UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

ALERTAS TEMPRANAS Y REDES EN CÁNCER

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA:

ANGEL JUAREZ FLORES

TUTOR PRINCIPAL DE TESIS: Dr. José Valenzuela Marco Antonio
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM.
COMITÉ TUTOR: Dr. García Carrancá Alejandro Manuel
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM.
Dr. Miramontes Vidal Pedro Eduardo
FACULTAD DE CIENCIAS, UNAM.

Ciudad Universitaria, CD. MX. Enero, 2021

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

ALERTAS TEMPRANAS Y REDES EN CÁNCER

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA:

ANGEL JUAREZ FLORES

TUTOR PRINCIPAL DE TESIS: Dr. José Valenzuela Marco Antonio
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM.
COMITÉ TUTOR: Dr. García Carrancá Alejandro Manuel
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM.
Dr. Miramontes Vidal Pedro Eduardo
FACULTAD DE CIENCIAS, UNAM.

Ciudad Universitaria, CD. MX. 2021

COORDINACIÓN DEL POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

OFICIO CPCB/010/2021

ASUNTO: Oficio de Jurado

M. en C. Ivonne Ramírez Wence
Directora General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **09 de noviembre de 2020** se aprobó el siguiente jurado para el examen de grado de **DOCTOR EN CIENCIAS** del estudiante **JUÁREZ FLORES ÁNGEL** con número de cuenta **308031968** con la tesis titulada **"Alertas tempranas y redes en cáncer",** realizada bajo la dirección del **DR. MARCO ANTONIO JOSÉ VALENZUELA,** quedando integrado de la siguiente manera:

Presidente:	Dr. Alejandro Alagón Cano
Vocal:	Dr. Felipe Vaca Paniagua
Secretario:	Dr. Pedro Eduardo Miramontes Vidal
Suplente:	Dra. Marcela Lizano Soberón
Suplente:	Dra. Aliesha Araceli González Arenas
Emergente:	Dr. Alejandro Manuel García Carrancá

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cd. Universitaria, Cd. Mx., a 06 de enero de 2021

COORDINADOR DEL PROGRAMA

DR. ADOLFO GERARDO NAVARRO SIGÜENZA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

AGRADECIMIENTOS INSTITUCIONALES

Agradezco al Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México (UNAM), por haberme brindado su apoyo para mi formación como doctor y la oportunidad de continuar el camino de la investigación científica.

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por su apoyo económico en mi formación como investigador durante mis estudios de doctorado en el Programa de Posgrado en Ciencias Biológicas de la UNAM a través del otorgamiento de una beca CONACYT (número 775924).

Al comité tutoral:
Dr. Marco Antonio José Valenzuela por la orientación recibida durante todo el doctorado para el crecimiento como investigador e individuo. También por el apoyo tanto de ideas como para la resolución de dudas y brindar literatura científica para el enriquecimiento de las investigaciones y mejor comprensión de técnicas y análisis, así como para la redacción de los artículos científicos.

Dr. Pedro Eduardo Miramontes Vidal por el apoyo para mi crecimiento como investigador, brindando críticas constructivas acerca de la investigación durante los exámenes tutorales e informes que permitieron una mejor consolidación de los resultados, así como el apoyo en mi preparación para el examen de candidatura y las críticas durante este para una mejora.

Dr. Alejandro Manuel García Carrancá por su apoyo para mi crecimiento como investigador, las críticas constructivas que permitieron mejorar el aterrizaje de conceptos, ideas y resultados durante mi investigación. También el apoyo en mi preparación para el examen de candidatura, la literatura brindada vía artículos científicos y sus consejos durante los exámenes tutorales.

AGRADECIMIENTOS A TÍTULO PERSONAL

Agradezco a los miembros del laboratorio de Biología Teórica: Gabriel, Miriam, Normita, Eréndira, Juan Román por sus consejos y críticas durante el desarrollo de mi investigación.

En el caso de Juan Román Bobadilla, adicionalmente le agradezco por su apoyo en el soporte técnico para el mantenimiento de los equipos hardware y software utilizados durante la realización de las investigaciones además de su asesoría.

A los miembros del jurado de mi examen de candidatura: Dr. Germinal Cocho†, Dra. Aliesha González Arenas, Dr. Pedro Eduardo Miramontes Vidal, Dr. Alfredo Hidalgo Miranda y el Dr. Ángel Alfonso Zarain Herzberg, cuyos comentarios permitieron la mejora del proyecto de investigación.

DEDICATORIAS

A mi mamá por su apoyo y cariño durante toda mi vida y el tiempo de mis estudios. Así por haber creído en mi potencial.

A Dios por brindarme la oportunidad de obtener este logro y la vida.

Al Sr. Daniel Farías y a la Sra. Patricia Farías por el apoyo brindado durante mis estudios, así como sus consejos.

A mi papá Q.E.P.D. por el apoyo que me llegaste a dar.

A mi prima Gaudencia y su esposo Sabás por su apoyo.

A los miembros de biología teórica por su apoyo y amistad: Gabriel, Miriam, Normita, Eréndira, Juan Román, Dr. Marco José, el Dr. Morgado.

A Kendra† y a Grace por el tiempo de distracción que me brindaron.

La imaginación es la clave del descubrimiento… Winston, Overwatch.

ÍNDICE

RESUMEN

El cáncer sigue siendo uno de los principales problemas de salud pública en el mundo, que afecta a un gran número de personas de cualquier edad. El término cáncer, comprende un grupo de enfermedades cuyo número de personas afectadas y muertas, aumenta cada año. Este grupo de enfermedades comparten un comportamiento en donde se da una división anormal de las células afectadas que tienen la capacidad de diseminarse y eventualmente llegar a lugares distantes dentro del cuerpo (proceso denominado metástasis). Además, presentan características similares entre ellas, por ejemplo: una independencia de factores del crecimiento, alargamiento de los telómeros, evasión de la muerte celular, evasión de la respuesta inmune entre otras. En el presente trabajo se usaron distintas estrategias y enfoques teóricos, aplicados a una serie de datos de expresión de genes de pacientes afectados por estas enfermedades, los cuales fueron obtenidos de distintas bases de datos y se encuentran públicos en sus respectivos repositorios. El trabajo de investigación se enfocó principalmente en cuatro tipos de cáncer, dentro de los cuales se incluyen: El hepatocarcinoma, el carcinoma de células escamosas de pulmón, el cáncer de páncreas y el melanoma. Cada uno de estos tienen una gran relevancia dado la mortalidad y número de pacientes afectados. El análisis permitió encontrar y proponer distintos conjuntos de genes dentro de cada uno de estos, que teóricamente pueden tener una gran relevancia dada la función biológica que tienen los productos de estos. También se propuso un nuevo término conceptual, cuya finalidad es apoyar para una mejor interpretación del análisis de redes cuyo origen es biológico y que en este trabajo logró fungir como un filtro adicional para determinar genes cuya importancia fuera relevante en el desarrollo del carcinoma de células escamosas de pulmón.

ABSTRACT

Cancer is one of the main public health problems in the world since it affects many people of any age; the word cancer refers to a group of diseases whose numbers of affected and death people increase every year. This group of diseases share a behavior where there is an abnormal division of the affected cells that have the ability to spread and eventually reach distant places within the body. In addition, they have similar characteristics among them, for example, an independence of growth factors, telomere elongation, avoidance of cell death, evasion of the immune response, among others. In this work, different strategies and theoretical approaches have been applied to a series of gene expression data from patients affected by these diseases, which were obtained from different databases and are publicly available in their respective repositories. The focus has been mainly on four types of cancer, including: hepatocarcinoma, squamous cell carcinoma of the lung, pancreatic cancer, and melanoma. Each of these are highly relevant given the mortality and number of affected patients. The Analysis found and proposed different sets of genes that theoretically have great relevance given their biological function that the products of these have. We propose a new conceptual term whose purpose is to support a better interpretation of the analysis of networks whose origin is biological. In this work we used it as an additional filter to unveil genes which importance are relevant for the squamous cell carcinoma of the lung development.

INTRODUCCIÓN GENERAL

El cáncer es uno de los principales problemas de salud pública de la humanidad. La Organización Mundial de la Salud (OMS), menciona que el cáncer es la segunda causa de muertes a nivel mundial. Cifras mostradas por la OMS en 2018 le atribuyen 9.6 millones de muertes. En México el cáncer es considerada la tercera causa de muerte (*El Cáncer En El Mundo y México*, s/f).

En estados Unidos, el Instituto Nacional del Cáncer (*NCI*, por sus siglas en inglés), define al cáncer de la siguiente manera: "Un grupo de enfermedades relacionadas en donde algunas de las células del cuerpo presentan una división anormal sin detenerse la cual llega a diseminarse en los tejidos adyacentes" (*What Is Cancer?*, s/f). Típicamente, hay una mayor tasa de división, una desregulación en el crecimiento y tiene la capacidad de diseminarse a los tejidos circundantes y, finalmente, a los tejidos distantes. (*What Is Cancer?*, s/f; *WHO | Cancer*, s/f). A través del tiempo gracias a las diversas investigaciones realizadas se han logrado develar una serie de características y procesos denominados "*hallmarks"* (palabra en inglés cuyo significado en español es "sellos") que comparten esta serie de enfermedades denominadas como cáncer. Los llamados *hallmarks* del cáncer, son considerados como pilares fundamentales para el desarrollo de esta enfermedad los cuales a su vez han sido utilizados para el desarrollo de estrategias farmacéuticas para el tratamiento. Seis *hallmarks* son los que en un inicio fueron contemplados y son los siguientes: proliferación sostenida, evasión de los supresores del crecimiento, invasión y metástasis, inmortalidad replicativa, angiogénesis y resistencia a la muerte celular. Adicionalmente se han propuesto dos nuevos *hallmarks*: desregulación del metabolismo celular, evasión de la eliminación por el sistema inmune, inestabilidad genómica y la inflamación promotora del tumor (Hanahan & Weinberg, 2011). Cada uno de estos *hallmarks* involucran una serie de vías y procesos biológicos, que en conjunto permiten el desarrollo y eventual diseminación de la enfermedad en un organismo. Las distintas maneras que aborda la investigación, para esta serie de enfermedades, han sido experimentales, sin embargo, con el advenimiento de la era informática, genómica y el avance científico y tecnológico en distintas áreas del conocimiento, una nueva serie de herramientas y estrategias se han ido incorporando de manera gradual para el apoyo en la investigación. Una de las nuevas ramas de la ciencia, que se crearon con dichos avances, fue la biología de sistemas la cual es una rama de la ciencia que integra técnicas de las matemáticas, física, química, ciencia de la computación, ingeniería y teoría de la información para modelar diversos fenómenos biológicos desde un punto de vista integral. El objetivo de la biología de sistemas es el entendimiento de un sistema biológico a través de modelos matemáticos y computacionales. Para el estudio de este sistema compuesto, es necesaria la identificación de las diversas interacciones entre los componentes del sistema; para cumplir dicho propósito la teoría de redes es una pieza fundamental (Marcus, 2008; Singh & Dhar, 2015). Las propiedades de las redes pueden ser modeladas y analizadas usando métodos computacionales, sin embargo, estos son indiferentes hacia lo que representan los nodos y las conexiones. Existen diversas mediciones estadísticas que permiten la comparación y caracterización de las redes complejas. Es posible determinar las implicaciones biológicas de la estructura de una red, la posible conexión entre el rol de una proteína dentro de la red y su importancia biológica como una medida de su esencialidad en la supervivencia de un organismo (Dehmer, 2011). Para esto último, es necesario un conocimiento y un dominio específico de la naturaleza, de lo que representan los nodos y las uniones entre estos (Ghasemi et al., 2014; Prokop & Csukás, 2013; Raman, 2010). Muchos de estos sistemas biológicos considerados como complejos, exhiben umbrales críticos (puntos de inflexión) en los cuáles la dinámica cambia abruptamente de un estado a otro. Ejemplos en medicina son: las fallas sistémicas espontáneas como los ataques de asma o las crisis epilépticas en donde se puede detectar un cambio súbito en la varianza de las señales obtenidas desde un electroencefalograma antes de que suceda un

