi,W. and GK,S. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

53. UniProt Consortium. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

54. Gene Ontology Consortium. (2018) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.

55. Baker,S., Ali,I., Silins,I., Pyysalo,S., Guo,Y., Högberg,J., Stenius,U. and Korhonen,A. (2017) Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics*, **33**, 3973–3981.

56. Bowman,R.L.,Wang,Q., Carro,A., Verhaak,R.G. and Squatrito,M. (2016) GlioVis data portal for visualization and analysis of brain tumor expression datasets. *Neuro-Oncology*, **19**, 139–141.

57. Wang,Q., Hu,B., Hu,X., Kim,H., Squatrito,M., Scarpace,L., deCarvalho,A.C., Lyu,S., Li,P., Li,Y. *et al.* (2017) Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell*, **32**, 42–56.

58. Zhang,L. and Zhang,S. (2020) Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **17**, 376–389.

59. van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

60. Kaka,N., Hafazalla,K., Samawi,H., Simpkin,A., Perry,J., Sahgal,A. and Das,S. (2019) Progression-free but no overall survival benefit for adult patients with bevacizumab therapy for the treatment of newly diagnosed glioblastoma: a systematic review and meta-analysis. *Cancers*, **11**, 1723.

61. Ohno,M., Miyakita,Y., Takahashi,M., Igaki,H., Matsushita,Y., Ichimura,K. and Narita,Y. (2019) Survival benefits of hypofractionated radiotherapy combined with temozolomide or temozolomide plus bevacizumab in elderly patients with glioblastoma aged ≥75 years. *Radiat. Oncol.*, **14**, 200.

62. Soubéran,A., Brustlein,S., Gouarné,C., Chasson,L., Tchoghandjian,A.,Malissen,M. and Rougon,G. (2019) Effects of VEGF blockade on the dynamics of the inflammatory landscape in glioblastoma-bearing mice. *J. Neuroinflammation*, **16**, 191.

63. Chen,Y. and Xu,R. (2016) Drug repurposing for glioblastoma based on molecular subtypes. *J. Biomed. Inform.*, **64**, 131–138.

64. Fatai,A.A. and Gamieldien,J. (2018) A 35-gene signature discriminates between rapidly- and slowly-progressing glioblastoma multiforme and predicts survival in known subtypes of the cancer. *BMC Cancer*, **18**, 377.

65. Zhou,X., Li,G., An,S., Li,W.-X., Yang,H., Guo,Y., Dai,Z., Dai,S., Zheng,J. and Huang,J. (2018) A new method of identifying glioblastoma subtypes and creation of corresponding animal models. *Oncogene*, **37**, 4781–4791.

66. Chen,Z. and Hambardzumyan,D. (2018) Immune microenvironment in glioblastoma subtypes. *Front. Immunol.*, **9**, 1004.

67. Gao,Y.-H., Li,C.-X., Shen,S.-M., Li,H., Chen,G.-Q., Wei,Q. and Wang,L.-S. (2013) Hypoxia-inducible factor 1_ mediates the down-regulation of superoxide dismutase 2 in von Hippel-Lindau deficient renal clear cell carcinoma. *Biochem. Biophys. Res. Commun.*, **435**, 46–51.

68. Feitelson,M.A., Arzumanyan,A., Kulathinal,R.J., Blain,S.W., Holcombe,R.F., Mahajna,J., Marino,M., Martinez-Chantar,M.L., Nawroth,R. and Sanchez-Garcia,I. (2015) In: *Seminars in Cancer Biology*, Vol. **35**, Elsevier, Amsterdam, pp. S25–S54.

69. Mayer,A., Schneider,F., Vaupel,P., Sommer,C. and Schmidberger,H. (2012) Differential expression of HIF-1 in glioblastoma multiforme and anaplastic astrocytoma. *Int. J. Oncol.*, **41**, 1260–1270.

70. Milinkovic,V., Bankovic,J., Rakic,M., Milosevic,N., Stankovic,T., Jokovic,M., Milosevic,Z., Skender-Gazibara,M., Podolski-Renic,A., Pesic,M. *et al.* (2012) Genomic instability and p53 alterations in patients with malignant glioma. *Exp. Mol. Pathol.*, **93**, 200–206.

71. Moon,J.J., Lu,A. and Moon,C. (2019) Role of genomic instability in human carcinogenesis. *Exp. Biol. Med.*, **244**, 227–240.

72. Maximchik,P.V., Kulikov,A.V., Zhivotovsky,B.D. and Gogvadze,V.G. (2016) Cellular energetics as a target for tumor cell elimination. *Biochemistry (Moscow)*, **81**, 65–79.

73. Deryugina,E.I. and Quigley,J.P. (2006) Matrix metalloproteinases and tumor metastasis. *Cancer Metastasis Rev.*, **25**, 9–34.

74. Wang,Y., Shi,J., Chai,K., Ying,X. and Zhou,B.P. (2013) The role of Snail in EMT and tumorigenesis. *Curr. Cancer Drug Targets*, **13**, 963–972.

75. Roomi,M.W., Kalinovsky,T., Rath,M. and Niedzwiecki,A. (2017) Modulation of MMP-2 and MMP-9 secretion by cytokines, inducers and inhibitors in human glioblastoma T-98G cells. *Oncol. Rep.*, **37**, 1907–1913.

76. Nduom,E.K., Weller,M. and Heimberger,A.B. (2015) Immunosuppressive mechanisms in glioblastoma. *Neuro-Oncology*, **17**(Suppl. 7), vii9–vii14.

77. Alvarado,A.G., Thiagarajan,P.S.,Mulkearns-Hubert,E.E., Silver,D.J., Hale,J.S., Alban,T.J., Turaga,S.M., Jarrar,A., Reizes,O., Longworth,M.S. *et al.* (2017) Glioblastoma cancer stem cells evade innate immune suppression of self-renewal through reduced TLR4 expression. *Cell Stem Cell*, **20**, 450–461.

78. Verhaak,R.G.W., Hoadley,K.A., Purdom,E., Wang,V., Qi,Y., Wilkerson,M.D.,Miller,C.R., Ding,L., Golub,T., Mesirov,J.P. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.

79. Friedmann-Morvinski,D. (2014) Glioblastoma heterogeneity and cancer cell plasticity. *Crit. Rev. Oncog.*, **19**, 327–336.

80. Boumahdi,S. and de Sauvage,F.J. (2019) The great escape: tumour cell plasticity in resistance to targeted therapy. *Nat. Rev. Drug Discov.*, **19**, 39–56.

81. Shaffer,S.M., Dunagin,M.C., Torborg,S.R., Torre,E.A., Emert,B., Krepler,C., Beqiri,M., Sproesser,K., Brafford,P.A. and Xiao,M. (2017) Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, **546**, 431–435.

82. Miura,K., Oba,T., Hamanaka,K. and Ito,K.-I. (2019) FGF2-FGFR1 pathway activation together with thymidylate synthase upregulation is induced in pemetrexed-resistant lung cancer cells. *Oncotarget*, **10**, 1171–1192.

83. Satow,R., Nakamura,T., Kato,C., Endo,M., Tamura,M., Batori,R., Tomura,S., Murayama,Y. and Fukami,K. (2017) ZIC5 drives melanoma aggressiveness by PDGFD-mediated activation of FAK and STAT3. *Cancer Res.*, **77**, 366–377.

84. Smith,B.D., Kaufman,M.D., Lu,W.-P., Gupta,A., Leary,C.B., Wise,S.C., Rutkoski,T.J., Ahn,Y.M., Al-Ani,G. and Bulfer,S.L. (2019) Ripretinib (DCC-2618) is a switch control kinase inhibitor of a broad spectrum of oncogenic and drug-resistant KIT and PDGFRA variants. *Cancer Cell*, **35**, 738.e9–751.e9.

85. Cao,R., Brakenhielm,E., Pawliuk,R., Wariaro,D., Post,M.J., Wahlberg,E., Leboulch,P. and Cao,Y. (2003) Angiogenic synergism, vascular stability and improvement of hind-limb ischemia by a combination of PDGF-BB and FGF-2. *Nat. Med.*, **9**, 604–613.

86. Noch,E.K., Ramakrishna,R. and Magge,R. (2018) Challenges in the treatment of glioblastoma: multisystem mechanisms of therapeutic resistance. *World Neurosurg.*, **116**, 505–517.

87. Wick,W., Osswald,M., Wick,A. and Winkler,F. (2018) Treatment of glioblastoma in adults. *Ther. Adv. Neurol. Disord.*, **11**, doi:10.1177/1756286418790452.

Capítulo 5

Discusión general

5. DISCUSIÓN GENERAL

5.1 CARACTERIZACIÓN DEL ESTADO DE ACTIVACIÓN DE ELEMENTOS REGULADORES

Los mecanismos regulatorios confieren a la célula la capacidad de adaptarse a los cambios que ocurren dentro y fuera de la misma. Estos mecanismos también se encargan de regular el ciclo celular y decidir cuándo una célula puede dividirse. El entramado que conforman los diferentes elementos regulatorios forma un complejo equilibrio cuya alteración puede desencadenar varias enfermedades. En cáncer, la detección de los reguladores perturbados que tienen un efecto en la enfermedad es crucial para entender sus mecanismos de inicio y desarrollo, así como para el diagnóstico o la toma de decisiones terapéuticas. Por ello, esta tesis se ha centrado en el estudio de 2 elementos reguladores: los factores de transcripción y las rutas de señalización. Sin embargo, para la caracterización de sus niveles de activación hubo que hacer frente a ciertas problemáticas que presentan tanto la gran complejidad biológica inherente a la regulación celular como las herramientas disponibles actualmente.

Los niveles de expresión de los factores de transcripción no constituyen una medida fidedigna de su estado de activación debido a que, además de tener una expresión basal fluctuante y baja, pueden no ser representativos de los correspondientes niveles de proteína funcional (Lemon y Tjian, 2000). Además, el nivel de expresión de proteínas de un factor de transcripción no necesariamente indica que vaya a haber actividad transcripcional por parte del mismo debido las modificaciones post-traduccionales que pueden tener y a que, para la expresión de un cierto

gen, no basta la unión del factor de transcripción a su zona promotora, también es necesaria la participación de ciertos cofactores (Reiter et al., 2017). Por esta razón se estimó la actividad transcripcional efectiva de los factores de transcripción a partir de la expresión de los genes a los que regulan mediante una regresión logística. Una complicación a la hora de emplear la metodología desarrollada consistió en determinar cuáles eran los genes regulados por los distintos factores. Existen varios métodos para identificar las asociaciones entre factores de transcripción y sus genes diana. Uno de los más extendidos consiste en la identificación en el genoma de sitios de unión de factores de transcripción (TFBS, por sus siglas en inglés) mediante inmunoprecipitación de la cromatina acoplado a la secuenciación de los fragmentos aislados, técnica conocida como ChIP-seq (Johnson et al., 2007). Empleando estas técnicas, se han desarrollados proyectos importantes para el estudio de regiones reguladoras, como ENCODE, donde se han agrupado y estandarizado distintos análisis experimentales, computacionales y estadísticos para descubrir genes, tránscritos, proteínas de unión al DNA, incluidos factores de transcripción, que interaccionan con las regiones reguladoras del genoma (ENCODE, 2011). Sin embargo, a pesar de estos grandes esfuerzos, la detección de los genes diana para un determinado factor de transcripción sigue siendo una tarea complicada debido a que éstos se unen a una región promotora que puede estar a varias kilobases (kb) del gen regulado, por lo que la detección de un TFBS, no implica detectar a qué gen está regulando desde esa posición. En esta tesis, para detectar los TFBS se usó la base de datos de Ensembl, ya que recopila información de distintos consorcios y los procesa uniformemente para detectar el mayor número posible de regiones reguladoras (Zerbino et al., 2015), y posteriormente, como ya se ha hecho en otros estudios (Veerla et al., 2010; Vaes et al., 2006), se determinó que existía una relación entre factor de transcripción y gen si había una distancia inferior a 5 kb entre el TFBS y

el inicio del gen. Hay que tener en cuenta que este método puede estar introduciendo algunos falsos positivos junto con los verdaderos TFBS que regulan a las dianas reales, lo que puede reducir el poder de detección del análisis. También hay que tener en cuenta que a pesar de que existen cerca de 2000 TFs, solo hay relativamente pocos factores de transcripción bien caracterizados (Vaquerizas et al., 2009). Por este motivo se están desarrollando continuamente nuevas herramientas y bases de datos que permiten recoger mejor la relaciones entre factores de transcripción y sus genes diana. Por ejemplo, se han creado algunas bases de datos supervisadas y cuidadosamente anotadas a partir de literatura o experimentos, como TRRUST (Han et al., 2015), con pocas interacciones validadas, pero a medida que se vayan actualizando (Han et al., 2018) constituirán una importante y valiosa fuente de datos. Otra fuente de datos potencialmente importante procede del cruce de todas las fuentes anteriores, como en el trabajo realizado por Garcia-Alonso y colaboradores (2019), donde se extrajeron las interacciones factor de transcripción-diana de fuentes curadas por literatura, experimentos ChIP-seq, predicciones de TFBS y métodos de inferencia por ingeniería reversa, para ordenarlas por nivel de confianza según el número de evidencias que recogían esa interacción.

A pesar de estos problemas, se ha conseguido identificar una cantidad significativa de asociaciones, tanto nuevas como conocidas, entre la actividad de los algunos factores de transcripción y la progresión y supervivencia de algunos tipos de cáncer. Si a todo esto le sumamos el mejor conocimiento que se va adquiriendo, así como la creciente disponibilidad de herramientas que permiten detectar las relaciones entre los factores de transcripción y sus genes regulados, la aplicación de la metodología descrita en este trabajo, junto con esas nuevas fuentes de

datos, pueden suponer una importante herramienta para la identificación de nuevos patrones de activación de factores de transcripción en cáncer.

Otra limitación de esta metodología consiste en la necesidad de tener un grupo control para realizar una comparación sobre la cual emplear la regresión logística, de manera que esta metodología se puede emplear únicamente en experimentos comparativos. Aun así, la aplicación de este método puede ser de utilidad para el estudio de numerosas enfermedades ya que el uso de un grupo control como referencia suele ser una práctica habitual en los diseños experimentales.

El segundo estudio se centró en determinar el estado de activación, a nivel de células individuales, de otro elemento regulador, las rutas de señalización, ya que hasta ahora la única manera que había para estudiar las rutas de señalización en este tipo de experimentos consistía en simples tests de enriquecimiento (Fan et al., 2016). Esta aproximación tiene ciertas limitaciones ya que no tiene en cuenta aspectos de las rutas de señalización tan importantes como son la topología o las relaciones activadoras o inhibidoras entre sus elementos. Por este motivo, en este trabajo se ha buscado aplicar en un experimento de scRNAseq el modelo mecanístico con mejor sensibilidad y especificidad desarrollado hasta la fecha (Amadoz et al., 2019), Hipathia. Este modelo mecanístico emplea los mapas de rutas de señalización de la base de datos KEGG (Kanehisa et al., 2002) para definir los circuitos de señalización que conectan cualquier posible receptor con una proteína efectora. Posteriormente se usan valores de expresión génica para modelar la señal que pasa a través de los nodos que conforman la red, de manera que cada muestra tiene su propio valor del estado de activación de los distintos circuitos. La aplicación de modelos mecanísticos en experimentos de scRNAseq abre un nuevo abanico de posibilidades a la hora de realizar análisis desde puntos de vista interesantes e innovadores, como ya ha ocurrido en

experimentos transcriptómicos clásicos, donde se se ha empleado este tipo de modelos para el estudio de mecanismos celulares de diversas enfermedades como obesidad (Razzoli et al., 2016) o cáncer (Cubuk et al., 2018), para estudiar los mecanismos de acción de fármacos (Esteban-Medina et al., 2019) ,como biomarcadores para la predicción de pronósticos clínicos (Fey et al., 2015; Hidalgo et al., 2017) o la interpretación del efecto que tienen variantes genómicas en enfermedades complejas (Peña-Chilet et al., 2019). Para aplicar esta metodología, primero hubo que solventar una de las principales problemáticas de los experimentos de scRNAseq, y es que los métodos de secuenciación actuales solo capturan entre el 10-40% de los 100.000 a 400.000 mRNAs que tiene una célula (Wu et al., 2014; Islam et al., 2011). Esto es debido en parte a la poca eficiencia y la dificultad de realizar las reacciones necesarias para la secuenciación, como la PCR, en volúmenes del orden de nanolitros (Hashimshony et al., 2012). Como consecuencia, la mayoría de genes con poca expresión suelen ser representados como ceros en la matriz de expresión. A estos ceros técnicos, que no son ceros biológicos reales, se les conoce como dropouts (Lähnemann et al., 2020). Como Hipathia usa la expresión de los genes en las muestras como una representación de la capacidad de las correspondientes proteínas para transmitir la señal a través de las rutas de señalización definidas dentro del pathway, para poder emplear el modelo mecanístico se decidió realizar una imputación de los dropouts empleando herramientas ya desarrolladas, ya que la gran cantidad de ceros técnicos impedía calcular la propagación de la señal a través de dichas rutas de señalización. La siguiente problemática surgió a la hora de elegir el método de imputación, ya que, para abordar el problema de los dropouts, han surgido numerosos métodos distintos que intentan predecir el valor real de los ceros técnicos mediante diferentes abordajes matemáticos, como por ejemplo mediante machine learning o fijándose en la expresión de células cercanas. Para elegir qué

método de imputación se iba a emplear en el experimento de scRNAseq de glioblastoma, se realizó una comparativa de las agrupaciones (*clusterings*) obtenidas mediante 3 métodos de imputación más usados hasta entonces con el *clustering* del artículo original, que se utilizó como referencia. Cuanto más parecido fuera el *clustering* resultante al original menos habría influido la imputación sobre variabilidad biológica subyacente observada en el gráfico inicial. DrImpute fue el método de imputación elegido tras el benchmarking, aunque cabe destacar que la elección del método de imputación a partir de la comparación realizada en este estudio no es extrapolable a otros experimentos de scRNAseq por diversos motivos. En primer lugar, el análisis del transcriptoma a nivel de células es una técnica bastante reciente, por lo que existe un constante desarrollo de nuevas técnicas que permitan explotar este tipo de experimentos (Hodzic, 2016), incluyendo nuevos métodos de imputación. Por ejemplo, en el último año han salido numerosos nuevos métodos que afrontan el problema desde diferentes aproximaciones, como el uso de autoencoders (Eraslan et al., 2019), aproximaciones bayesianas (Tang et al., 2020), una mezcla de estos dos métodos (Wang et al., 2019), usando información de otras células con patrones similares mediante bootstrapping (Tracy et al., 2019) o el uso de deep neural-networks (Arisdakessian et al., 2019). En segundo lugar, con un experimento con más células se podrían haber empleado otros métodos para comparar los diferentes resultados de las imputaciones, como crear dropouts artificialmente sobre valores conocidos y comparar ambos valores. Este método de selección no se pudo llevar a cabo porque el número de células era demasiado ajustado para lo recomendado por los desarrolladores de algunos métodos, de manera que una reducción en la cantidad de información de la matriz de expresión podría hacer que un método no funcionase tan bien y por lo tanto la comparación no fuese tan justa. Por último, el óptimo funcionamiento de cada método de imputación depende

en gran medida de la estructura de los datos, el nivel de ruido, el porcentaje de dropouts y la robustez de diferentes métodos (Zhang y Zhang, 2018). Por ejemplo, los métodos de imputación bayesianos, como SAVER (Huang et al., 2018), deben ser aplicados sobre experimentos con un elevado número de células, o métodos que han sido desarrollados para recuperar información en experimentos de RNAseq como LLSimpute (Kim et al., 2005) pueden ser también empleados para imputar dropouts, pero se recomienda aplicar en datos con baja heterogeneidad y dispersión (Zhang y Zhang, 2018).

Por estos motivos, a pesar de haber determinado que DrImpute era el mejor método de imputación en su momento para nuestros datos, se recomienda hacer una nueva comparativa de métodos de imputación antes de aplicar la metodología desarrollada a un nuevo experimento de scRNAseq.

5.2 HETEROGENEIDAD REGULATORIA EN CÁNCER

A pesar de haberse identificado varias características funcionales en común para la mayoría de tumores (ver sección 1.1.2), existe una gran heterogeneidad de mecanismos causantes de dichos fenotipos moleculares. Esta heterogeneidad se refleja tanto entre los distintos tumores (heterogeneidad intertumoral) como dentro de un mismo tumor (heterogeneidad intratumoral). En esta tesis se ha abordado el estudio de la heterogeneidad desde distintos enfoques, para observar a diferentes niveles como varían los estados de activación de elementos reguladores.