episodio. Existen diversos sucesos que pueden ocurrir en una gran clase de sistemas conforme se aproximan a un punto crítico, y que son independientes de los detalles de cada sistema. Un ingrediente básico para un punto de inflexión es una retroalimentación positiva, esto quiere decir, que una vez que se ha alcanzado un punto crítico, éste impulsa el cambio hacia un estado alternativo (M. Scheffer et al., 2012; Marten Scheffer et al., 2009). Una red, tiene dos componentes básicos: nodos; los cuales pueden representar a un gen o una proteína, y las líneas que los unen, llamadas uniones, aristas o arcos, las cuales representan las interacciones entre los nodos (Dehmer, 2011). En la mayoría de las redes existen nodos que son más importantes o influyentes que otros. Dicha importancia puede ser cuantificada usando las medidas de centralidad u otras métricas (Borgatti, 2005; Ghasemi et al., 2014; Singh & Dhar, 2015). En el cáncer, el uso de redes ha permitido mejorar la comprensión de distintas vías moleculares, la relación entre distintos componentes de estas, y el proponer posibles usos para los hallazgos encontrados. Una de las medidas que ha resultado de utilidad en el análisis de redes en cáncer ha sido la entropía de la información, cuya definición conceptual proviene de la teoría de la información y la cual está relacionada al concepto de entropía en mecánica estadística (Cover & Thomas, 2006). La entropía puede definirse: como la auto-información de una variable aleatoria, una medida de la incertidumbre promedio de una sola variable en bits, el número de bits en promedio que se requieren para describir a una variable aleatoria o el número de preguntas necesarias para identificar a una variable (Cover & Thomas, 2006). Algunos de los trabajos que fungen como un antecedente a lo realizado en las publicaciones que se muestran en este trabajo, y en donde se utilizaron redes para el estudio del cáncer son: 1) un trabajo en donde al usar redes de vías desreguladas en cáncer, se relacionó la tasa de sobrevivencia en distintos tipos de cáncer con la entropía resultante del análisis de las redes de alteraciones moleculares correspondientes a cada cáncer. Se observó, que la entropía disminuye cuando aumenta la probabilidad de sobrevivencia a 5 años (Breitkreutz et al., 2012). Adicionalmente, usando una métrica conocida como Intermediación, propuso posibles blancos terapéuticos (Breitkreutz et al., 2012). 2) En otro estudio más reciente, se logró identificar señales de alerta temprana en las transiciones críticas durante el progreso de una enfermedad (sano-enfermo), usando datos de alto rendimiento del transcriptoma. En dicho estudio se observó un punto de inflexión (*Tipping point*), en el cual la entropía local de la red estudiada disminuye y es en esta fase donde se identificó un grupo de proteínas conocidas como red dinámica de biomarcadores (*DNB*, por sus siglas en inglés), los cuales son propuestos como los principales actores dentro de la estructura total de la red, dado que en estos nodos es donde existe una mayor disminución en la medida de entropía (Liu et al., 2012, 2012; Marten Scheffer et al., 2009). 3) En 2014, un grupo de investigadores propuso una medición la cual denominaron *signalling entropy,* como un enfoque capaz de analizar e interpretar datos ómicos, cuyo resultado permitió la discriminación de diferentes células dependiendo de su estado potencial de diferenciación en cáncer. Así como una relación entre esta medida y la resistencia de estas células a terapias (Teschendorff et al., 2014). 4) En otro trabajo de investigación al realizar un análisis de la Leucemia Mieloide Crónica aplicando el cálculo de la entropía, permitió separar diferentes estadios de progreso en la enfermedad (Brehme et al., 2016).

En el presente trabajo se ha dado uso de distintas estrategias y enfoques teóricos, aplicados a una serie de datos de distintas bases de datos, los cuales fueron obtenidos de manera experimental y se encuentran públicos en sus respectivos repositorios. El trabajo se ha enfocado en cuatro tipos de cáncer: 1) el carcinoma hepatocelular (HCC); es el tipo más común de cáncer primario de hígado. En 2018, datos del GLOBOCAN, muestran que este cáncer es la cuarta causa más común de muertes relacionadas al cáncer (Bray et al., 2018; Raza, 2014; Waller et al., 2015). La detección temprana, tiene un beneficio en la supervivencia del paciente. Sin embargo, el diagnóstico en la mayoría de las ocasiones se realiza en estadios avanzados (Balogh et al., 2016). El mejor tratamiento es el trasplante

de hígado; este tratamiento tiene una recurrencia entre el 10-20% (Waller, 2015). 2) El cáncer de páncreas; es uno de los tipos de cáncer más letales. En 2018, fue reportado como la séptima causa más frecuente de muertes relacionadas con el cáncer y cuya incidencia es cada vez mayor en países como: Estados Unidos de América, Nueva Zelanda y Europa (Bray et al., 2018). Algunas de las condiciones que se consideran incrementan el riesgo para desarrollar esta enfermedad son: fumar, la pancreatitis crónica, obesidad, diabetes y un historial familiar de parientes con cáncer de páncreas. En muchos casos el diagnóstico es realizado en los estadios más avanzados de la enfermedad. Sin embargo, cuando el diagnóstico es realizado en las etapas tempranas, la cirugía es el mejor tratamiento y cuando esta no es posible o se diagnóstica en los estadios más avanzados se usa quimioterapia y/o radiación (Moore & Donahue, 2019). 3) El carcinoma de células escamosas de pulmón (*Squamous cell carcinoma of the lung)*; es un tipo de cáncer de pulmón que representa aproximadamente el 30% de todos los casos de este tipo de cáncer. El cáncer de pulmón hasta 2018, era el tipo de cáncer con más nuevos casos y muertes en todo el mundo (Bray et al., 2018); usualmente se clasifica en dos grandes grupos al cáncer de pulmón: el primero es como cáncer de pulmón de células pequeñas y el segundo es como cáncer de pulmón de células no pequeñas (NSCLC, por sus siglas en inglés). El NSCLC representa el 85% de todos los cánceres de pulmón (Gridelli et al., 2015; Inamura, 2017), dentro de estos se encuentran el adenocarcinoma y el carcinoma de células escamosas como los más prevalentes (Drilon et al., 2012; Gandara et al., 2015). La tasa de supervivencia de los pacientes con cáncer de pulmón es inferior al 5% después de cinco años, y tienden a metastatizar, por lo que el diagnóstico precoz es importante, sin embargo, la mayoría de las veces el diagnóstico se realiza en los estadios avanzados (Derman et al., 2015; Heist et al., 2012). El mejor tratamiento es la cirugía, pero su efectividad está ligada a su uso en los estadios tempranos de la enfermedad (Herbst et al., 2018; Hirsch et al., 2017). El fumar cigarro está fuertemente asociado a la aparición de este tipo de cáncer (Gandara et al., 2015). 4) El melanoma; es un tipo de cáncer dermatológico que solo representa una pequeña proporción de todos esto, pero es la principal causa de muerte debido a cáncer de piel, con aproximadamente 80% del total. Además, ha tenido una tasa de incidencia creciente en los últimos años, la cuál ha sido asociada la exposición al sol que es considerado como el principal factor de riesgo (Miller & Mihm, 2006; Schadendorf et al., 2015). Es un tipo de cáncer que es detectable de manera visual sin la necesidad de una inspección invasiva y que una simple escisión puede curarlo en las etapas más tempranas. La auto examinación ha mostrado ser un método efectivo para apoyar el diagnóstico de esta enfermedad (O'Neill & Scoggins, 2019).

El principal objetivo de este trabajo fue encontrar algún suceso que funcionara como alerta temprana para la detección de cáncer, así como proponer moléculas cuyo posible uso fuera como blanco terapéutico o biomarcador. Lo anterior se realizó haciendo uso de los distintos repositorios de información públicos para su análisis. La primera publicación presentada en este trabajo, verso sobre un análisis *in silico* de datos de expresión genética de pacientes, obtenidos por microarreglos, en los cuatro tipos de cáncer previamente mencionados: hepatocarcinoma, cáncer de páncreas, melanoma y carcinoma de células escamosas de pulmón. En este trabajo, se hizo uso de la métrica conocida como entropía de la información para el descubrimiento de un posible grupo de genes que podrían estar involucrados en la transición precancerígena-cancerígena, así como en cada uno de los estadios del proceso carcinogénico. Estudios previos de algunos de los genes propuestos, han mostrado, que la modificación de su regulación puede tener consecuencias en el desarrollo de los respectivos tipos de cáncer en donde fueron estudiados. Además, se realizó una propuesta de genes totalmente nuevos, cuya función biológica normal indicaba que estos también pueden formar una parte importante para el desarrollo de la enfermedad; se espera que estos resultados puedan ayudar para captar el interés sobre estos genes para posteriores investigaciones a nivel experimental. La segunda publicación se centró en un análisis de datos de pacientes con el carcinoma de células escamosas de pulmón, este

hasta la fecha de su publicación y bajo el conocimiento de los autores, fue el primer tipo de análisis *in silico* realizado a este tipo de cáncer contemplando cada una de las fases del proceso carcinogénico, este análisis permitió comprender un poco más cada uno de los estadios, a través de la obtención de un grupo de genes desregulados en cada estadio. También permitió conocer con más detalle cuales son los procesos biológicos y funciones moleculares que teóricamente tienen una mayor relevancia dentro de cada estadio; también se introdujo un nuevo concepto el cual fue llamado *network gatekeepers* a través del análisis de una red generada mediante el apoyo de bases de datos de interacción proteína-proteína y sus valores de expresión.

Justificación del proyecto, hipótesis, objetivo general y objetivos particulares

El proyecto nace de la necesidad general que existe en el cáncer de encontrar nuevas formas que permitan la detección temprana, el tratamiento personalizado y moléculas de monitoreo, ya sea para el apoyo en el diagnóstico o tratamiento. El proyecto principalmente se centró en cuatro tipos de cáncer, los cuales en los últimos años han tenido un incremento en el número de casos y muertes debidos a estos.

La hipótesis que se formuló para este proyecto fue la siguiente: Si el cáncer es un sistema complejo, cuya expresión de genes e interacciones proteína-proteína (PPI) cambian dependiendo del estadio, entonces existe la posibilidad de detectar señales que funcionen como alertas tempranas integrando dichos datos en redes.

El objetivo general fue la búsqueda de una señal que permitiera la detección temprana en estos cuatro tipos de cáncer. Los objetivos particulares fueron dos: 1) la búsqueda de grupos de genes que permitieran la detección temprana y que pudieran participar en el desarrollo del cáncer. 2) Encontrar grupos de genes en cada una de las etapas del proceso carcinogénico cuya función biológica en su forma proteica pudieran estar relacionados con el desarrollo de dicha etapa y el avance a la siguiente.

Principales hallazgos encontrados en el Capítulo 1 y 2

En el capítulo 1 se utilizan los datos de expresión de genes de los diferentes estadios que conforman el proceso carcinogénico para cuatro tipos de cáncer. Usando un enfoque de redes aunado con un análisis de expresión diferencial y la medición de una medida conocida como entropía de la información, se logró proponer diferentes grupos de genes que pudieran estar relacionados con el cambio de transición entre los estadios pre-cancerígenos/cancerígenos; además por sí misma la señal obtenida de la entropía funciona como un apoyo para detectar el cambio. Los grupos de genes propuestos hasta ese momento no habían sido reportados de importancia para cada uno de los tipos de cáncer, sin embargo, estos no fueron los únicos detectados; otros grupos de genes previamente reportados se utilizaron como un apoyo de validación para los resultados.

En el capítulo 2 se realizó un análisis bioinformático enfocado al proceso carcinogénico del carcinoma de células escamosas de pulmón. Dicho análisis hasta ese momento y de acuerdo con nuestro conocimiento, fue el primero en su tipo al abordar todo el proceso carcinogénico de la enfermedad. En este proyecto se logró describir cada uno de los estadios basados en los grupos de genes y procesos biológicos que encontramos enriquecidos en el análisis aunado con un análisis de redes. Adicionalmente se realizó la propuesta de un grupo de genes acuñando el término de *gatekeepers* en el análisis de redes, para genes cuya importancia está en su función sobre el desarrollo del cáncer; sugerimos que estos genes funcionan como puertas para diferentes vías dada la conexión que tienen entre estos.