A un nivel más amplio y holístico, se ha presentado el primer catálogo de los estados de activación de 52 factores de transcripción para 11 tipos distintos de tumor calculados mediante un modelo logístico. Para este tipo

de visión es imprescindible proyectos como los deI ICGC o el TCGA, que aportan a la comunidad investigadora una gran cantidad de datos ómicos y clínicos de numerosas muestras de distintos tipos de cáncer que no sería posible conseguir de otra manera. Además, estas muestras se obtienen con unos protocolos estandarizados que reducen el ruido experimental y permiten observar mejor la variabilidad biológica (ICGC y TCGA, 2020). Es evidente que la clasificación por tipos no representa toda la heterogeneidad observable en cáncer, ya que incluso dentro de un tipo se pueden distinguir varios subtipos, por ejemplo, dentro del cáncer de mama se pueden diferenciar los subtipos luminales, *HER2* o triple negativo atendiendo a las distintas características moleculares que presenten los tumores. Sin embargo, la caracterización molecular mediante un análisis de pan-cáncer permite obtener una visión general de los eventos o alteraciones de elementos moleculares en común de distintos tipos de cáncer, así como alteraciones características de uno o un grupo de ellos. De hecho, en esta tesis se han observado ciertos factores de transcripción (*E2F6*, *E2F4*, *MYC*, *MYC:MAX* y *NRF1*) que están significativamente más activados en todos los tipos de cáncer si se comparan con sus respectivas muestras normales. También se pudo comprobar cómo la heterogeneidad observada no dependía excesivamente de la progresión del tumor, ya que al realizar el mismo análisis separando las muestras por los estadíos del tumor la activación de los factores de transcripción parecía consistente a lo largo de los distintos estadíos, salvo en algunas excepciones. En este análisis, a parte de mostrar hechos contrastados como la activación de *MYC* en todos los tumores analizados (Littlewood et al., 2012), se resalta la alteración específica para algunos tipos de cáncer de la actividad de ciertos factores de transcripción que pueden servir como base para la elección de candidatos como marcadores de diagnóstico.

Por otro lado, la metodología desarrollada ha servido para poder describir el escenario regulatorio, de los factores de transcripción analizados, a nivel de paciente. Se puede observar una heterogeneidad a nivel de pacientes tanto en el desarrollo de la enfermedad como en ciertos factores clínicos como la supervivencia. Es común en la práctica clínica, la búsqueda de biomarcadores que permitan reducir esta heterogeneidad y poder distinguir entre los pacientes de alto y bajo riesgo, según su pronóstico. La obtención de valores individuales para cada paciente del estado de activación de los factores de transcripción ha permitido analizar su correlación con la heterogeneidad en la supervivencia de los distintos tipos de cáncer. Se han encontrado patrones de activación significativamente asociados con la supervivencia que pueden servir como base para su utilización como marcadores de pronóstico.

Para este análisis también se ha comprobado la posible implicación de la heterogeneidad de tipos celulares dentro del tumor, conocido como pureza del tumor, en la supervivencia de los pacientes. En un tumor, además de células neoplásicas, se encuentran una variedad de células no cancerosas que principalmente incluyen células inmunes, fibroblastos y células de soporte para los vasos sanguíneos. La composición de tipos celulares de un tumor varía según el tipo de cáncer e incluso por individuos (Aran et al., 2015) y tiene un papel importante en procesos claves del cáncer como la metástasis (Joyce y Pollard, 2009). Por este motivo se comprobó si la pureza de los tumores influía en los resultados del análisis de supervivencia. La pureza del tumor se puede detectar de varias maneras ya sea fijándose en la metilación de islas CpG inmuno-específicas (Johan et al., 2019), en la detección de CNVs somáticos (Carter et al., 2012) o en la expresión de genes inmunes y estromales (Yoshihara et al., 2013), aunque en este trabajo se empleó una medida consenso de las estimaciones obtenidas por distintos métodos (Aran et al., 2015). Tras añadir la pureza

de cada uno de los pacientes como una variable más en el modelo de regresión de Cox, se encontró que solo en 3 de los cánceres estudiados tenía un efecto significativo sobre la supervivencia. A pesar de este hallazgo, no parece que la pureza tenga un efecto en el poder de detección del método, ya que en KIRC y HNSC, aunque son dos de los tumores con menor porcentaje de pureza, es donde se observa un mayor número de factores de transcripción significativos. El principal factor que influye en la detección de factores de transcripción relacionados con la supervivencia es el número de pacientes con datos de fallecimientos. En los últimos años, también se ha descrito la importancia de tener en cuenta la pureza del tumor en análisis genómicos y transcriptómicos, especialmente en genes implicados en la inmunidad celular (Rhee et al., 2018), por lo que habría que tener especial cuidado a la hora de interpretar los resultados tras aplicar la metodología descrita en este trabajo sobre factores de transcripción implicados en la respuesta inmune como AP-1 (Atsaves et al., 2019) en tumores con un bajo porcentaje de pureza. Una posible solución a este problema sería emplear alguno de los métodos de deconvolución desarrollados recientemente (Peng et al., 2019), donde se obtienen para cada gen pesos y relevancias de acuerdo a cada tipo celular detectado en la muestra mediante el método. De esta manera se podría discriminar del grupo de genes diana, que representan la actividad del factor de transcripción, aquellos genes cuya representación se deba principalmente a tipos celulares no neoplásicos. También se podría discriminar las muestras de aquellos pacientes en los que la expresión de la mayoría de genes diana no estén representados por células cancerosas, aunque esto habría que aplicarlo en aquellos tipos de tumor con un considerable número de muestras ya que reduciría su tamaño, y por tanto, el poder de detección.

Por último, a nivel de las células del tumor, se ha conseguido determinar la heterogeneidad en el estado de activación de las rutas de señalización de las células de cuatro tumores de glioblastoma. El uso de experimentos de scRNAseq, con los que se obtienen datos transcriptómicos para células únicas, confieren una ventaja importante a la hora de estudiar la heterogeneidad intratumoral y permite discernir los diferentes eventos moleculares que pueden ocurrir dentro de un tumor. El *clustering* inicial realizado sobre las muestras de glioblastoma refleja claramente la diversidad de tipos celulares que se pueden encontrar además de las células cancerosas. A pesar de esta heterogeneidad celular, las células normales de un mismo tipo se agrupan juntas, independientemente del paciente, mientras que las células neoplásicas forman 3 *clusters* claramente separados. Este comportamiento suele observarse en los distintos experimentos de scRNAseq en muestras de cáncer (Tirosh et al., 2016; Puram et al., 2017; Chung et al., 2017), y muestra cómo la heterogeneidad molecular existente entre las células de un mismo tumor es mayor que la observada en células sanas. Al realizar el *clustering* con los valores de activación de los circuitos calculados se sigue manteniendo el mismo comportamiento, lo que sugiere que el elevado nivel de heterogeneidad observado a nivel de expresión de genes también es observable a nivel de las rutas de señalización. Tras comparar los distintos grupos de células neoplásicas para detectar qué rutas estaban alteradas con respecto a las células normales, se detectaron diferentes circuitos alterados para activar un mismo hallmark, lo que sugiere que existen diferentes mecanismos para el desarrollo del tumor. Además dos de los tres grupos de células neoplásicas tienen activados un mayor número de alteraciones en circuitos relacionados con un desarrollo más agresivo, como por ejemplo rutas relacionadas con la metástasis. Complementar este estudio con un experimento que contuviera muestras de recidiva sería altamente interesante para observar si los diferentes perfiles encontrados están detrás

de mecanismos de resistencia a un tratamiento. Complementariamente, se subtipó a las células neoplásicas de acuerdo a los subtipos consenso en glioblastoma descritos por Wang y colaboradores (2017) para ver si coincidían con los patrones observados. Como ya se ha descrito en otros tipos de cáncer (Chung et al., 2017; Valdes-Mora et al.,2018), un tumor está formado por una mezcla de subtipos. Los dos grupos de células con un perfil más agresivo son las que contienen un mayor número de células con subtipo mesenquimal, ligado a una peor prognosis (Huse et al., 2011). Aunque es cierto que en cada tumor se observa la predominancia de un subtipo determinado, es probable que la presencia de poblaciones celulares residuales de un subtipo distinto al mayoritario pueda explicar la recaída y resistencia de algunos tumores a ciertas terapias dirigidas a combatir un subtipo específico (Boumahdi y Sauvage, 2020). Esta idea también se ha visto reforzada tras medir la heterogeneidad en la respuesta a tratamientos, ya que gracias al uso de un modelo mecanístico, fue posible realizar intervenciones in silico simulando el efecto de diversas drogas sobre sus dianas terapéuticas. Por ejemplo, en la simulación del efecto de bevacizumab sobre las células neoplásicas de glioblastoma se identificó un grupo de células con una respuesta más débil al bloqueo del receptor *VEGFA* por parte del fármaco que el resto de células. Tras observar en detalle los circuitos en los que intervenía el fármaco, se pudo observar que estas células con menor respuesta, pertenecientes en su mayoría al subtipo proneural, activaban dichos circuitos mediante la expresión de otros receptores como *PDGFD*, *KITLG* y *FGF2*.

Para tener una imagen completa de la heterogeneidad intratumoral sobre los elementos reguladores más importantes, la metodología desarrollada en este trabajo podría ser complementada con el uso de otras herramientas desarrolladas específicamente para determinar el estado de activación de

los factores de transcripción en experimentos de scRNAseq (Aibar et al.,2017; Ding et al., 2018).

En definitiva, la metodología desarrollada a lo largo de esta tesis ha permitido retratar la heterogeneidad en cáncer sobre distintos factores moleculares y clínicos y a distintos niveles. Si bien el estudio de la heterogeneidad intertumoral ha permitido realizar grandes avances en el estudio del cáncer, cada vez parece más evidente la necesidad de incorporar el análisis de la heterogeneidad intratumoral a la hora de desentrañar la gran complejidad que caracteriza al cáncer y poder dilucidar mecanismos incomprendidos hasta ahora, como la resistencia a ciertos tratamientos o las diferencias observadas en el desarrollo del tumor de distintos pacientes para un mismo tipo de cáncer (Fisher et al., 2013). Si el análisis de la heterogeneidad regulatoria en cuatro pacientes de glioblastoma ha permitido comprender mejor algunos mecanismos regulatorios de esta enfermedad, un esfuerzo similar al del proyecto TCGA pero recogiendo análisis de scRNAseq bajo unos estándares de calidad comunes, sería una fuente de datos inmensurable para arrojar luz sobre la complejidad regulatoria del cáncer.