CAPÍTULO 1

Juarez-Flores, A., & José, M. (2018). Multivariate Entropy Characterizes the Gene Expression and Protein-Protein Networks in Four Types of Cancer. Entropy, 20(3), 154. https://doi.org/10.3390/e20030154

entropy

MDPI

Article

Multivariate Entropy Characterizes the Gene Expression and Protein-Protein Networks in Four Types of Cancer

Angel Juarez-Flores [1,2] and Marco V. José [2,*]

[1] Posgrado en Ciencias Biológicas, Unidad de Posgrado, Circuito de Posgrados, Ciudad Universitaria, Universidad Nacional Autónoma de México, CP 04510, Mexico City, Mexico; angel.juarez.bqd@gmail.com
[2] Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, CP 04510, Mexico City, Mexico
* Correspondence: marcojose@biomedicas.unam.mx

Received: 2 November 2017; Accepted: 23 February 2018; Published: 28 February 2018

Abstract: There is an important urgency to detect cancer at early stages to treat it, to improve the patients' lifespans, and even to cure it. In this work, we determined the entropic contributions of genes in cancer networks. We detected sudden changes in entropy values in melanoma, hepatocellular carcinoma, pancreatic cancer, and squamous lung cell carcinoma associated to transitions from healthy controls to cancer. We also identified the most relevant genes involved in carcinogenic process of the four types of cancer with the help of entropic changes in local networks. Their corresponding proteins could be used as potential targets for treatments and as biomarkers of cancer.

Keywords: multivariate entropy; cancer; protein-protein networks; gene expression; local networks; average network entropy; biomarkers; early warning

1. Introduction

Cancer is a generic term given to a collection of related diseases that can arise in every part of the organism. Typically, there is an increased division rate, dysregulation in growth and the capacity to spread into surrounding tissues and eventually on distant tissues. The latter process is known as metastasis, and is the main cause of death by cancer [1,2]. Entropy can be defined as the measure of the average uncertainty of a single random variable in bits, whereas the differential entropy is the entropy of a continuous random variable with an important characteristic that it only depends on the probability density of the random variable [3]. Unlike discrete entropy, the differential entropy can be negative [3]. Differential entropy can be conceived as the logarithm of the equivalent side length of the smallest set that contains most of the probability. Hence, low entropy implies that the random variable is confined to a small effective volume and high entropy indicates that the random variable is widely dispersed [3]. Entropy can be used as a descriptive and comprehensive measure of multivariate variability especially when data are non-Gaussian, since it can capture higher-order statistics and information content of the data [4].

Liver cancer is one of the main causes of cancer-related deaths worldwide [5]. Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer and the third most common cause of cancer-related deaths [6,7]. Its incidence and mortality continue to rise, chronic viral hepatitis and cirrhosis being some risk factors. Early screening has a survival benefit for patients. Available methods for HCC screening are radiographic, but unfortunately diagnosis is often made at advanced states of the disease, when effectiveness of treatment has poor prognosis. Sorafenib is the recommended treatment with patients in advanced stages, but due to its side effects it is difficult for the patient to tolerate it [8]. It is a disease with 10 to 20% recurrence even in patients with liver transplantation,

which is the most successful treatment [6]. There was a previous work in which entropy was used to detect an early warning in HCC, where the transition occurs at a very early stage of HCC [9].

Pancreatic cancer is one of the deadliest types of cancer. It was reported as the fourth cause of death cancer-related in developed countries in 2012. The GLOBOCAN estimations of new cases were 94,700 and 92,800 and estimated deaths were 93,100 and 91,300, in males and females, respectively. In 2015 for ductal adenocarcinoma, the most common type of pancreatic cancer, 367,000 new cases were diagnosed, and 359,000 patients died that same year. A pancreatic adenoma is followed by a neoplasm which leads to a carcinoma [10]. For diagnosis, there is currently no proper biomarker with enough specificity and sensitivity, yet the detection of the earliest stages of pancreatic cancer are urgently needed to improve the outcomes of resection, which is so far, the best treatment. Surgery is the only option for cure but only 10–15% of the newly diagnosed patients are eligible [11]. Due to the resistance of pancreatic cancer to therapy, even with resection, most of the patients will relapse and eventually succumb to the disease [12].

Lung cancer has the highest mortality rate. The rate of survival of patients with lung cancer is less than 5% after five years, and it tends to metastasize, thereby early diagnosis is important. Squamous cell carcinoma represents approximately 30% of all the cases [13,14]. Usually the stages of carcinogenesis are: squamous metaplasia → dysplasia → carcinoma in situ → squamous cell carcinoma of lung [15,16]. Squamous cell carcinoma (SCC) of the lung is the most common histologic subtype of non-small cell lung cancer (NSCLC). It accounts for 400,000 new cases annually worldwide, with cigarette smoking as the principal risk factor. For advanced stages, the standard care consists of a platinum-based doublet as a palliative systemic therapy [17]. Diagnosis is made by histological analysis of small biopsies or cytological specimens as fine needle aspirates or bronchial brushings. There are new promising therapies that improve patient's survival [14].

Melanoma is only one among various dermatological cancers, but it is the principal cause of death due to skin cancer, accounting for approximately 80%, and with an increasing incidence in the last years with sun exposure as the main risk factor [18,19]. All melanomas originate from melanocytes which represents a minority of the cell population within the basilar epidermis, but it can also be found in hair follicles and other tissues. Melanocytes provide the pigment melanin to their neighboring keratinocytes. Pigment production is stimulated by UV radiation-induced DNA damage to keratinocytes [20]. The gold standard for diagnosis is histopathological assessment [19]. There is an ideal model of melanoma carcinogenesis ranging from benign naevi → dysplastic naevi → melanoma in situ → invasive melanoma [20].

Teschendorff et al. [21] proposed signaling entropy as a novel approach for analyzing and interpreting omics data, such as discriminating cells according to their differentiation potential and cancer status. In other works, some driver genes were found to be associated with reductions in network entropy [9,21,22]. More recently, Brehme et al. [23] carried out an analysis in chronic myeloid leukemia using entropy dynamics and separated the progression stages. They revealed an important difference in the chronic phase (CP) which allowed to separate it into two phases: "early" from "late" CP [23]. That same year Park et al. [24], using an entropy-based distance metric, were able to successfully measure the intratumor heterogeneity and propose it as a useful tool to characterize it at the RNA level using transcriptome and network information.

Cancer research has a wide variety of approaches, mainly for therapeutics usage, ranging from analysis of massive data, such as the genome or transcriptome, to more detailed analyses as of single genes or proteins that are involved in cancer hallmarks [25]. A major challenge is the opportune detection of the early stages of the disease, which is the most desirable scenario, with better options for patients treatment and an improved outcome as seen in pancreatic cancer and lung cancer, where cancer is commonly detected in the last stages of the disease [12–14]. Finding new therapeutics targets are important due to cancer resistance to therapies and to have more repertoires of targets that could be modulated by immunotherapy, miRNAs therapy, gene therapy, or other treatments [26,27].

In general, cancer progression can be divided into three states [9]: a normal state, a pre-disease state (or a critical state), and a disease state. In the normal state, the disease is under control (immune system) and dynamically it has high resilience and robustness to perturbations. The pre-disease state is defined as the limit of the normal state, which occurs before the imminent phase transition point is reached, but it has low resilience and robustness due to its dynamical structure [9]. The disease state represents a seriously deteriorated stage possibly with high resilience and robustness, where the system usually finds it difficult to recover or return to the normal state even after treatment, which contrasts with the pre-disease state. Therefore, it is crucial to detect the pre-disease state to prevent qualitative deterioration and to further elucidate its molecular mechanism. In cancer, it is a daunting task to predict a pre-disease state because the state of the system may change little before the bifurcation point or the critical transition is reached. There may be slight differences between the normal and pre-disease. The detection of early-warning signals can involve a myriad of genetic factors. There are leading networks in critical transitions, which make the first move from a normal state to a disease state [9]. The leading network is the first subnetwork that breaks down the limit of a normal state to move into a disease state, which means that they are clearly related to the causal or driving genes in a disease network, in contrast to the differential gene expression that results from the disease. Therefore, identifying the leading networks during a critical transition can signal the emergence of a pre-disease state to make the early diagnosis on the disease, and help to disentangle the mechanisms of disease initiation and progression at the network level. The leading networks are dynamical signals that herald the pre-disease state, rather than the disease state detected by the traditional static biomarkers.

Herein, we contend that multivariate entropy is a useful filter for detecting driver genes. Therefore, we calculated the multivariate entropy of the gene expression profiles in four types of cancer, to wit, pancreatic cancer, melanoma, HCC, and squamous cell carcinoma of the lung. For these cancers, we also constructed their corresponding protein-protein interaction (PPI) networks considering the disease stages. We calculated the multivariate entropy for the local networks of PPI from which we estimated the average entropy of PPI networks. In general, the reliable identification of local leading networks and pre-disease stages is not easy to achieve with noisy data and a small number of samples. The identification of biomarkers and the critical states may be inaccurate. In this work, we validated our proposed biomarkers using different statistical tests, such as double filter for the differentially gene expression, from which we selected the genes for constructing the local networks. The Wilcoxon rank sum test was used to test the statistical differences in the average network entropy between the pre-disease and diseased states with the normal state. We searched the biological function of those genes whose entropy changed and some of them are already considered potential therapeutic targets, for example see [28]. With our analyses, we also found new potential targets whose biological functions are relevant to the normal cell function. We successfully identified the most relevant genes involved in the carcinogenic processes of the four types of cancer with the help of entropic changes in local networks. Their corresponding proteins could be used as potential targets for treatments and as biomarkers of cancer.

2. Materials and Methods

Four series of raw transcriptomic data of the carcinogenesis process were retrieved from the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI). The first one: hepatocellular carcinoma with GEO accession: GSE6764, which has 75 tissue samples (platform: Affymetrix Human Genome U133 Plus 2.0 Array [29]). The second: melanoma, GEO accession: GSE4587, with 18 samples (platform: Affymetrix Human Genome U133 Plus 2.0 Array [30]). The third: pancreatic cancer, GEO accession: GSE19650, with 22 samples (platform: Affymetrix Human Genome U133 Plus 2.0 Array [31]). The fourth: squamous cell carcinoma of the lung, GEO accession: GSE33479, with 122 samples (platform: Agilent-014850 Whole Human Genome Microarray 4 $\times$ 44 K G4112F).

We used the R software with Limma package [32] according to its manual and the Bioconductor Manual to preprocess and process all the transcriptomic data. Only melanoma and squamous cell

carcinoma of the lung were retrieved preprocessed (background corrected, normalized) from GEO. The Limma package was used for differentially gene expression analysis obtaining the respective adjusted *p*-value for False Discovery Rate (*FDR*) [32,33]. Differential gene expression analysis was made between each stage and the normal one. Fold changes were also calculated, and we used a double filter to select the differentially expressed genes. Criteria consist in selecting genes with an adjusted p-value < 0.05 and Fold Change > 1.5 [32–35]. To create the networks of Protein-Protein interactions (PPIs) of each cancer, we used the APID database [36] which provides protein interaction data for a wide variety of species with a controlled quality using PPI found by experimental evidence. We used data set from quality level 1 (all known interactions). Herein, we used the *Homo sapiens* data which were processed and cleaned using the Cytoscape Software version 3.2 [37]. Cleanup consisted in the deletion of data of other species based on its taxonomic tag, and in the deletion of duplicated edges and nodes. The proteins retrieved from the database were the ones detected by the double filter applied to the differentially gene expression but due to the lack of information of the Human interactome, we only retrieved at least 50% of the proteins coded by the genes detected by the double filter analysis. Then, we obtained the first neighbors for each node in the network using Rcy3 [38] which allows a connection between R and Cytoscape. The result was used to create the *local networks* of each node (protein). We tested the hypothesis that the distribution of expression levels across all the selected genes for the four cancers followed a normal or a log-normal distribution (Appendix A) [39]. The density function of the multivariate normal distribution is given by [3,38]:

$$f(x) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\}, \quad (1)$$

where x is a random vector, μ is the mean vector, and Σ is the covariance matrix.

In each stage, the local networks data were matched with their respective gene expression of each sample to create matrices of *local networks* where the genes expression were the rows and each column a sample. For each *local network* matrix, a covariance matrix was calculated and then we applied Equation (2) to obtain the multivariate entropy of each *local network*. Each sample is a set of vectors X_1, X_2 and so on. The expression level of the genes are elements of the vectors.