5.3 HETEROGENEIDAD Y MEDICINA PERSONALIZADA

A medida que en la lucha contra el cáncer se ha ido describiendo la gran variedad de mecanismos moleculares que hay detrás de la formación y el desarrollo del tumor, se ha hecho más evidente la necesidad de establecer decisiones terapéuticas de forma personalizada, es decir, en función de las características genómicas y moleculares del tumor de cada paciente. El estudio de la heterogeneidad en cáncer adquiere un papel muy importante a la hora de aplicar la medicina personalizada en pacientes, ya que, por un

lado, sirve para caracterizar mejor las diferencias entre tumores y adaptar el tratamiento en consecuencia, pero por otro lado, también está detrás de algunos mecanismos de resistencia (Dagogo-Jack y Shaw, 2018).

El análisis de la heterogeneidad intertumoral, ha permitido mejorar la estratificación de los pacientes en distintos subtipos en función de sus características moleculares. La clasificación por subtipos moleculares ha supuesto una gran mejora a la hora de tratar y diagnosticar numerosos tipos de cáncer. Por ejemplo, en cáncer de pulmón de células no pequeñas, la identificación de mutaciones en el gen *EGFR* (Paez et al., 2004) y la translocación *EML4-ALK* (Soda et al., 2007) han supuesto un progreso significativo en el estudio de los estadios avanzados de la enfermedad. También en cáncer de mama, es una práctica común por parte de los oncólogos observar si el tumor de un paciente sobreexpresa el gen *HER2*, asociado a un peor pronóstico, pero con una mejor respuesta a la medicación trastuzumab (Gutierrez y Schiff, 2011).

Esta clasificación por subtipos se puede basar en varios factores genéticos, como la expresión de uno o varios genes, o de una determinada mutación o alteración estructural. Para un mismo tipo de tumor pueden existir varias clasificaciones que pueden ir variando según el tipo de heterogeneidad que se quiera explicar. Por ejemplo, para las muestras de glioblastoma empleadas en nuestro estudio, existe una clasificación original de 4 subtipos (Verhaak et al., 2010), basada en la expresión de 200 genes para la clasificación de cada subtipo: classical, neural, proneural y mesenquimal. Algunas de las clasificaciones pueden estar sesgadas, al incluir en el análisis parte del transcriptoma de células sanas asociadas a tumor (Sidaway, 2017), por lo que se propuso la reclasificación del glioblastoma en 3 subtipos (Wang et al., 2017), eliminando el subtipo neural. Sin embargo, se han propuesto varias clasificaciones más en este cáncer basándose en alteraciones genéticas (Zhou el at., 2018), diferencias

en el microambiente del tumor (Chen y Hambardzumyan, 2018) o creando subtipos que explicaran mejor ciertas variables clínicas como la progresión (Fatai y Gamieldien, 2018) o el pronóstico (Park et al., 2019). Este hecho refleja como uno de los mayores retos de la medicina personalizada consiste en la correcta selección de marcadores para una mejor caracterización de los pacientes.

Las metodologías desarrolladas a lo largo de esta tesis, también permiten detectar perturbaciones en los patrones de activación de elementos regulatorios a nivel de paciente y proporcionan una valiosa fuente de información para su aplicación en la medicina personalizada, ya que permite la detección de diversos marcadores diagnósticos, pronósticos o de resistencia. De hecho, algunos de los resultados obtenidos en el análisis de pan-cáncer ya han sido propuestos como marcadores diagnósticos con anterioridad, como el caso de los receptores X retinoides (*RXR*) en cáncer de tiroides (Hoftijzer et al., 2009) o también factores de transcripción de la familia *PAX* para cáncer de riñón (Barr et al., 2015). Además, mediante el análisis de pan-cáncer también ha sido posible la identificación de alteraciones comunes entre diferentes linajes de tumores, lo cual puede ser de utilidad para extender una terapia determinada a varios tipos de tumor con perfiles fenotípicos similares. Por ejemplo, recientemente se ha observado mediante este tipo de análisis como varios tipos de tumores presentan mutaciones $BRAF^{V600}$, y se ha estudiado cómo la inhibición de *RAF* puede ser beneficiosa incluso para aquellos tumores que presentan la mutación, pero para los que todavía no es considerado como un tratamiento estándar (Subbiah et al., 2020).

Para detectar posibles marcadores de supervivencia, con los patrones de activación de factores de transcripción por individuo se calcularon curvas Kaplan-Meier y un modelo de Cox para cada tipo de tumor. Con las curvas de Kaplan-Meier se consiguieron detectar 19 factores de transcripción en

3 tipos de cáncer cuya actividad estaba relacionada con la supervivencia del paciente. El modelo de regresión de Cox, acoplado a un algoritmo stepwise para la selección de variables, permitió detectar que combinaciones de activaciones en los factores de transcripción explicaban la supervivencia observada, además de poder observar la relevancia que tenían en el pronóstico otras variables como la pureza del tumor. Sin embargo, puede que los resultados del modelo de regresión de Cox no sean directamente aplicables en práctica clínica como marcadores de supervivencia, ya que a pesar de que en conjunto consiguen explicar mejor la supervivencia, puede que la aplicación como marcador de cada uno de ellos por separado no tenga suficiente poder para predecir por sí solos la supervivencia de los pacientes, como se puede observar en el hecho que algunos de los factores de transcripción significativos en el modelo de Cox no lo son en el de Kaplan-Meier.

Una de las deficiencias que se suelen encontrar a la hora de detectar los distintos marcadores es la falta de datos clínicos de los pacientes (Liu et al., 2018). En el análisis de supervivencia realizado con los estados de activación de los factores de transcripción hemos conseguido detectar más marcadores de pronóstico en aquellos tumores con un mayor número de fallecimientos reportados. Esto evidencia cómo la integración de datos genómicos y otros datos clínicos en un registro de salud electrónico sería ampliamente beneficioso para el avance de la medicina personalizada (Abul-Husn y Kenny, 2019). Esta integración además permitiría, no solo tener datos de muchos pacientes y muchas enfermedades, si no la capacidad de hacer diversos estudios, como análisis de comorbilidades, estudios transversales y prospectivos, entre otros, donde se puede estudiar la evolución de la heterogeneidad del tumor tras los distintos tratamientos o recidivas.

También existen varias evidencias que muestran los grandes beneficios de implantar la medicina personalizada en el sistema público de salud, a pesar de los desafíos que conlleva (Mathur y Sutton, 2017; Goodsaid et al., 2020). Actualmente existen varias iniciativas de distintos países que pretenden implementar la secuenciación genómica en el sistema de salud público y la práctica clínica. Para conseguirlo se han invertido cerca de 4.000 millones de dólares en al menos 14 países (Stark et al., 2019). Concretamente en España, aplicar la medicina personalizada en todo el ámbito nacional supone un reto complicado debido a la descentralización del sistema de salud, y a la falta en algunos casos de disponibilidad electrónica del historial de salud clínica. A pesar de ello, Andalucía ha sido la primera región en apostar por poner en marcha un programa de medicina personalizada en todo su territorio e integrada en su sistema de salud. Aunque también existen otros proyectos en España que son interesantes, como, entre otros, el proyecto NAGEN en Navarra en el cual pretenden analizar 1000 genomas de pacientes y sus familiares con enfermedades raras y cáncer familiar.

Otro tipo de marcadores útiles para la práctica de la medicina personalizada, son aquellos capaces de discernir qué pacientes pueden beneficiarse de un determinado tratamiento. Por ejemplo, en pacientes con leucemia que tienen una translocación cromosómica t(9;22)(q34;q11) se observa el doble de tasa de supervivencia tras el tratamiento con imatinib que en pacientes que no tienen dicha alteración (Druker et al., 2001). También se ha observado un aumento en la supervivencia de aquellos pacientes de cáncer colorrectal tratados con cetuximab portadores de una mutación en el gen *EGFR,* pero no en *KRAS* (Karapetis et al., 2008). Aunque hasta ahora la mayoría de marcadores de respuesta se fijaban en la heterogeneidad intertumoral, se ha observado cómo los tumores con un mayor nivel de heterogeneidad intratumoral pueden predisponer a los

pacientes a peores resultados clínicos debido a la expansión de subpoblaciones clonales existentes o la evolución de células tolerantes a un tratamiento dirigido como resultado de la presión selectiva terapéutica (Menon et al., 2015; Sharma et al., 2010; Roesch et al., 2013). Por este motivo, el análisis de la heterogeneidad intratumoral está ganando peso en el estudio del tratamiento de diversos tipos de cáncer (Sexton et al., 2020; Haynes et al., 2017; Hernandez et al., 2019; Su et al., 2017). Sin embargo, en ocasiones falta profundidad en la interpretación de los resultados obtenidos del análisis de la heterogeneidad para entender mejor los mecanismos detrás de procesos como el desarrollo del tumor o la resistencia a tratamientos (Cirillo y Valencia, 2019; Fryburg et al., 2014). En este sentido, la aplicación de modelos mecanísticos a experimentos de scRNAseq no solo permite interpretar mejor la heterogeneidad intratumoral presente en un tumor, si no que además puede servir para detectar posibles mecanismos de resistencia y sugerir marcadores de resistencia para su aplicación en medicina personalizada. Tras realizar una simulación en datos de scRNAseq de glioblastoma del efecto de bevacizumab, se observó que en las células menos afectadas por la droga un grupo de genes, en especial *PDGFD*, tenían una mayor expresión que la diana terapéutica (*VEGFA*) con la que comparten función. Este comportamiento podría sugerir que en este pequeño grupo de células estos genes están relevando a *VEGFA* en su función por lo que podrían convertirse en una semilla para una futura emergencia de un tumor resistente. Aunque estos resultados requieren una validación in vivo (Rambow et al., 2018), esta herramienta es una gran aliada a la hora de detectar poblaciones potencialmente resistentes a ciertos tratamientos dirigidos.

En definitiva, las terapias dirigidas han conseguido mejorar el tratamiento de muchos tipos de cáncer. Los distintos niveles de heterogeneidad juegan

un papel muy importante para la implementación de la medicina personalizada, por lo que su estudio y comprensión permitirá grandes avances en el tratamiento de la enfermedad. Con ese propósito, en esta tesis se han presentado varias metodologías y análisis que pueden servir como fuente de recursos para la detección de posibles marcadores diagnósticos, pronósticos y de respuesta que reflejen la heterogeneidad existente en estudios de cáncer.