We also group nodes (proteins) to create subnetworks. Based on the maximum and minimum calculated entropy values of the healthy states, we established the limit from where a preset range was applied as follows: HCC in ranges of 10 units resulting in 11 subnetworks, Melanoma in ranges of three units resulting in eight subnetworks, Pancreatic in ranges of five units resulting in nine subnetworks, Squamous cell lung carcinoma in ranges of 10 units resulting in 19 subnetworks. The starting point was the 0 unit. This healthy-generated groups were kept in the four types of cancer resulting in that each successive stage has the same groups with same nodes with a unique variation in their entropy values. To calculate the entropy for each local network, we consider the entropy of multivariate normal distribution given by [3,40]:

$$h(X_1, X_2, \ldots, X_n) = h(N_n(\mu, K)) = \frac{1}{2}\log_2(2\pi e)^n|\kappa|, \quad (2)$$

where $|\kappa|$ is the determinant of the covariance matrix; X is the random variable (set of vectors); n is the size of the covariance matrix.

Entropy of a local network is calculated with Equation (3), where a single value is calculated from the data with all genes within the same local network, and colored rectangles (Supplementary Materials) represent the entropy change for the *local network*.

The average network entropy, $H(t)$ is given by:

$$H(t) = -\frac{1}{n}\sum_{i=1}^{n} h_i(t), \quad (3)$$

where n is the number of nodes in the network (differentially expressed genes) and $h_i(t)$ is the entropy of the local network i. Equation (3) is the same as the one used in [9], but with a negative sign to obtain positive values of entropy and to have units of information in bits.

Gene expression entropies (Figure 2, Tables 1–4) were calculated using all selected genes by the double filter to create a single matrix for every stage in each cancer, without using the *local networks* data. The level expression of the genes are elements of the vectors. From the matrix, a covariance matrix was calculated and then Equation (2) was applied to obtain the gene expression entropies.

For each entropy calculation, the following number of samples in each stage were used: HCC: seven samples; melanoma: two samples; pancreatic cancer: three samples; squamous cell lung cancer: 12 samples. The identification was made by looking at the color changes in the nodes. Colors represent a range of entropy values, so if a node has a change in color from one to another we record it.

2.1. Statistical Analysis of Local Network Entropy

Wilcoxon Rank Sum test was used to compare the results of local network entropy in each stage of the carcinogenic process versus its respective control for the four types of cancers (see Appendix A). Calculations were made with the R package. This is a non-parametric statistical test to compare two populations and allows to determine if two medians have significant differences.

2.2. Local Networks

Multivariate entropy of gene expression profiles is influenced only by gene expression (Figure 2). Multivariate entropy of local networks of PPIs is also influenced by gene expression with the addition of another level of complexity: the PPIs of each node. Entropy values of local networks will be influenced by which nodes constitute the local network and their respective gene expression values (Figure 1)

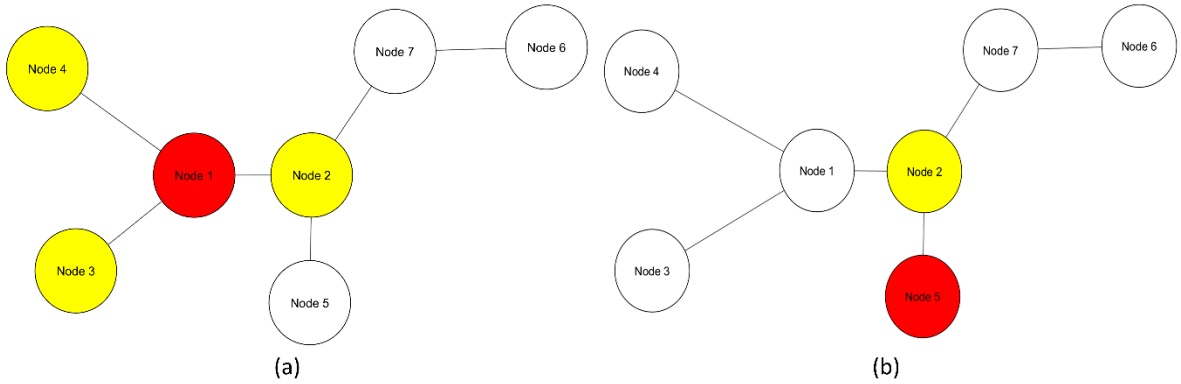

Figure 1. Graph of basic network with seven nodes. (**a**) A local network of node 1 (red) consists of its first neighbors (yellow) and node 1; (**b**) Local network of node 5 (red) is constituted by its first, and in this case, only neighbor node 2 (yellow) and node 5 itself.

3. Results

Herein we illustrate how entropy changes throughout the distinct stages of the carcinogenesis process in the four types of cancer. We used Equation (2) to obtain the global change in entropy using the gene expression of the selected genes by the double filter criteria. We found that each cancer stage displays a characteristic entropic value when compared with pre-cancerous stages.

We observe that the last part of the carcinogenesis processes, the gene expression possesses the largest positive entropy than all previous stages with the only exception of melanoma in situ. The latter could be ascribed to different pathways and inherent heterogeneities among the different regions of the body where biological carcinogenesis of melanoma ensues (Figure 2 and Tables 1–4) [20].

In Figure 2a we observed variations in entropy values throughout all carcinogenic process of Melanoma, with the most evident change occurring at in situ stage in which its entropy value reaches a minimum. This change can be used as a first cancer early warning. Notice in Figure 2b that in HCC carcinogenic process the entropy values in pre-cancerous stages were decreasing until a sudden change in the transition point from a pre-cancerous to cancer stage. This change is in agreement with a first cancer early warning [9]. In Figure 2c we observe for pancreatic cancer that entropy increases

In Figure 2a we observed variations in entropy values throughout all carcinogenic process of Melanoma, with the most evident change occurring at in situ stage in which its entropy value reaches a minimum. This change can be used as a first cancer early warning. Notice in Figure 2b that in HCC carcinogenic process the entropy values in pre-cancerous stages were decreasing until a sudden change in the transition point from a pre-cancerous to cancer stage. This change is in agreement with a first cancer early warning [9]. In Figure 2c we observe for pancreatic cancer that entropy increases as the carcinogenic process progresses. A sudden change occurs between intraductal papillary-mucinous adenoma and intraductal papillary-mucinous neoplasm stages that are not cancer stages and therefore we need to be careful to talk about a cancer early warning. In Figure 2d we observe that for squamous cell lung carcinoma, entropy is increasing gradually in the pre-cancerous stages. The last stages correspond to cancer and they appear to have a greater change in entropy with an obvious change in entropy between carcinoma in situ and the squamous cell carcinoma of the lung. The change between severe dysplasia and carcinoma in situ could be a cancer early warning.

Entropy **2018**, *20*, x FOR PEER REVIEW 6 of 27

and therefore we need to be careful to talk about a cancer early warning. In Figure 2d we observe that for squamous cell lung carcinoma, entropy is increasing gradually in the pre-cancerous stages. The last stages correspond to cancer and they appear to have a greater change in entropy with an obvious change in entropy between carcinoma in situ and the squamous cell carcinoma of the lung. The change between severe dysplasia and carcinoma in situ could be a cancer early warning.

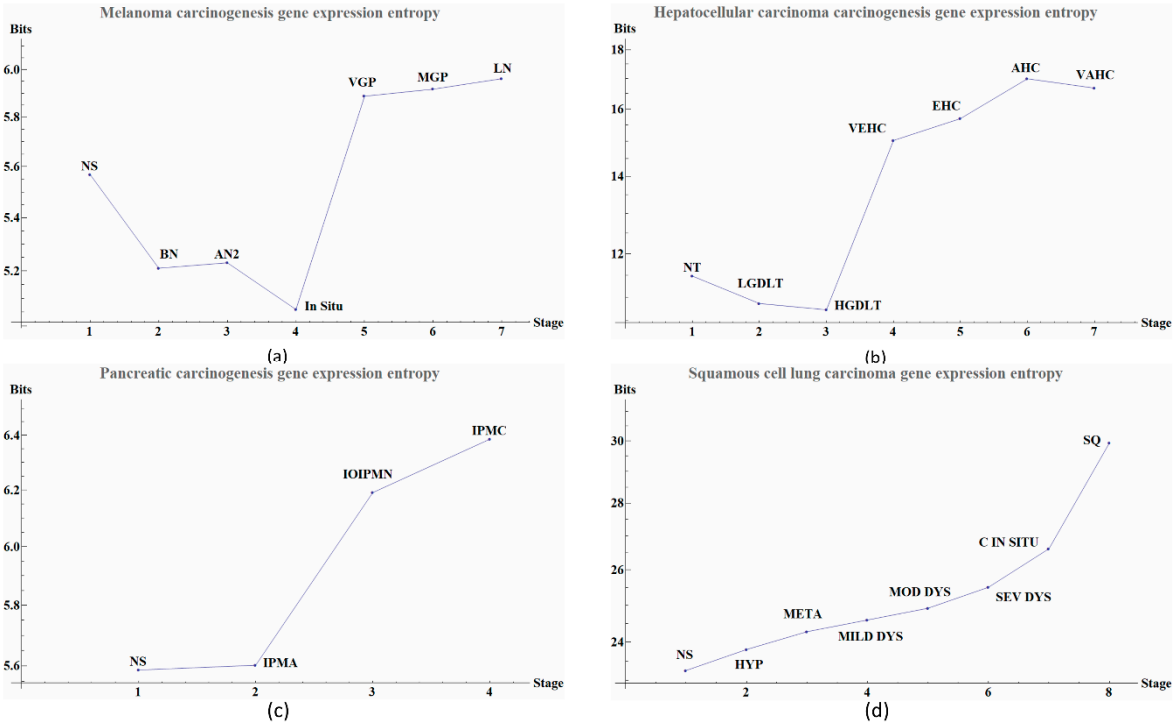

Figure 2. Gene expression-entropy of four types of cancer. Scale of the axes vary depending upon the type of cancer and sample size. (**a**) Melanoma carcinogenesis gene expression-entropy. Each point in the graph denote a stage during carcinogenic process the last four points denote cancer and the previous are pre-cancerous stage. The meanings of the abbreviated labels are given in Table 1. (**b**) Hepatocellular carcinoma carcinogenesis gene expression-entropy. Each point in the graph denote a stage during carcinogenic process the last four points denote cancer and the previous are pre-cancerous stage. The labels abbreviations are given in Table 2. (**c**) Pancreatic carcinogenesis gene expression-entropy. Each point in the graph denote a stage during carcinogenic process the last point denote cancer and the previous are pre-cancerous stage. The label abbreviation meanings are given in Table 3. (**d**) Squamous cell lung carcinoma carcinogenesis gene expression-entropy. Each point in the graph denote a stage during carcinogenic process the two points denote cancer and the previous are pre-cancerous stage. The label abbreviation meanings are given in Table 4.

Table 1. Melanoma carcinogenesis process and their respective entropy values.

Abbreviation	Stage	Entropy Value
NS	Normal skin	5.5685
BN	Benign nevi	5.2100
AN2	Atypical nevi	5.2303
In situ (INS)	Melanoma in situ	5.0595
VGP	VGP melanoma Vertical growth phase melanoma	5.8874

Table 1. Melanoma carcinogenesis process and their respective entropy values.

Abbreviation	Stage	Entropy Value
NS	Normal skin	5.5685
BN	Benign nevi	5.2100
AN2	Atypical nevi	5.2303
In situ (INS)	Melanoma in situ	5.0595
VGP	VGP melanoma Vertical growth phase melanoma	5.8874
MGP	MGP melanoma Metastatic growth phase melanoma	5.9169
LN	Lymph node metastasis	5.9612

Table 2. Hepatocellular carcinoma carcinogenesis process and their respective entropy values.

Abbreviation	Stage	Entropy Value
NT	Normal skin	11.4778
LGDLT	Low grade dysplasia	10.8693
HGDLT	High grade dysplasia	10.7351
VEHC	Very early hepatocellular carcinoma	15.0311
EHC	Early hepatocellular carcinoma	15.6988
AHC	Advanced hepatocellular carcinoma	16.9940
VAHC	Very advanced hepatocellular carcinoma	16.6807

Table 3. Pancreatic carcinogenesis process and their respective entropy values.

Abbreviation	Stage	Entropy Value
NS	Normal main pancreatic duct	5.5868
IPMA	Intraductal papillary-mucinous adenoma	5.6022
IOIPMN	Intraductal papillary-mucinous neoplasm	6.1912
IPMC	intraductal papillary-mucinous carcinoma	6.3856

Table 4. Squamous carcinogenesis process and their respective entropy values.