Capítulo 6

Conclusiones generales

6. CONCLUSIONES GENERALES

- La aplicación de modelos mecanísticos en experimentos de scRNAseq ha servido para caracterizar la heterogeneidad intratumoral en la activación de las rutas de señalización.

- El uso de un modelo logístico en experimentos comparativos ha permitido describir el escenario regulatorio de los factores de transcripción.

- Mediante el estudio de la activación de los factores de transcripción en un análisis de pan-cáncer se ha presentado el primer catálogo de la activación de 52 factores de transcripción en 11 tipos de cáncer distintos, donde se encuentran algunas alteraciones descritas en la literatura y otras aún sin explorar. Además la adaptación de la metodología para obtener valores de activación individuales por pacientes ha permitido observar la relación entre la activación de factores de transcripción y la supervivencia.

- El uso de modelos mecanísticos en un experimento de scRNAseq de 4 pacientes de glioblastoma ha facilitado la interpretación de la heterogeneidad regulatoria intratumoral observada en este tipo de cáncer. También ha permitido realizar simulaciones de distintas intervenciones terapéuticas dirigidas donde se han observado distintos patrones de respuesta a fármacos en algunas células, sugiriendo posibles mecanismos de resistencia.

- A pesar de las limitaciones en las metodologías desarrolladas, como son la imputación, la detección de los genes diana para los factores de transcripción o la falta de datos clínicos, se han detectado múltiples asociaciones significativas entre la actividad de mecanismos regulatorios y algunas características del cáncer. Aun así, se espera que una mejora en estas limitaciones aumente el potencial de detección de las herramientas.

- Debido a la detección de potenciales marcadores de diagnóstico, pronóstico y respuesta, tanto las metodologías como los estudios realizados en esta tesis constituyen un valioso recurso para la medicina personalizada.

REFERENCIAS

Abul-Husn, N. S., & Kenny, E. E. (2019). Personalized medicine and the power of electronic health records. Cell, 177(1), 58-69.

Adams, J. M., & Cory, S. (2007). The Bcl-2 apoptotic switch in cancer development and therapy. Oncogene, 26(9), 1324-1337.

Adelman, K., & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nature Reviews Genetics, 13(10), 720-731.

Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J., Geurts, P., Aerts, J., Van den Oord, J., Atak, Z., Wouters, J., & Aerts, S. K. (2017). SCENIC: single-cell regulatory network inference and clustering. Nature methods, 14(11), 1083-1086.

Al-Shahrour, F., Minguez, P., Tárraga, J., Medina, I., Alloza, E., Montaner, D., & Dopazo, J. (2007). FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic acids research, 35(suppl_2), W91-W96.

Amadoz, A., Hidalgo, M. R., Cubuk, C., Carbonell-Caballero, J., & Dopazo, J. (2019). A comparison of mechanistic signaling pathway activity analysis methods. Briefings in bioinformatics, 20(5), 1655-1668.

André, F., Ciruelos, E., Rubovszky, G., Campone, M., Loibl, S., Rugo, H. S.,Iwata, H., Conte, P., Mayer, I. A., Kaufman, B., Yamashita, T., & Lu, Y. (2019). Alpelisib for PIK3CA-mutated, hormone receptor–positive advanced breast cancer. New England Journal of Medicine, 380(20), 1929-1940.

Anastasiadou, E., Jacob, L. S., & Slack, F. J. (2018). Non-coding RNA networks in cancer. Nature Reviews Cancer, 18(1), 5.

Aran, D., Sirota, M., & Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. Nature communications, 6(1), 1-12.

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., & Garmire, L. X. (2019). DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome biology, 20(1), 1-14.

Baeriswyl, V., & Christofori, G. (2009, October). The angiogenic switch in carcinogenesis. In Seminars in cancer biology (Vol. 19, No. 5, pp. 329-337). Academic Press.

Barr, M. L., Jilaveanu, L. B., Camp, R. L., Adeniran, A. J., Kluger, H. M., & Shuch, B. (2015). PAX-8 expression in renal tumours and distant sites: a useful marker of primary and metastatic renal cell carcinoma?. Journal of clinical pathology, 68(1), 12-17.

Bhagwat, A. S., & Vakoc, C. R. (2015). Targeting transcription factors in cancer. Trends in cancer, 1(1), 53-65.

Blasco, M. A. (2005). Telomeres and human disease: ageing, cancer and beyond. Nature Reviews Genetics, 6(8), 611-622.

Boumahdi, S., & de Sauvage, F. J. (2020). The great escape: tumour cell plasticity in resistance to targeted therapy. Nature Reviews Drug Discovery, 19(1), 39-56.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 68(6):394–424. https://doi.org/10.3322/caac.21492 PMID:30207593

Brown, K. F., Rumgay, H., Dunlop, C., Ryan, M., Quartly, F., Cox, A., Deas, A., Ellis-Brookes, L., Gavin, A., Hounsome, L., Huws, D., Orminton-Smith, N., Shelton, J., White, C., & Parkin, D. M. (2018). The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. British journal of cancer, 118(8), 1130-1141.

Bugarín-González, R., & Carracedo, A. (2018). Genética y medicina de familia. Medicina de Familia. SEMERGEN, 44(1), 54-60.

Cantley, L. (2012). Principles of Cell Signaling. Seldin and Giebisch's The Kidney: Physiology and Pathophysiology, 369.

Cavallaro, U., & Christofori, G. (2004). Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. Nature Reviews Cancer, 4(2), 118-132.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M. & Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology, 30(5), 413-421.

Chen, Y. F., Jobanputra, P., Barton, P., Jowett, S., Bryan, S., Clark, W., & Burls, A. (2006). A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness. In NIHR Health Technology Assessment programme: Executive Summaries. NIHR Journals Library.

Chen, Z., & Hambardzumyan, D. (2018). Immune microenvironment in glioblastoma subtypes. Frontiers in immunology, 9, 1004.

Cheng, N., Chytil, A., Shyr, Y., Joly, A., & Moses, H. L. (2008). Transforming growth factor-β signaling–deficient fibroblasts enhance hepatocyte growth factor signaling in mammary carcinoma cells to promote scattering and invasion. Molecular Cancer Research, 6(10), 1521-1533.

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

Chung, W., Eum, H. H., Lee, H. O., Lee, K. M., Lee, H. B., Kim, K. T., Ryu, H. S., Kim, S., Lee, J. E., Park, Y. H., Kan, Z, Han, W. & Park, W. Y. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nature communications, 8(1), 1-12.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, G., Lander, E. S., & Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology, 31(3), 213-219.

Cirillo, D., & Valencia, A. (2019). Big data analytics for personalized medicine. Current opinion in biotechnology, 58, 161-167.

Cookson, J., Gilaberte, I., Desaiah, D., & Kajdasz, D. K. (2006). Treatment benefits of duloxetine in major depressive disorder as assessed by number needed to treat. International clinical psychopharmacology, 21(5), 267-273.

Corey, L., Valente, A., & Wade, K. (2020). Personalized medicine in gynecologic cancer: fact or fiction?. Surgical Oncology Clinics, 29(1), 105-113.

Currie, G. P., Lee, D. K., & Lipworth, B. J. (2006). Long-Acting β 2-Agonists in Asthma. Drug safety, 29(8), 647-656.

Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. Nature reviews Clinical oncology, 15(2), 81.

Davies, M. A., & Samuels, Y. (2010). Analysis of the genome to personalize therapy for melanoma. Oncogene, 29(41), 5545-5555.

de Anda-Jáuregui, G., & Hernández-Lemus, E. (2020). Computational Oncology in the Multi-Omics Era: State of the Art. Frontiers in Oncology, 10, 423.

DeBerardinis, R. J., Lum, J. J., Hatzivassiliou, G., & Thompson, C. B. (2008). The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. Cell metabolism, 7(1), 11-20.

de Bruin, E. C., McGranahan, N., & Swanton, C. (2015). Analysis of intratumor heterogeneity unravels lung cancer evolution. Molecular & cellular oncology, 2(3), e985549.

Dekker, A., Bulley, S., Beyene, J., Dupuis, L. L., Doyle, J. J., & Sung, L. (2006). Meta-analysis of randomized controlled trials of prophylactic granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor after autologous and allogeneic stem cell transplantation. Journal of Clinical Oncology, 24(33), 5207-5215.

Demichelis, F., Setlur, S. R., Banerjee, S., Chakravarty, D., Chen, J. Y. H., Chen, C. X., Huang, J., Beltran, H., Oldridge, D. A., Kitabayashi, N., Stenzel, B., Schaefer, G., Horninger, W., Bektic, J., Chinnaiyan, A. M., Goldenberg, S., Siddiqui, J., Regan, M. M., Kearney, M., Sanda, M. A., Bartsch, G., Lee, C., Klocker, H., & Rubin, M.A. (2012). Identification of functionally active, low frequency copy number variants at 15q21. 3 and 12q21. 31 associated with prostate cancer risk. Proceedings of the National Academy of Sciences, 109(17), 6686-6691.

Ding, H., Douglass, E. F., Sonabend, A. M., Mela, A., Bose, S., Gonzalez, C., Canoll, P. D., Sims, P. A., Alvarez, M. J. & Califano, A. (2018). Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. Nature communications, 9(1), 1-10.

Dreesen, O., & Brivanlou, A. H. (2007). Signaling pathways in cancer and embryonic stem cells. Stem cell reviews, 3(1), 7-17.

Dubitzky, W., Granzow, M., Downes, C. S., & Berrar, D. (2003). Introduction to microarray data analysis. In A practical approach to microarray data analysis (pp. 1-46). Springer, Boston, MA.

Druker, B. J., Sawyers, C. L., Kantarjian, H., Resta, D. J., Reese, S. F., Ford, J. M., Capdeville, R., & Talpaz, M. (2001). Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. New England Journal of Medicine, 344(14), 1038-1042.

Dzau, V. J., & Ginsburg, G. S. (2016). Realizing the full potential of precision medicine in health and health care. Jama, 316(16), 1659-1660.

El-Sayeh, H. G., & Morganti, C. (2006). Aripiprazole for schizophrenia. Cochrane Database of Systematic Reviews, (2).

ENCODE Project Consortium. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biol, 9(4), e1001046.

England, G. (2016). The 100,000 genomes project. The, 100, 0-2.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nature communications, 10(1), 1-14.

Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., Kaper, F., Fan, J., Zhang, K., Chun, J., & Kharchenko, P. V. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nature methods, 13(3), 241-244.

Fane, M., & Weeraratna, A. T. (2019). How the ageing microenvironment influences tumour progression. Nature Reviews Cancer, 1-18.

Fatai, A. A., & Gamieldien, J. (2018). A 35-gene signature discriminates between rapidly-and slowly-progressing glioblastoma multiforme and predicts survival in known subtypes of the cancer. BMC cancer, 18(1), 377.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram I., & Bray, F. (2018). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Disponible en: https://gco.iarc.fr/today. Consulta: [10 Junio 2020].

Fey, D., Halasz, M., Dreidax, D., Kennedy, S. P., Hastings, J. F., Rauch, N., Muñoz, A. G., Pilkington, R., Fischer, M., Westermann, F., Kolch, W., Kholodenko, B. N., & Croucher, D. R. (2015). Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. Sci. Signal., 8(408), ra130-ra130.

Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. British journal of cancer, 108(3), 479-485.

Flavahan, W. A., Gaskell, E., & Bernstein, B. E. (2017). Epigenetic plasticity and the hallmarks of cancer. Science, 357(6348).

Freedman, M. S., Hughes, B., Mikol, D. D., Bennett, R., Cuffel, B., Divan, V., & Ahmad, A. S. (2008). Efficacy of disease-modifying therapies in relapsing remitting multiple sclerosis: a systematic comparison. European neurology, 60(1), 1-11.

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., Maathuis, M.H., Moreau, Y., Murphy, S. A., Przytycka, T. M., Rebhan, M., Röst, H., Schuppert, A., Schwab, M., Spang, R., Steckhoven, D., Sun, J., Weber, A., Ziemek, D., & Zupan, B. (2018). From hype to reality: data science enabling personalized medicine. BMC medicine, 16(1), 150.

Fryburg, D. A., Song, D. H., Laifenfeld, D., & de Graaf, D. (2014). Systems diagnostics: anticipating the next generation of diagnostic tests based on mechanistic insight into disease. Drug Discovery Today, 19(2), 108-112.

Fuda, N. J., Ardehali, M. B., & Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature, 461(7261), 186-192.

Galanina, N., Bossuyt, V., & Harris, L. N. (2011). Molecular predictors of response to therapy for breast cancer. The Cancer Journal, 17(2), 96-103.

Gallyas Jr, F., Sumegi, B., & Szabo, C. (2020). Role of Akt activation in PARP inhibitor resistance in cancer. Cancers, 12(3), 532.

Gambardella, V., Tarazona, N., Cejalvo, J. M., Lombardi, P., Huerta, M., Roselló, S., Fleitas, T., Roda, D., & Cervantes, A. (2020). Personalized Medicine: Recent Progress in Cancer Therapy. Cancers, 12(4), 1009.

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome research, 29(8), 1363-1375.

Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., & Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. BMC bioinformatics, 19(1), 1-10.

Goodsaid, F., Frueh, F., & Burczynski, M. E. (2020). Personalized Medicine. Drug Discovery and Evaluation: Methods in Clinical Pharmacology, 425-438.

Gralnek, I. M., Dulai, G. S., Fennerty, M. B., & Spiegel, B. M. (2006). Esomeprazole versus other proton pump inhibitors in erosive esophagitis: a meta-analysis of randomized clinical trials. Clinical Gastroenterology and Hepatology, 4(12), 1452-1458.

Grivennikov, S. I., Greten, F. R., & Karin, M. (2010). Immunity, inflammation, and cancer. Cell, 140(6), 883-899.

Gutierrez, C., & Schiff, R. (2011). HER2: biology, detection, and clinical implications. Archives of pathology & laboratory medicine, 135(1), 55-62.

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T.,Kim, H., Kim, K., Yang, S., Bae, D., Yun, A., Kim, S., Kim, C. Y., Cho, H. J., Kang, B., Shin, S., & Lee, I. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. Scientific reports, 5, 11432.

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H., Jung, H., Nam, S., Chung,

M., Kim, J., & Lee, I.. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic acids research, 46(D1), D380-D386.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. cell, 100(1), 57-70.

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

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell reports, 2(3), 666-673.

Haynes, B., Sarma, A., Nangia-Makker, P., & Shekhar, M. P. (2017). Breast cancer complexity: implications of intratumoral heterogeneity in clinical management. Cancer and Metastasis Reviews, 36(3), 547-555.

Hernandez, A. L., Wang, Y., Somerset, H. L., Keysar, S. B., Aisner, D. L., Marshall, C., Bowles, D. W., Karam, S. D., Raben, D., Jimeno, A., Varella-Garcia, M., & Wang, X. (2019). Inter-and intra-tumor heterogeneity of SMAD4 loss in head and neck squamous cell carcinomas. Molecular carcinogenesis, 58(5), 666-673.

Hidalgo, M. R., Cubuk, C., Amadoz, A., Salavert, F., Carbonell-Caballero, J., & Dopazo, J. (2017). High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget, 8(3), 5160.

Hodzic, E. (2016). Single-cell analysis: Advances and future perspectives. Bosnian journal of basic medical sciences, 16(4), 313.

Hoftijzer, H. C., Liu, Y. Y., Morreau, H., van Wezel, T., Pereira, A. M., Corssmit, E. P., Romijn, J. A., & Smit, J. W. (2009). Retinoic acid receptor and retinoid X receptor subtype expression for the differential diagnosis of thyroid neoplasms. European journal of endocrinology, 160(4), 631.

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, Schadendorf, D., & Kumar, R. (2013). TERT promoter mutations in familial and sporadic melanoma. Science, 339(6122), 959-961.

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., & Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science, 339(6122), 957-959.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., & Zhang, N. R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. Nature methods, 15(7), 539-542.

Hyman, D. M., Taylor, B. S., & Baselga, J. (2017). Implementing genome-driven oncology. Cell, 168(4), 584-599.

Hynes, N. E., & MacDonald, G. (2009). ErbB receptors and signaling pathways in cancer. Current opinion in cell biology, 21(2), 177-184.

Jiang, Y., Shen, H., Xiao'an Liu, J. D., Jin, G., Qin, Z., Chen, J., Wang, S., Wang, X., Hu, Z., & Shen, H. (2011). Genetic variants at 1p11. 2 and breast cancer risk: a two-stage study in Chinese women. PloS one, 6(6).

I. C. G. C., of Whole Genomes Consortium, T. P. C. A. (2020). Pan-cancer analysis of whole genomes. Nature, 578(7793), 82.

International Cancer Genome Consortium (ICGC). About us [Internet] Disponible en: https://icgc.org/about-us. Consulta: [24-07-2020]

Ignatiadis, M., & Sotiriou, C. (2013). Luminal breast cancer: from biology to treatment. Nature reviews Clinical oncology, 10(9), 494.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J. B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome research, 21(7), 1160-1167.

Jackson, S. P., & Bartek, J. (2009). The DNA-damage response in human biology and disease. Nature, 461(7267), 1071-1078.

Jacob, L., Neuvial, P., & Dudoit, S. (2012). More power via graph-structured tests for differential expression of gene networks. The Annals of Applied Statistics, 6(2), 561-600.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science, 316(5830), 1497-1502.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., & Stein, L. (2005). Reactome: a knowledgebase of biological pathways. Nucleic acids research, 33(suppl_1), D428-D432.

Joyce, J. A., & Pollard, J. W. (2009). Microenvironmental regulation of metastasis. Nature reviews cancer, 9(4), 239-252.

Kanehisa, M., Goto, S., Kawashima, S., &, Nakaya, A. (2002). The KEGG databases at GenomeNet. Nucleic acids research, 30(1), 42-46.

Karapetis, C. S., Khambata-Ford, S., Jonker, D. J., O'Callaghan, C. J., Tu, D., Tebbutt, N. C., Simes, J. R., Chalchal, H., Shapiro, J. D., Robitaille, S., Price, T. J. &,

Shepherd, L. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. New England Journal of Medicine, 359(17), 1757-1765.

Kim, H., Golub, G. H., & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics, 21(2), 187-198.

Kim, R. (2007). Cancer immunoediting: from immune surveillance to immune escape. In Cancer Immunotherapy (pp. 9-27). Academic Press.

Kitano, H. (2002). Systems biology: a brief overview. science, 295(5560), 1662-1664.

Kristensen, L. E., Christensen, R., Bliddal, H., Geborek, P., Danneskiold-Samsøe, B., & Saxne, T. (2007). The number needed to treat for adalimumab, etanercept, and infliximab based on ACR50 response in three randomized controlled trials on established rheumatoid arthritis: a systematic literature review. Scandinavian journal of rheumatology, 36(6), 411-417.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamakis, A., Attolini, C. S., Aparicio, S., Baaijens, J., Balvert, M., De Barbanson, B., Capuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T., Lelieveldt, B. P. F., Mandoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., De Ridder, J., Saliba, A., Somarakis, A., Stegle, O., Thes, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., & Schönhuth, A.. (2020). Eleven grand challenges in single-cell data science. Genome biology, 21(1), 1-35.

Lafzi, A., Moutinho, C., Picelli, S., & Heyn, H. (2018). Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. Nature protocols, 13(12), 2742-2757.

Lee, E. Y., & Muller, W. J. (2010). Oncogenes and tumor suppressor genes. Cold Spring Harbor perspectives in biology, 2(10), a003236.

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. Cell, 152(6), 1237-1251.

Lemon, B., & Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. Genes & development, 14(20), 2551-2569.

Levy-Lahad, E. & Friedman, E. (2007). Cancer risks among BRCA1 and BRCA2 mutation carriers. British journal of cancer, 96(1), 11-15.

Li, X., Shen, L., Shang, X., & Liu, W. (2015). Subpathway analysis based on signaling-pathway impact analysis of signaling pathway. PloS one, 10(7), e0132813.

Littlewood, T. D., Kreuzaler, P. & Evan, G. I. (2012). All things to all people. Cell, 151(1), 11-13.

Liu, G., Gramling, S., Munoz, D., Cheng, D., Azad, A. K., Mirshams, M., Chen, Z., Xu, W., Roberts, H., Shepherd, F. A., Tao, M. S. & Reisman, D. (2011). Two novel BRM insertion promoter sequence variants are associated with loss of BRM expression and lung cancer risk. Oncogene, 30(29), 3295-3304.