Abbreviation	Stage	Entropy Value
NS	Normal	23.2530
HYP	Hyperplasia	23.8010
META	Metaplasia	24.2762
MILD DYS	Mild dysplasia	24.5936
MOD DYS	Moderate dysplasia	24.9167
SEV DYS	Severe dysplasia	25.5035
C IN SITU	Carcinoma in situ	26.6175
SQ	Squamous cell carcinoma	29.9367

To find out which genes had major changes in every stage and how they are related between them, we constructed a protein-protein network for each carcinogenic process and assigned their respective local entropies for each stage. Then we calculated the average network entropy using the local entropy values and built groups based in health controls and for each stage of the carcinogenic process. The PPIs networks of melanoma are displayed in Figure 3. The PPIs networks with detail of melanoma, HCC, pancreatic cancer and squamous cell carcinoma of the lung are displayed in Supplementary Materials (Figures S1–S26) The entropy values were ranked and graded by colors. Notice that color variations in each group correspond to variations in entropy.

The calculated network entropy of PPIs and the entropy values calculated from expression data are positive (Figures 2 and 4). The average network entropy of local networks of melanoma and HCC exhibit a concave pattern in its entropy values of only gene expression, although the magnitudes are greater for HCC than for melanoma possibly due to the number of samples in each case. The interesting

behaviors of the calculated average network entropy are seen in Figure 4c,d. In Figure 4c of the pancreatic carcinogenic process, we can identify an early warning due to a sudden change observed between the non-cancer to cancer stage which is statistically significant by Wilcoxon rank sum test with continuity correction (p-value = 4.537×10^{-7}, see Appendix A), albeit this was not observed with its entropy values from only gene expression. Something similar occurs with Figure 2d, in which there is a smooth tendency of increasing entropy as the process progresses, but this pattern is not the same with the one observed in Figure 3d. In this case, there is an evident variation among stages with three major changes in metaplasia with an increased entropy value, carcinoma in situ stage with a slight decrease compared with SQ in which entropy decreases drastically and this change is statistically significant (calculated p-value = 0.0001274, see Appendix A). The melanoma carcinogenic process (Figure 3a) also denotes variations. There is an important change, melanoma in situ stage increases its entropy and the following stages decrease it and these changes are statistically significant ($p = 2.2 \times 10^{-16}$, see Appendix A). Not statistically significant changes for HCC were found (data shown). Dunnett's test were applied for each cancer with no statistically significant results (Appendix A, Tables A5–A8). For these cases, the average network entropy is useful as a first observation, whereas the most important data come from the local network entropy. Local network entropy values permit us to dissect each cancer to visualize the most important variations at each stage.

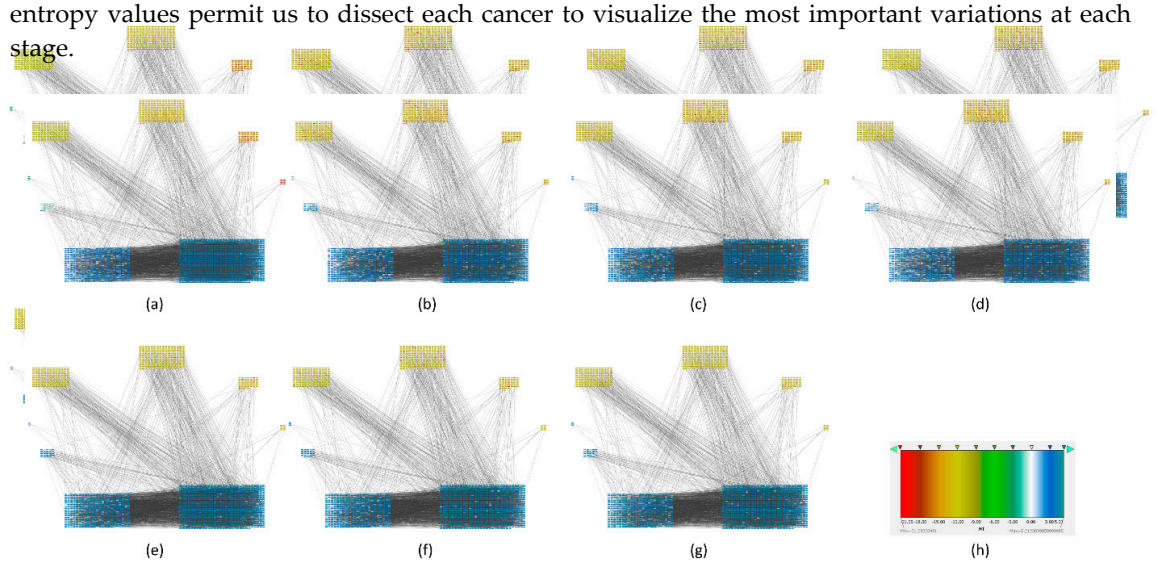

Figure 3. Melanoma carcinogenesis PPIs with local entropy. PPIs networks in which every node has an associated value that corresponds to its local network entropy. (**a**) Normal PPIs network (**b**) Benign nevi PPIs network (**c**) Atypical nevi PPIs network (**d**) In situ melanoma PPIs network (**e**) Vertical growth phase melanoma PPIs network (**f**) Metastatic growth phase melanoma PPIs network (**g**) Lymph node metastasis PPIs network (**h**) Represent scale of entropy and the color assigned to every group.

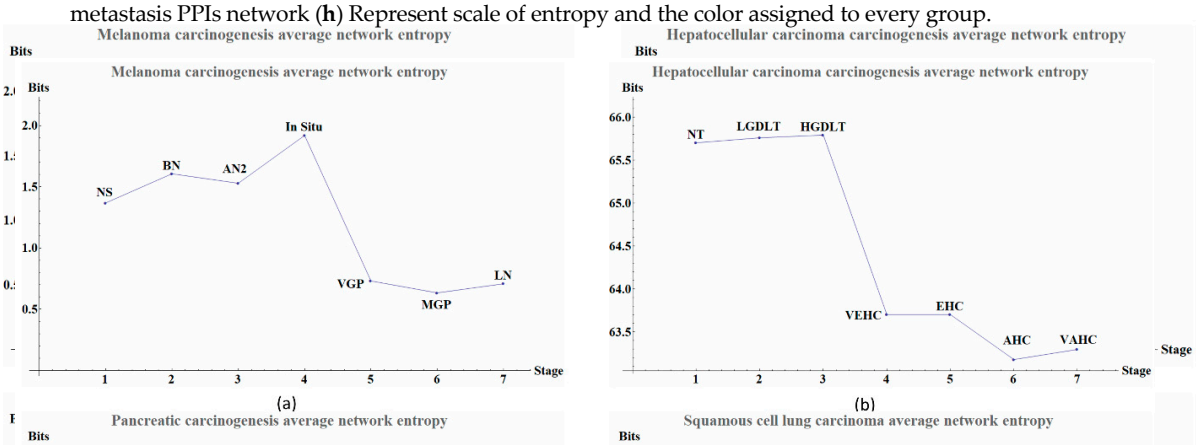

Figure 4. *Cont.*

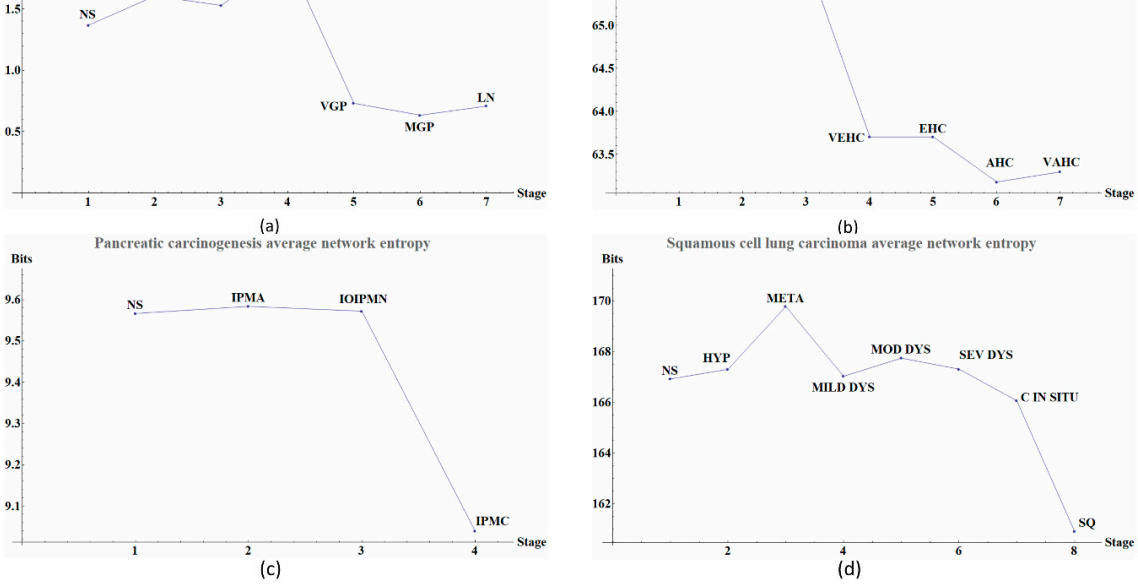

Figure 4. Average network entropy by local networks of four types of cancers. Scale of the axes vary according to the type of cancer and sample size. (**a**) Melanoma carcinogenesis average network entropy by local networks. Graph shows entropy values in each stage. Each point in the graph denotes a stage during carcinogenic process the last four points denote cancer and the previous are pre-cancerous stage. The label abbreviation meanings are given in Table 5 with their respective value for each stage. (**b**) Hepatocellular carcinoma carcinogenesis average network entropy by local networks. Graph shows entropy values in each stage. Each point in the graph denotes a stage during carcinogenic process the last four points denote cancer and the previous are pre-cancerous stage. The meanings of the label abbreviations are in given in Table 6 with their respective value for each stage. (**c**) Pancreatic carcinogenesis average network entropy by local networks. Graph shows entropy values in each stage. Each point in the graph denotes a stage during the carcinogenic process. The last point denotes cancer and the previous points are pre-cancerous stages. The abbreviation meanings are given in Table 7 with their respective value for each stage. (**d**) Squamous cell lung carcinoma average network entropy by local networks. Graph shows entropy values in each stage. Each point in the graph denotes a stage during carcinogenic process the last two points denote cancer and the previous are pre-cancerous stage. The abbreviation meanings are given in Table 8 with their respective values for each stage. Some descriptive statistics as the interquartile range, standard error and median about average network entropy are found in Appendix A (Tables A1–A4). Some characteristics of *local networks* of each stage, as the number of local networks and size of them between groups, are found in Appendix B (Tables A13–A16).

Table 5. Melanoma carcinogenesis process and their respective average network entropy values from PPIs.

Abbreviation	Stage	Entropy Value
NS	Normal skin	1.3677
BN	Benign nevi	1.6061
AN2	Atypical nevi	1.5287
INS	Melanoma in situ	1.9190
VGP	VGP melanoma Vertical growth phase melanoma	0.7308
MGP	MGP melanoma Metastatic growth phase melanoma	0.6339
LN	Lymph node metastasis	0.7081

Table 6. Hepatocellular carcinogenesis process and their respective average network entropy values from PPIs.

Abbreviation	Stage	Entropy Value
NT	Normal skin	65.7023
LGDLT	Low grade	65.7600
HGDLT	High grade	65.7916
VEHC	Very early HCC	63.7020
EHC	Early HCC	63.7023
AHC	Advanced HCC	63.1783
VAHC	Very advanced HCC	63.2960

Table 7. Pancreatic cancer carcinogenesis process and their respective average network entropy values from PPIs.

Abbreviation	Stage	Entropy Value
NS	Normal main pancreatic duct	9.5664
IPMA	Intraductal papillary-mucinous adenoma	9.5835
IOIPMN	Intraductal papillary-mucinous neoplasm	9.5720
IPMC	Intraductal papillary-mucinous carcinoma	9.0398

Table 8. Squamous cell lung carcinoma carcinogenesis process and their respective average network entropy values from PPIs.