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., & Thorsson, V. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell, 173(2), 400-416.

Losic, B., Craig, A. J., Villacorta-Martin, C., Martins-Filho, S. N., Akers, N., Chen, X., Ahsen, M. E., Von Felden, J., Labgaa, I., D'Avola, D., Allette, K., Lira, S. A., Furtado, G. C., Garcia-Lezana, T., Restrepo, P., Stueck, A., Ward, S. C., Fiel, M. I., Hiotis, S. P., Gunasekaran, G., Sia, D., Schadt, E. E., Sebra, R., Schwartz, M., Llovet, J. M., Thung, S., Stolovitzky, G. &, Villanueva, A. (2020). Intratumoral heterogeneity and clonal evolution in liver cancer. Nature communications, 11(1), 1-15.

Lubbe, S. J., Pittman, A. M., Olver, B., Lloyd, A., Vijayakrishnan, J., Naranjo, S., Dobbins, S., Broderick, P., Gómez-Skarmeta, J. L. & Houlston, R. S. (2012). The 14q22. 2 colorectal cancer variant rs4444235 shows cis-acting regulation of BMP4. Oncogene, 31(33), 3777-3784.

Luskin, M. R., Murakami, M. A., Manalis, S. R., & Weinstock, D. M. (2018). Targeting minimal residual disease: a path to cure?. Nature Reviews Cancer, 18(4), 255.

Maleki, F., Ovens, K., Hogan, D. J., & Kusalik, A. J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. Frontiers in Genetics, 11, 654.

Martelotto, L. G., Ng, C. K., Piscuoglio, S., Weigelt, B., & Reis-Filho, J. S. (2014). Breast cancer intra-tumor heterogeneity. Breast Cancer Research, 16(3), 210.

Marusyk, A., Janiszewska, M., & Polyak, K. (2020). Intratumor heterogeneity: The rosetta stone of therapy resistance. Cancer cell, 37(4), 471-484.

Marusyk, A., & Polyak, K. (2010). Tumor heterogeneity: causes and consequences. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1805(1), 105-117.

Mathur, S., & Sutton, J. (2017). Personalized medicine could transform healthcare. Biomedical reports, 7(1), 3-5.

Matrone, A., Campopiano, M. C., Nervo, A., Sapuppo, G., Tavarelli, M., & De Leo, S. (2020). Differentiated Thyroid Cancer, From Active Surveillance to Advanced Therapy: Toward a Personalized Medicine. Frontiers in endocrinology, 10, 884.

Menon, D. R., Das, S., Krepler, C., Vultur, A., Rinner, B., Schauer, S., Kashofer, K., Wagner, K., Zhang, G., Rad, E. B., Haass, N. K., Soyer, H. P., Gabrielli, B., Somasundarram, R., Hoefler, G., Herlyn, M., & Schaider, H. (2015). A stress-induced early innate response causes multidrug tolerance in melanoma. Oncogene, 34(34), 4448-4459.

Micalizzi, D. S., Farabaugh, S. M., & Ford, H. L. (2010). Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression. Journal of mammary gland biology and neoplasia, 15(2), 117-134.

Miething, C., Scuoppo, C., Bosbach, B., Appelmann, I., Nakitandwe, J., Ma, J., Wu, G., Lintault, L., Auer, M., Premsrirut, P. K., Teruya-Feldstein, J., Hicks, J., Benveniste, H., Speicher, M. R., Downings, J. R. &, Lowe, S. W. (2014). PTEN action in leukaemia dictated by the tissue microenvironment. Nature, 510(7505), 402-406.

Montaner, D., Minguez, P., Al-Shahrour, F., & Dopazo, J. (2009). Gene set internal coherence in the context of functional profiling. BMC genomics, 10(1), 197.

Mukherjee, D., & Topol, E. J. (2002). Pharmacogenomics in cardiovascular diseases. Progress in cardiovascular diseases, 44(6), 479-498.

Negrini, S., Gorgoulis, V. G., & Halazonetis, T. D. (2010). Genomic instability—an evolving hallmark of cancer. Nature reviews Molecular cell biology, 11(3), 220-228.

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, T. J., Naoki, K., Sasaki, H., Fujii, Y., Eck, M. J., Sellers, W. R., Johnson, B. E., & Meyerson, M. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science, 304(5676), 1497-1500.

Park, J., Shim, J. K., Yoon, S. J., Kim, S. H., Chang, J. H., & Kang, S. G. (2019). Transcriptome profiling-based identification of prognostic subtypes and multi-omics signatures of glioblastoma. Scientific reports, 9(1), 1-11.

Peng, X. L., Moffitt, R. A., Torphy, R. J., Volmar, K. E., & Yeh, J. J. (2019). De novo compartment deconvolution and weight estimation of tumor samples using DECODER. Nature communications, 10(1), 1-11.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. Nature Reviews Genetics, 14(4), 288-295.

Peña-Chilet, M., Esteban-Medina, M., Falco, M. M., Rian, K., Hidalgo, M. R., Loucera, C., & Dopazo, J. (2019). Using mechanistic models for the clinical interpretation of complex genomic variation. Scientific reports, 9(1), 1-12.

Pomerantz, M. M., Ahmadiyeh, N., Jia, L. I., Herman, P., Verzi, M. P., Doddapaneni, H., Beckwith, C. A., Chan, J. A., Hills, A., Davis, M., Yao, K., Kehoe, S. M., Lenz, H., Haiman, C. A., Yan, C., Henderson, B. E., Frenkel, B., Barretina, J., Bass, A., Tabernero, J., Baselga, J., Regan, M. M., Manak, J. R., Shivdasani, R., Coetzee, G. A. & Freedman, M. L. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature genetics, 41(8), 882-884.

Pritykin, Y., Ghersi, D., & Singh, M. (2015). Genome-wide detection and analysis of multifunctional genes. PLoS Comput Biol, 11(10), e1004467.

Prunier, C., Baker, D., ten Dijke, P., & Ritsma, L. (2019). TGF-β family signaling pathways in cellular dormancy. Trends in cancer, 5(1), 66-78.

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C. L., Mroz, E. A., Emerick, K. S.,Deschler, D. G., Varvares, M. A., Mylvaganam, R., Rozenblatt-Rosen, O., Rocco, J. W., Faquin, W. C., & Lin, D. T. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell, 171(7), 1611-1624.

Qian, B. Z., & Pollard, J. W. (2010). Macrophage diversity enhances tumor progression and metastasis. Cell, 141(1), 39-51.

Rambow, F., Rogiers, A., Marin-Bejar, O., Aibar, S., Femel, J., Dewaele, M., Karras, P., Brown, D., Chang, Y. H., Debiec-Rychter, M., Adriaens, C., Radaelli, E., Wolter, P., Bechter, O., Dummer, R., Levesque, M., Piris, A., Frederick, D. T., Boland, G., Flaherty, K. T., Van den Oord, J., Voet, T., Aerts, S., Lund, A. W., & Marine, J. (2018). Toward minimal residual disease-directed therapy in melanoma. Cell, 174(4), 843-855.

Razzoli, M., Frontini, A., Gurney, A., Mondini, E., Cubuk, C., Katz, L. S., Cero, C., Bolan, P. J., Dopazo, J., Vidal-Puig, A., Cinti, S., & Bartolomolucci, A. (2016). Stress-induced activation of brown adipose tissue prevents obesity in conditions of low adaptive thermogenesis. Molecular metabolism, 5(1), 19-33.

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), 52.

Reiter, F., Wienerroither, S., & Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Current opinion in genetics & development, 43, 73-81.

Rhee, J. K., Jung, Y. C., Kim, K. R., Yoo, J., Kim, J., Lee, Y. J., Ko, Y. H., Lee, H. H., Cho, B. C., & Kim, T. M. (2018). Impact of tumor purity on immune gene expression and clustering analyses across multiple cancer types. Cancer immunology research, 6(1), 87-97.

Ridker, P. M., MacFadyen, J. G., Fonseca, F. A., Genest, J., Gotto, A. M., Kastelein, J. J., Koenig, W., Libby, P., Lorenzatti, A., Nordestgaard, B., Shepherd, J., Willerson, J. & Glynn, R. (2009). Number needed to treat with rosuvastatin to prevent first cardiovascular events and death among men and women with low low-density lipoprotein cholesterol and elevated high-sensitivity C-reactive protein: justification for the use of statins in prevention: an intervention trial evaluating rosuvastatin (JUPITER). Circulation: Cardiovascular Quality and Outcomes, 2(6), 616-623.

Roesch, A., Vultur, A., Bogeski, I., Wang, H., Zimmermann, K. M., Speicher, D., Körbel, C., Laschke, M. W., Gimotty, P. A., Philipp, S. E., Krause, E., Pätzold, S., Villanueva, J., Krepler, C., Fukunaga-Kalabis, M., Hoth, M., Bastian. B. C., Vogt, T., & Herlyn, M. (2013). Overcoming intrinsic multidrug resistance in melanoma by blocking the mitochondrial respiratory chain of slow-cycling JARID1Bhigh cells. Cancer cell, 23(6), 811-825.

Schneider, M. V., & Orchard, S. (2011). Omics technologies, data and bioinformatics principles. In Bioinformatics for omics Data (pp. 3-30). Humana Press.

Schödel, J., Bardella, C., Sciesielski, L. K., Brown, J. M., Pugh, C. W., Buckle, V., Tomlinson, I. P., Ratcliffe, P. J., & Mole, D. R. (2012). Common genetic variants at the 11q13. 3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. Nature genetics, 44(4), 420.

Schork, N. J. (2015). Personalized medicine: time for one-person trials. Nature, 520(7549), 609-611.

Sebastian-Leon, P., Vidal, E., Minguez, P., Conesa, A., Tarazona, S., Amadoz, A., Armero, C., Salavert, F., Vidal-Puig, A., Montaner, D., & Dopazo, J. (2014). Understanding disease mechanisms with models of signaling pathway activities. BMC systems biology, 8(1), 121.

Sexton, R. E., Hallak, M. N. A., Uddin, M. H., Diab, M., & Azmi, A. S. (2020). Gastric Cancer Heterogeneity and Clinical Outcomes. Technology in Cancer Research & Treatment, 19, 1533033820935477.

Shalgi, R., Lieber, D., Oren, M., & Pilpel, Y. (2007). Global and local architecture of the mammalian microRNA–transcription factor regulatory network. PLoS Comput Biol, 3(7), e131.