Abbreviation	Stage	Entropy Value
NS	Normal	166.9220
HYP	Hyperplasia	167.3054
META	Metaplasia	169.7816
MILD DYS	Mild dysplasia	167.0286
MOD DYS	Moderate dysplasia	167.7428
SEV DYS	Severe dysplasia	167.3106
C IN SITU	Carcinoma in situ	166.0658
SQ	Squamous cell carcinoma	160.9167

Some genes like CYP2C9, FDX1, MUT, VAMP4, IL33, EMP2, DENND4A have drastic changes in its entropy during the transition from atypical nevi to melanoma in situ, as observed in Figure S10 and S11 (see Supplementary Materials). In the case of IL33, which is implicated in maturation of Th2 cell and the activation of MPK signaling pathway through IL1RL1/ST2 receptor and this pathway improves cell proliferation [41]. Previous studies tried to associate this protein with cancer promotion, but different results were found [28]. CYP2C9 polymorphisms were associated with colorectal cancer risk [42] and may influence breast cancer [43]. EMP2 is a protein implicated in progression and survival in endometrial cancer, and recently it was proposed as a possible oncoprotein [44]. We highlight that there are not studies in cancer for FDX1, MULT, VAMP4 and DENND4A genes or its protein products.

We found some genes in the transition of HGDLT to VECH in HCC like SOX6, ASPH, UBAP2L, CEP41. SOX6 encodes a transcription factor with a key role in developmental processes and it has been associated to HCC progression by its decreased progression [45]. UBAP2L was associated with the metastatic ability in some HCC cell lines via SNAIL1 [46]. Recently ASPH was suggested as a potential biomarker in gliomas [47]. There are no studies of the role of CEP41 in cancer. In pancreatic cancer, the transition between IOIPMN to IPMC includes the following genes: FAR1, CEACAM1, HCCS. CEACAM1 has 11 different splice variants, as reported in some studies in vivo, and restoration of its expression abolishes oncogenicity of tumor cell lines, but when it is expressed de novo it increases the risk of metastasis [48] CEACAM1 has also been proposed as a potential biomarker for breast cancer [49]. There are no studies for HCCS and FAR1 in cancer. In the transition of squamous severe dysplasia to carcinoma in situ, the genes UBET2 PIH1D2, KIF23 showed a change in their entropy. KIF23 is a protein essential for cytokinesis in Rho-mediated signaling. Its overexpression is associated with lung cancer cell growth and has been suggested as a novel therapeutic target for patients with advanced lung cancer and primary lung tumors [50]. A recent study showed that a knockdown of UBET2 induced an inhibition in the progression of gastric cancer in vivo and in vitro via WNT signal pathway [51]. There are no studies of PIH1D2 in cancer. A summary of the proposed genes for each cancer is shown in Table 9.

Table 9. Proposed genes as potential biomarkers or therapeutic targets.

Cancer	Proposed Genes
Melanoma	FDX1, MUT, VAMP4, DENND4A
HCC	CEP41
Pancreatic cancer	FAR1, HCCS
Squamous cell carcinoma of the lung	PIH1D2

4. Discussion

In this work, we have succeeded in the identification of the most relevant genes involved in carcinogenic processes of four types of cancer with the help of entropic changes in local networks. We are validating the use of multivariate entropy by testing that the distributions of gene expressions can follow either a normal distribution (SCC of the lung and melanoma) or a log-normal distribution (HCC and pancreatic cancer) (Appendix A).

We found that cancer entropic values from the average network entropy were lower than the observed values in healthy controls, which agrees with previous analysis [21,22]. The entropic values of the networks at the final stages for all examined cancers fully comply with the latter observation. However, not all pre-advanced cancer stages followed this behavior as we observed in melanoma in situ where entropy is higher than the control. Overall, entropy values correlate to each cancer stage. Interestingly, entropy values from only gene expression have a different behavior than those from the PPIs. In some cases, they could look like a mirror image of each other as is the case of HCC, but this is not always the case as was observed in pancreatic cancer or squamous cell carcinoma of the lung. As we mentioned previously, the differential entropy can be negative [3].

Our analysis of the carcinogenic processes allowed us to identify initial stages of the four types of cancer at which entropy changed with respect the control. A similar early warning measured with entropy for the average network was observed in HCC but this change was not statistically significant [9]. In pancreatic cancer, however, we found that the sudden change in average network entropy is statistically significant.

We also characterized each stage by local network entropy of its genes, which permitted to identify the most important genes in each stage and the ones in the transition between them. We illustrated the transitions from pre-cancer to cancer stage in which some of the found genes have been previously reported and even proposed to be early biomarkers in cancer by experiments in vivo and/or in vitro [47,49]. We also proposed new genes as biomarkers and as potential therapeutic targets (Table 1) that have not been previously reported: for melanoma: FDX1 (essential for the synthesis of various steroid hormones [41]), MUT (involved in degradation of several amino acids, odd-chain fatty acids and cholesterol via propionyl-CoA to the tricarboxylic acid cycle [41]), VAMP4 (involved in the pathway that functions to remove an inhibitor of calcium-triggered exocytosis during the maturation of secretory granules [41]), DENND4A (promotes the exchange of GDP to GTP, converting inactive GDP-bound Rab proteins into their active GTP-bound form [41]); for HCC: CEP41 (required during ciliogenesis for tubulin glutamylation in cilium [41]); for pancreatic cancer: FAR1 (catalyzes the reduction of saturated and unsaturated C16 or C18 fatty acyl-CoA to fatty alcohols [41]) and HCCS (stress-activated component of a protein kinase signal transduction cascade and regulates the JNK and p38 pathways [41]); for squamous cell carcinoma of the lung: PIH1D2 (exhibits a Ral GTPase binging which means a selectively interacting and non-covalently with Ral protein [41]). Our proposed genes as potential biomarkers or therapeutic targets for melanoma and pancreatic cancer must be taken with caution due to their small sample sizes. Due to its biological relevance, we hope that our results of local networks strongly inspire further experimental work for testing the proposed genes as biomarkers and/or therapeutic targets. In the Supplementary Materials, we provide entropic values for all the local networks of genes from healthy stage to all stages of cancer. This work was focused only in protein coding genes and the PPIs of its products, that represent less than 1.5% of the human

genome [52]. There is a wide field of cancer research such as non-coding-DNA/RNA, single nucleotide polymorphisms, copy number variations, and epigenetic factors such as methylation and acetylation, that could lead us to a better understanding of the dynamics of cancer diseases.

Supplementary Materials: The following are available online at http://www.mdpi.com/1099-4300/20/3/154/s1.

Acknowledgments: Angel Juarez Flores is a doctoral student from Programa de Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México (UNAM) and a fellowship recipient from Consejo Nacional de Ciencia y Tecnología (CONACYT) (number: 775924) and this paper constitutes a partial fulfilment of the Graduate Program in Biological Science of the UNAM. MVJ was financially supported by DGAPA PAPIIT-IN224015, UNAM, México. We thank Juan R. Bobadilla for his technical computer support. We thank the anonymous reviewers for their helpful criticisms and suggestions.

Author Contributions: Conceived the whole work: Angel Juarez-Flores and Marco V. José. Gather and processing data Angel Juarez-Flores; Performed calculations: Angel Juarez-Flores; Conducted literature reviews Angel Juarez-Flores and Marco V. José; Wrote the paper: Angel Juarez-Flores and Marco V. José. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Statistical Analyses

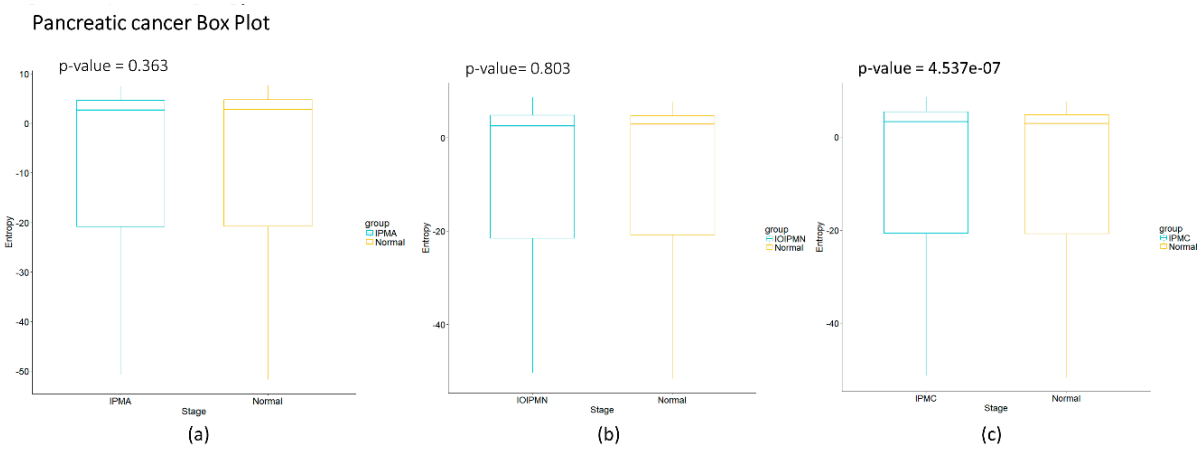

Figure A1. Pancreatic carcinogenesis process boxplot. The graph shows the comparison of the different stages vs. the control with boxplot and a *p*-value corresponding to the Wilcoxon rank sum test. (**a**) Intraductal papillary-mucinous adenoma; (**b**) Intraductal papillary-mucinous neoplasm; (**c**) Intraductal papillary-mucinous carcinoma.

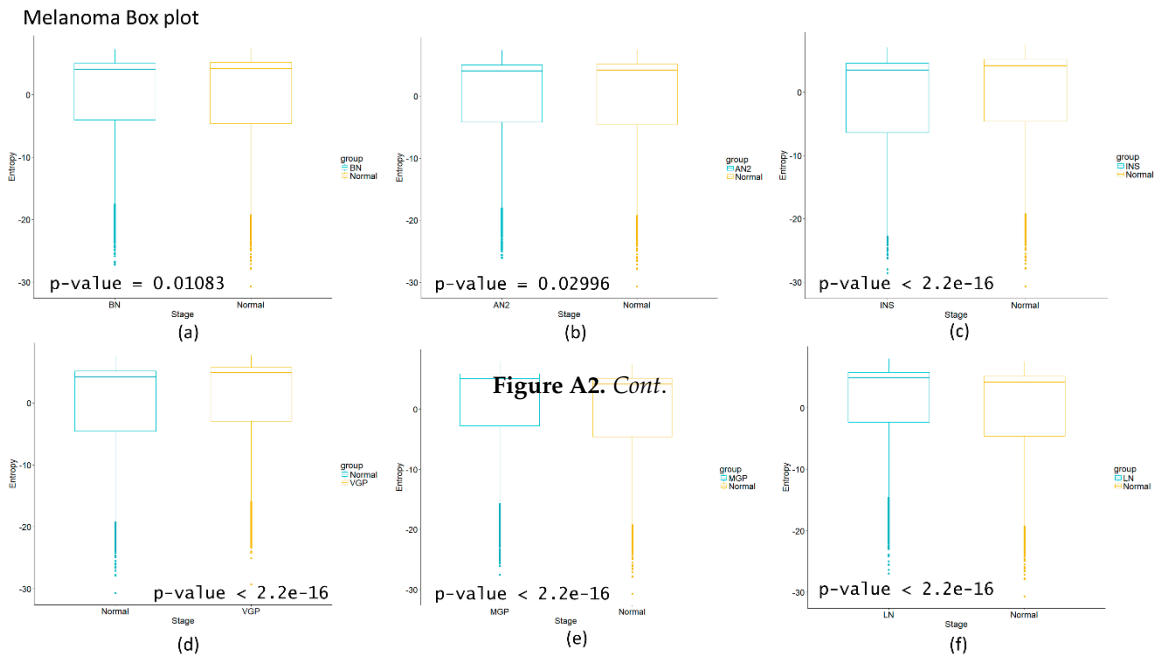

Figure A2. Melanoma carcinogenesis process boxplot. The graph shows the comparison of the different stages vs. the control with boxplot and a *p*-value corresponding to the Wilcoxon rank sum

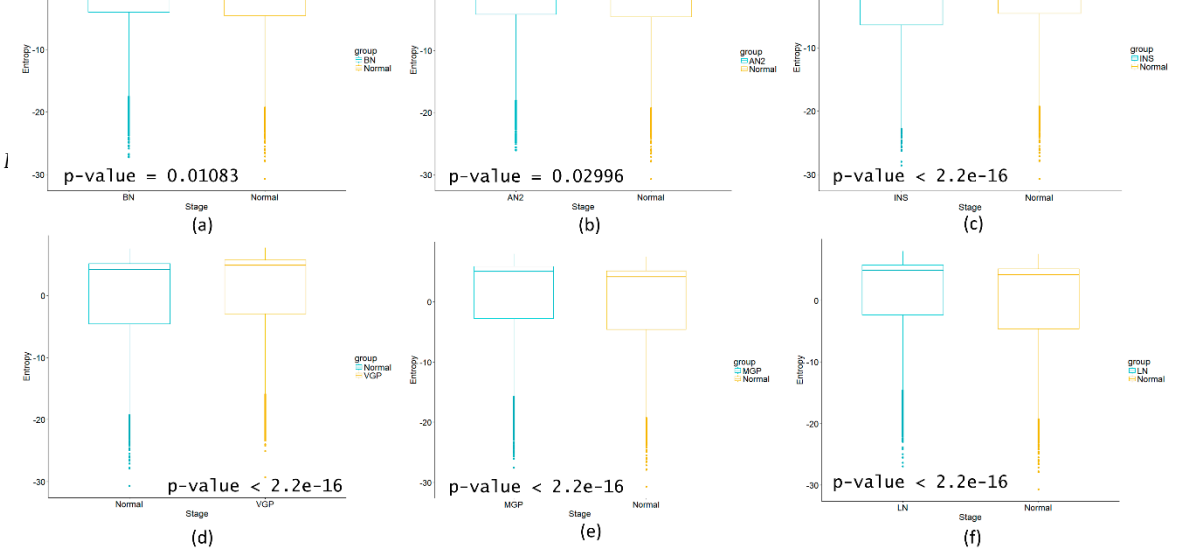

Figure A2. Melanoma carcinogenesis process boxplot. The graph shows the comparison of the different stages vs. the control with boxplot and a *p*-value corresponding to the Wilcoxon rank sum test. **(a)** Benign nevi; **(b)** Atypical nevi; **(c)** In situ melanoma; **(d)** Vertical growth phase melanoma; **(e)** Metastatic growth phase melanoma; **(f)** Lymph node metastasis.

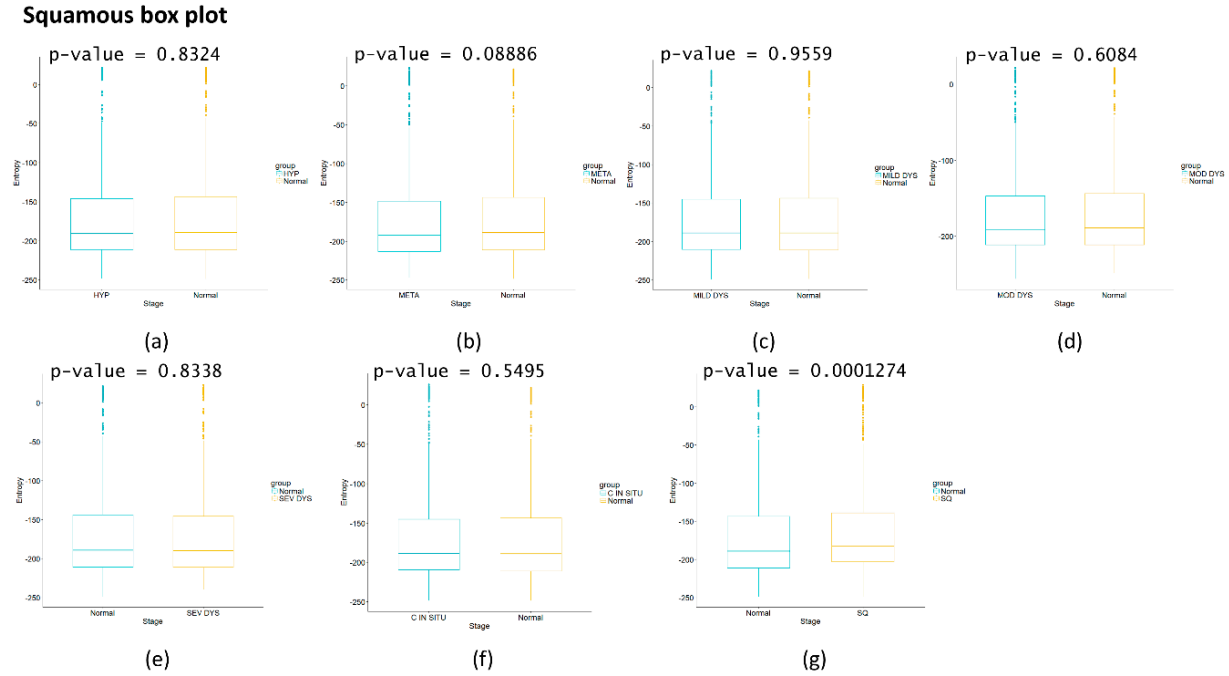

Figure A3. Squamous carcinogenesis process boxplot. The graph shows the comparison of the different stages vs. the control with boxplot and a *p*-value corresponding to the Wilcoxon rank sum test. **(a)** Hyperplasia; **(b)** Metaplasia; **(c)** Mild displasia; **(d)** Moderate displasia; **(e)** Severe Dysplasia; **(f)** Carcinoma in situ; **(g)** Squamous cell carcinoma.

Appendix A.1. Descriptive Statistics for the Average Network Entropy of Each Stage in the Four Types of Cancer

Statistics estimations were made using the R software and we used the following packages: bootstrap, plotrix, dplyr [53–55].

Table A1. Melanoma statistics.

Stage	Mean	Standard Error	MAD	Median	IQR	Jack.se When Applied to Mean	Jack.se When Applied to Variance
Normal	−1.367775	0.2259220	2.022601	4.154535	9.776100	0.225922	2.814388
BN	−1.606108	0.2303059	2.001127	4.019503	9.047087	0.2303059	2.896202
AN2	−1.528727	0.2292635	1.941641	4.050469	9.248341	0.2292635	2.849997
INS	−1.919017	0.2262065	2.086082	3.455508	10.95625	0.2262065	2.813802
VGP	−0.7308951	0.2282546	1.690730	4.886908	8.654485	0.2282546	2.815323
MGP	−0.6339521	0.2299159	1.659544	5.048833	8.677472	0.2299159	2.873484
LN	−0.7081158	0.2259220	2.022601	4.154535	9.77610	0.2281064	2.805363

MAD: Median Absolute Deviation, IQR: Interquartile Range. Jack.se: Jackknife standard error.

Table A1. Melanoma statistics.

Stage	Mean	Standard Error	MAD	Median	IQR	Jack.se When Applied to Mean	Jack.se When Applied to Variance
Normal	−1.367775	0.2259220	2.022601	4.154535	9.776100	0.225922	2.814388
BN	−1.606108	0.2303059	2.001127	4.019503	9.047087	0.2303059	2.896202
AN2	−1.528727	0.2292635	1.941641	4.050469	9.248341	0.2292635	2.849997
INS	−1.919017	0.2262065	2.086082	3.455508	10.95625	0.2262065	2.813802
VGP	−0.7308951	0.2282546	1.690730	4.886908	8.654485	0.2282546	2.815323
MGP	−0.6339521	0.2299159	1.659544	5.048833	8.677472	0.2299159	2.873484
LN	−0.7081158	0.2259220	2.022601	4.154535	9.77610	0.2281064	2.805363

MAD: Median Absolute Deviation, IQR: Interquartile Range. Jack.se: Jackknife standard error.

Table A2. Pancreatic cancer statistics.

Stage	Mean	Standard Error	MAD	Median	IQR	Jack.se When Applied to Mean	Jack.se When Applied to Variance
Normal	−9.566407	0.4283494	3.899366	2.901261	25.45503	0.4283494	8.116081
IPMA	−9.583562	0.4229802	4.222622	2.726233	25.45141	0.4229802	7.979798
IOIPMN	−9.572038	0.4243606	4.743738	2.520672	26.37048	0.4243606	7.967539
IPMC	−9.039864	0.4288906	4.647821	3.350183	26.11237	0.4288906	8.078527

MAD: Median Absolute Deviation, IQR: Interquartile Range. Jack.se: Jackknife standard error.

Table A3. Squamous cell carcinoma statistics.

Stage	Mean	Standard Error	MAD	Median	IQR	Jack.se When Applied to Mean	Jack.se When Applied to Variance
NS	−166.9220	2.07917	37.68617	−189.2606	67.45387	2.07917	242.6845
HYP	−167.3055	2.07897	35.41448	−190.7849	65.29746	2.07897	244.6484
META	−169.7816	2.075519	3538473	−192.9719	64.74367	2.075519	247.9728
MILD DYS	−167.0286	2.078943	37.16981	−189.2085	65.54929	2.078943	245.2105
MOD DYS	−167.7429	2.083012	34.75794	−191.3746	64.21930	2.083012	248.3033
SEV DYS	−167.3107	2.082539	35.95474	−189.9894	65.61071	2.082539	247.5942
C IN SITU	−166.0658	2.079962	36.71084	−188.7635	64.23117	2.079962	247.5762
SQ	−160.9168	2.05378	36.48999	−182.8355	63.88627	2.05378	238.1753

MAD: Median Absolute Deviation, IQR: Interquartile Range. Jack.se: Jackknife standard error.

Table A4. HCC statistics.

Stage	Mean	Standard Error	MAD	Median	IQR	Jack.se When Applied to Mean	Jack.se When Applied to Variance
NT	−45.54143	1.322529	37.53280	−51.37492	62.92547	1.908006	74.42024
LGDLT	−45.58139	1.319797	37.67159	−51.74393	62.86078	1.904064	74.35039
HGDLT	−45.60330	1.319819	37.25316	−5189218	63.02723	1.904096	74.39116
VEHC	−44.15488	1.353999	39.67826	−50351675	64.91867	1.953408	77.8041
EHC	−44.15512	1.353269	38.7171	−51.31032	64.92842	1.952355	77.72415
AHC	−43.79192	1.377034	40.60674	−50.27078	67.11002	1.98664	80.5827
VAHC	−43.87345	1.370975	40.65539	−50.52407	66.22477	1.977898	79.51231

MAD: Median Absolute Deviation, IQR: Interquartile Range. Jack.se: Jackknife standard error.

Appendix A.2. Dunnett's Test for the Mean

Dunnett's test was applied to compare the means of each stage vs. the control for the four types of cancer and effect size is also reported. No significant statistical results were found.

Cohen's formula was used in the calculation of effect size:

$$d = M_1 - M_2 / s_{pooled}$$

where:

$M_1 - M_2$ is the difference between the group mean (M)
M_1 is the mean of the control group
M_2 is the mean of either group that is not the control

s is the pooled standard deviation:

$$s_{pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}$$

where:

SD_1 is the standard deviation of the control group
SD_2 is the mean of either group that is not the control

Table A5. Melanoma Dunnett Contrasts and effect size.

Hypotheses	*t*-Value	Pr(>|t|)	Effect Size
Normal-BN	−0.738	0.947	−0.0233
Normal-AN2	−0.499	0.992	−0.0158
Normal-INS	−1.707	0.333	−0.0544
Normal-VGP	1.973	0.202	0.0626
Normal-MGP	2.273	0.105	0.07191
Normal-LN	2.043	0.175	0.0649

Table A6. Pancreatic cancer Dunnett Contrasts and effect size.

Hypotheses	*t*-Value	Pr(>|t|)	Effect Size
Normal-IPMA	−0.028	1.000	−0.000915
Normal-IOIPMN	−0.009	1.000	−0.000299
Normal-IPMC	0.874	0.713	0.027897

Table A7. Squamous cell carcinoma of the lung Dunnett Contrasts and effect size.

Hypotheses	*t*-Value	Pr(>|t|)	Effect Size
Normal-HYP	−0.131	1.000	−0.00629
Normal-META	−0.974	0.873	−0.04697
Normal-MILD DYS	−0.036	1.000	−0.00175
Normal-MOD DYS	−0.280	1.000	−0.01346
Normal-SEV DYS	−0.132	1.000	−0.00638
Normal-C IN SITU	0.292	1.000	0.01404
Normal-SQ	2.045	0.193	0.09915

Table A8. HCC Dunnett Contrasts and effect size.

Hypotheses	*t*-Value	Pr(>|t|)	Effect Size
Normal-LGDLT	−0.021	1.000	−0.00124
Normal-HGDLT	−0.033	1.000	−0.00192
Normal-VEHC	−0.729	0.950	0.04268
Normal-EHC	0.729	0.950	0.04269
Normal-AHC	0.919	0.868	0.05339
Normal-VAHC	0.877	0.890	0.05102

Appendix A.3. Goodness of Fit to Test Log-Normal Distribution for Pancreatic Cancer and HCC

Gene expression data were tested for a log-normal distribution or a normal distribution. The following results are only for one sample of each stage in the four types of cancer. The other samples render similar results with significant approximations to the tested distributions.