Sharma, S. V., Lee, D. Y., Li, B., Quinlan, M. P., Takahashi, F., Maheswaran, S., McDermott, U., Azizian, N., Zou, L., Fischbach, M. A., Wong, K. K., Brandstetter, K., Wittner, B., Ramaswamy, S., Classon, M., & Sttleman, J. (2010). A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell, 141(1), 69-80.

Shay, J. W., & Wright, W. E. (2000). Hayflick, his limit, and cellular ageing. Nature reviews Molecular cell biology, 1(1), 72-76.

Sidaway, P. (2017). Glioblastoma subtypes revisited. Nature Reviews Clinical Oncology, 14(10), 587-587.

Sherr, C. J., & McCormick, F. (2002). The RB and p53 pathways in cancer. Cancer cell, 2(2), 103-112.

Smyth, L. M., Piha-Paul, S. A., Won, H. H., Schram, A. M., Saura, C., Loi, S., Lu, J., Shapiro, G. I., Juric, D., Mayer, I. A., Arteaga, A. L., De la Fuente, M. I., Brufsky, A. M., Spanggaard, I., Mau-Sorensen, M., Arnedos, M., Moreno, V., Boni, V., Sohn, J., Schwartzberg, L. S., Gonzàlez-Farré, X., Cervantes, A., Bidard, F., Gorelick, A. N., Lanman, R. B., Nagy, R. J., Ulaner, G. A., Chandarlapaty, S., Jhaveri, K., Gavrila, E. I., Zimal, C., Selcuklu, S. D., Melcer, M., Samoila, A., Cai, Y., Scaltriti, M., Mann, G., Xu, F., Eli, L. D., Dujka, M., Lalani, A. S., Bryce, R., Badelga, J., Taylor, B. S., Solit, D. B., Meric-Bernstam, F., & Hyman, D. M. (2020). Efficacy and determinants of response to HER kinase inhibition in HER2-mutant metastatic breast cancer. Cancer discovery, 10(2), 198-213.

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., Bando, M., Ohno, S., Ishikawa, Y., Aburatani, H., Niki, T., Sohara, Y., Sugiyama, Y., & Mano, H. (2007). Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. Nature, 448(7153), 561-566.

Solinas, G., Vilcu, C., Neels, J. G., Bandyopadhyay, G. K., Luo, J. L., Naugler, W., Grivennikov, S., Wynshaw-Boris, A., Scadeng, M., Olefsky, J.M., & Karin, M. (2007). JNK1 in hematopoietically derived cells contributes to diet-induced

inflammation and insulin resistance without affecting obesity. Cell metabolism, 6(5), 386-397.

Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C., & Tavaré, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proceedings of the National Academy of Sciences, 110(10), 4009-4014.

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. Nature reviews genetics, 13(9), 613-626.

Stark, Z., Dolman, L., Manolio, T. A., Ozenberger, B., Hill, S. L., Caulfied, M. J., Levy, Y., Glazer, D., Wilson, J., Lawler, M., Boughtwood, T., Braithwaite, J., Goodhand, P., Birney, E., & North, K. N. (2019). Integrating genomics into healthcare: a global responsibility. The American Journal of Human Genetics, 104(1), 13-20.

Stefanick, M. L., Anderson, G. L., Margolis, K. L., Hendrix, S. L., Rodabough, R. J., Paskett, E. D., Lane, D. S., Hubbell, A., Assaf, A. R., Sarto, G. E., Schenken, R. S., Yasmeen, S., Lessin, L., & Chlebowski, R. T.(2006). Effects of conjugated equine estrogens on breast cancer and mammography screening in postmenopausal women with hysterectomy. Jama, 295(14), 1647-1657.

Su, Y., Wei, W., Robert, L., Xue, M., Tsoi, J., Garcia-Diaz, A., Moreno, B. H., Kim, J., Ng, R. H., Lee, J. W., Koya, R. C., Comin-Andoux, B., Graeber, T. G., Ribas, A., & Heath, J. R. (2017). Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance. Proceedings of the National Academy of Sciences, 114(52), 13679-13684.

Subbiah, V., Puzanov, I., Blay, J. Y., Chau, I., Lockhart, A. C., Raje, N. S., Wolf, J., Baselga, J., Meric-Bernstam, F., Roszik, J., Diamond, E. L., Riely, G. J., Sherman, E. J., Riehl, T., Pitcher, B., & Hyman, D. M. (2020). Pan-cancer efficacy of vemurafenib in BRAFV600-mutant non-melanoma cancers. Cancer discovery, 10(5), 657-663.

Suissa, S. (2015). Number needed to treat: enigmatic results for exacerbations in COPD.

Tang, W., Bertaux, F., Thomas, P., Stefanelli, C., Saint, M., Marguerat, S., & Shahrezaei, V. (2020). bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. Bioinformatics, 36(4), 1174-1181.

Testa, U., Pelosi, E., & Castelli, G. (2018). Colorectal cancer: genetic abnormalities, tumor progression, tumor heterogeneity, clonal evolution and tumor-initiating cells. Medical Sciences, 6(2), 31.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., Kazer, S. W., Gaillard, A., Kolb, K. E., Villani, A. Johannessen, C. M., Flaherty, K. T., Frederick, D. T., Jané-Valbuena, J., Yoon, C. H., Rozenblatt-Rosen, O., Shalek, A. K., Regev, A., & Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352(6282), 189-196.

Tracy, S., Yuan, G. C., & Dries, R. (2019). RESCUE: imputing dropout events in single-cell RNA-sequencing data. BMC bioinformatics, 20(1), 388.

Vaes, B. L., Ducy, P., Sijbers, A. M., Hendriks, J. M., van Someren, E. P., de Jong, N. G., Van den Heuvel, E., Olijve, W., Van Zoelen, E. J. J., & Dechering, K. J. (2006). Microarray analysis on Runx2-deficient mouse embryos reveals novel Runx2 functions and target genes during intramembranous and endochondral bone formation. Bone, 39(4), 724-738.

Vajdic, C. M., & van Leeuwen, M. T. (2009). Cancer incidence and risk factors after solid organ transplantation. International journal of cancer, 125(8), 1747-1754.

Valdes-Mora, F., Handler, K., Law, A. M., Salomon, R., Oakes, S. R., Ormandy, C. J., & Gallego-Ortega, D. (2018). Single-cell transcriptomics in cancer immunobiology: the future of precision oncology. Frontiers in immunology, 9, 2582.

van Dijk, D., Nainys, J., Sharma, R., Kaithail, P., Carr, A. J., Moon, K. R., Mazutis, L., Wolf, G., Krishnaswamy, S., & Pe'er, D. (2017). MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. BioRxiv, 111591.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics, 10(4), 252-263.

Veerla, S., Ringnér, M., & Höglund, M. (2010). Genome-wide transcription factor binding site/promoter databases for the analysis of gene sets and co-occurrence of transcription factor binding motifs. BMC genomics, 11(1), 145.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G.,

Perou, C. M., & Hayes, D. N. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell, 17(1), 98-110.

Wang, K., Grivennikov, S. I., & Karin, M. (2013). Implications of anti-cytokine therapy in colorectal cancer and autoimmune diseases. Annals of the rheumatic diseases, 72(suppl 2), ii100-ii103.

Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., & Zhang, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. Nature methods, 16(9), 875-878.

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., De Carvalho, A., Lyu, S., Li, P., Li, Y., Barthel, F., Cho, H. J., Lin, Y., Satani, N., Martinez-Ledesma, E., Zheng, S., Chang, E., Sauvé, C. G., Olar, A., Lan, Z. D., Finocchiaro, G., Phillips, J. J., Berger, M. S., Gabrusiewicz, K. R., Wang, G., Eskilsson, E., Hu, J., Mikkelsen, T., DePinho, R. A., Muller, F., Heimberger, A. B., Sulman, E. P., Nam, D., & Verhaak, R. G. W. (2017). Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. Cancer cell, 32(1), 42-56.

Wasserman, N. F., Aneas, I., & Nobrega, M. A. (2010). An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. Genome research, 20(9), 1191-1197.

Weinberg, R. A. (2013). The biology of cancer. Garland science.

WHO. World Health Organization. [Internet] Disponible en: https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf. Consulta: [10 Junio 2020]

Willems, L., Tamburini, J., Chapuis, N., Lacombe, C., Mayeux, P., & Bouscary, D. (2012). PI3K and mTOR signaling pathways in cancer: new data on targeted therapies. Current oncology reports, 14(2), 129-138.

Wilson, L. F., Antonsson, A., Green, A. C., Jordan, S. J., Kendall, B. J., Nagle, C. M., Neale, R. E., Olsen, C. M., Webb, P. M., & Whiteman, D. C. (2018). How many cancer cases and deaths are potentially preventable? Estimates for Australia in 2013. *International journal of cancer, 142*(4), 691-701.

Witsch, E., Sela, M., & Yarden, Y. (2010). Roles for growth factors in cancer progression. Physiology.

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., & & Quake, S. R. (2014).

Quantitative assessment of single-cell RNA-sequencing methods. Nature methods, 11(1), 41.

Yew, P. Y., Mushiroda, T., Kiyotani, K., Govindasamy, G. K., Yap, L. F., Teo, S. H., Lim, P. V., Govindaraju, S., Ratnavelu, K., Sam, C., Yap, Y. Y., Khoo, A. S., Pua, K., & Nakamura, Y. (2012). Identification of a functional variant in SPLUNC1 associated with nasopharyngeal carcinoma susceptibility among Malaysian Chinese. Molecular carcinogenesis, 51(S1), E74-E82.

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A., Carter, S. L., Getz, G., Stemke-Hale, K., Mills, G. B., & Verhaak, R. G. W. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications, 4(1), 1-11.

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., & Flicek, P. R. (2015). The ensembl regulatory build. Genome biology, 16(1), 56.

Zhang, L., & Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. IEEE/ACM transactions on computational biology and bioinformatics.

Zhou, X., Li, G., An, S., Li, W. X., Yang, H., Guo, Y., Dai, Z., Zheng, J., Huang, H., Iavarone, A., & Zhao, X. (2018). A new method of identifying glioblastoma subtypes and creation of corresponding animal models. Oncogene, 37(35), 4781-4791.

Zhou, Q., Li, T., Price, D. H. (2012). RNA polymerase II elongation control. Annual review of biochemistry, 81, 119-143.