In pancreatic cancer and HCC Kolmogorov-Smirnov test was made to probe the log-normal distribution.

H0: The sample follows a log-normal distribution
Ha: The sample does not follow a log-normal distribution
If $D_n < D_{n\alpha}$ then the data is a good fit with the log-normal distribution.

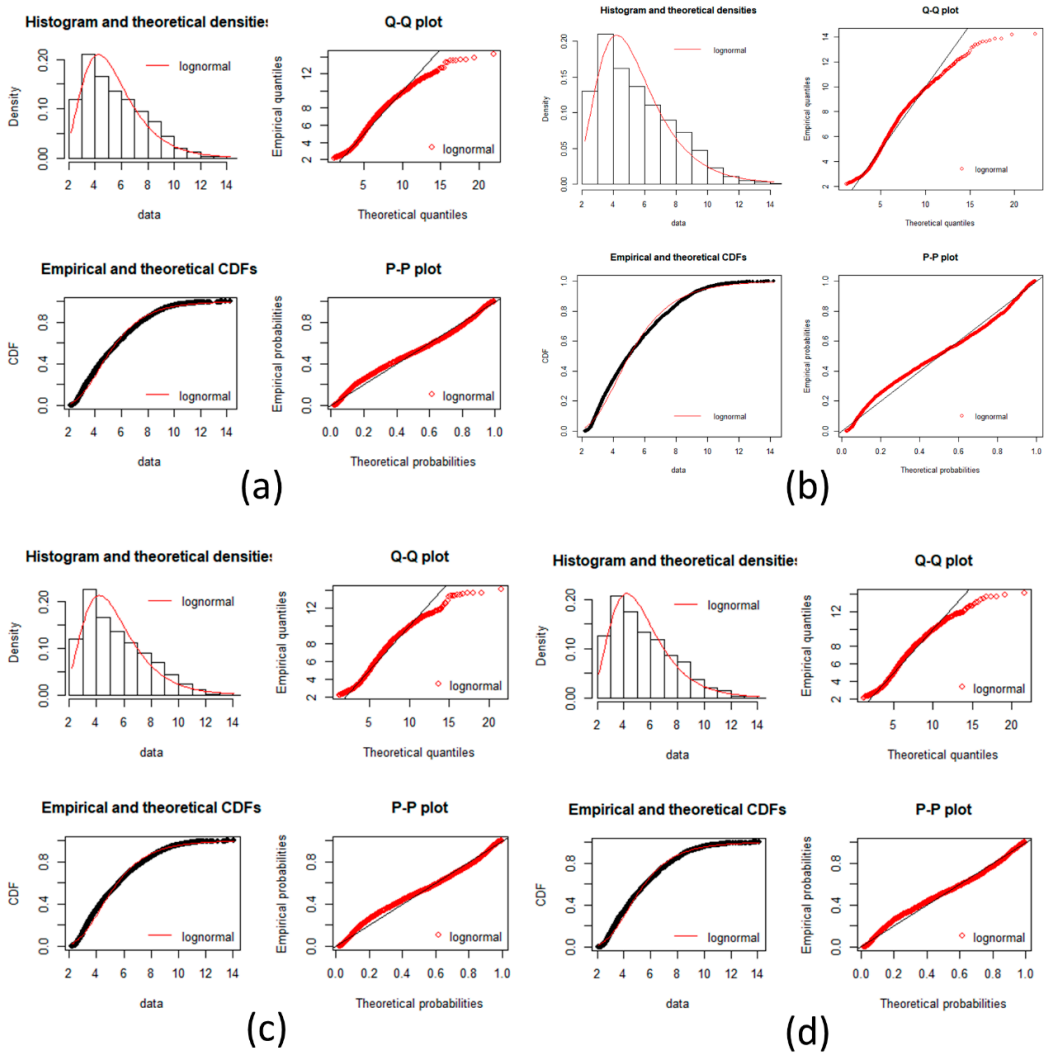

Figure A4. Pancreatic cancer goodness of fit test plot. Plots of goodness fits test for a log-normal distribution is showed in each incise for gene expression data. Histogram, Q-Q plot, CDFs and P-P plots are common plots to test a fitness of distribution. It can be seen that each stage fit well for a log normal distribution. (**a**) Normal stage; (**b**) IPMA stage; (**c**) IOIPMN stage; (**d**) IPMC stage.

In the plots, there is a very good fit with the log normal distribution.

$\alpha = 0.05$ for an $n > 35$; $D_{n,\alpha} = \frac{\sqrt{-0.5\ln\left(\frac{\alpha}{2}\right)}}{\sqrt{n}}$; $D_{n,\alpha} = \frac{1.358}{55.25} = 0.024$; n is the number of genes used.

$\alpha = 0.05$ for an $n > 35$; $D_{n,\alpha} = \frac{\sqrt{-0.5\ln\left(\frac{\alpha}{2}\right)}}{\sqrt{n}}$; $D_{n,\alpha} = \frac{1.358}{55.25} = 0.024$; n is the number of genes used.

Table A9. Kolmogorov-Smirnov test for Pancreatic cancer.

Cancer Stage	Kolmogorov-Smirnov Statistic (Calculated D_n Value)
Normal	0.03062797
IPMA	0.04904261
IOIPMN	0.04815737
IPMC	0.05123668

Table A9. Kolmogorov-Smirnov test for Pancreatic cancer.

Cancer Stage	Kolmogorov-Smirnov Statistic (Calculated D_n Value)
Normal	0.03062797
IPMA	0.04904261
IOIPMN	0.04815737
IPMC	0.05123668

Since $D_n < D_{n\alpha}$ we conclude by Kolmogorov-Smirnov that the data does not fit a log-normal distribution, but it is a good aproximation as seen by all plots and the near D values.

Since $D_n < D_{n\alpha}$ we conclude by Kolmogorov-Smirnov that the data does not fit a log-normal distribution, but it is a good aproximation as seen by all plots and the near D values.

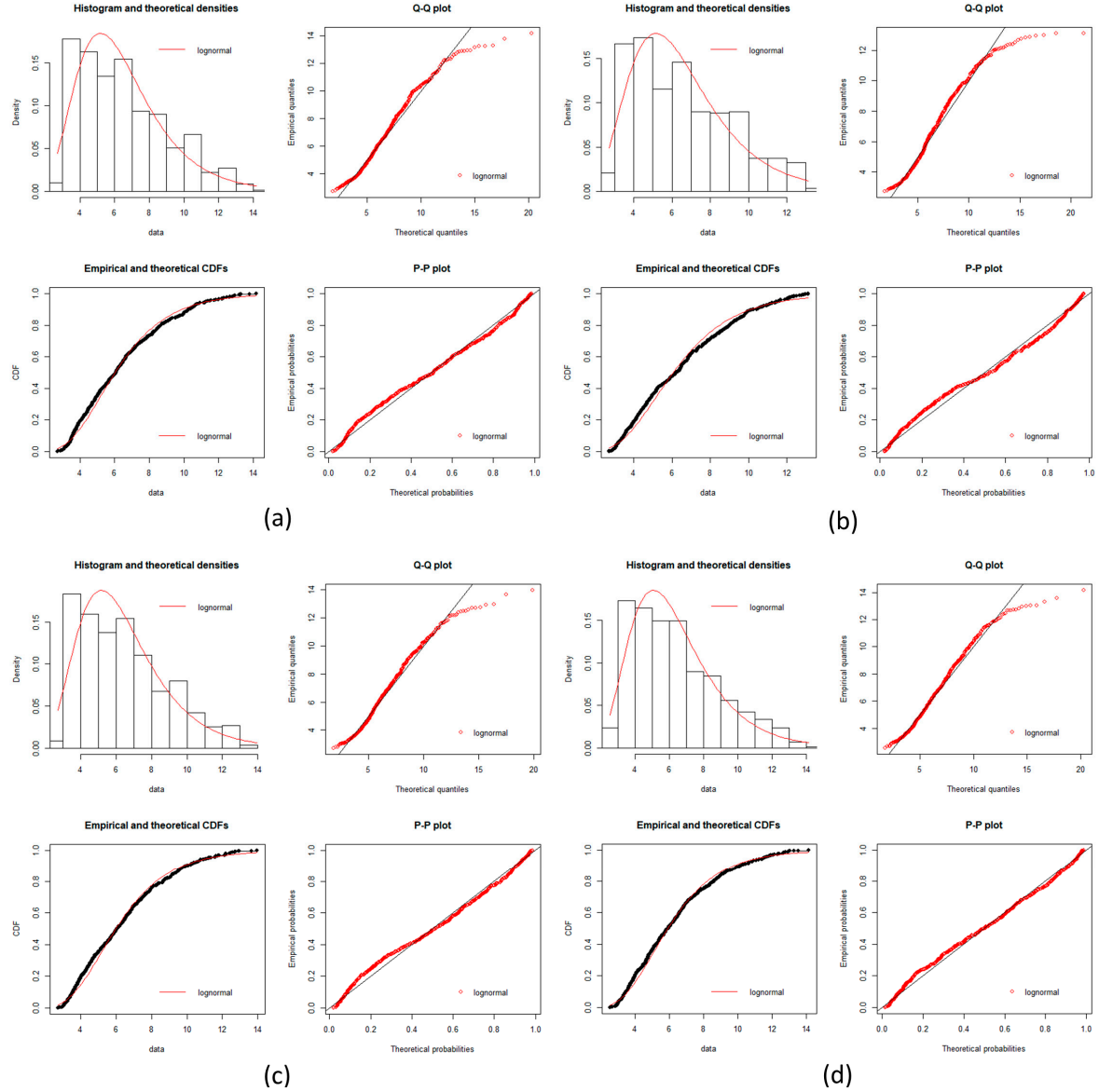

Figure A5. Hepatocellular carcinoma goodness of fit test plots part 1. Plots of goodness fits test for a log-normal distribution is showed in each incise for gene expression data. Histogram, Q-Q plot, CDFs and P-P plots are common plots to test a fitness of distribution. It can be seen that each stage fit well for a log normal distribution. (**a**) Normal stage; (**b**) LGDLT stage; (**c**) HGDLT stage; (**d**) VEHC stage.

$\alpha = 0.05$ for an $n > 35$; $D_{n,\alpha} = \frac{\sqrt{-0.5\ln\left(\frac{\alpha}{2}\right)}}{\sqrt{n}}$; $D_{n,\alpha} = \frac{1.358}{24.26} = 0.0597$.

$$\alpha = 0.05 \quad \text{for an} \quad n > 35; \quad D_{n,\alpha} = \frac{\sqrt{-0.5\ln\left(\frac{\alpha}{2}\right)}}{\sqrt{n}}; \quad D_{n,\alpha} = \frac{1.358}{24.26} = 0.0597.$$

Table A10. Kolmogorov-Smirnov test for HCC.

Cancer Stage	Kolmogorov-Smirnov Statistic (Calculated D_n Value)
Normal	0.05043655
LGDLT	0.05035987
HGDLT	0.05226466
VEHC	0.04894896
EHC	0.04728102
AHC	0.04770404
VAHC	0.04946804

Table A10. Kolmogorov-Smirnov test for HCC.

Cancer Stage	Kolmogorov-Smirnov Statistic (Calculated D_n Value)
Normal	0.05043655
LGDLT	0.05035987
HGDLT	0.05226466
VEHC	0.04894896
EHC	0.04728102
AHC	0.04770404
VAHC	0.04946804

Since $D_n < D_{n\alpha}$ we conclude that the data is a good fit with the log-normal distribution.

Since $D_n < D_{n\alpha}$ we conclude that the data is a good fit with the log-normal distribution.

In melanoma and squamous cell carcinoma of the lung, the Kolmogorov-Smirnov test was made to probe the normal distribution.

In melanoma and squamous cell carcinoma of the lung, the Kolmogorov-Smirnov test was made to probe the normal distribution.

H0: The sample follows a normal distribution

H0: The sample follows a normal distribution

Ha: The sample does not follow a normal distribution

Ha: The sample does not follow a normal distribution

If $D_n < D_{n\alpha}$ it implies that the data is a good fit with the normal distribution.

If $D_n < D_{n\alpha}$ it implies that the data is a good fit with the normal distribution.

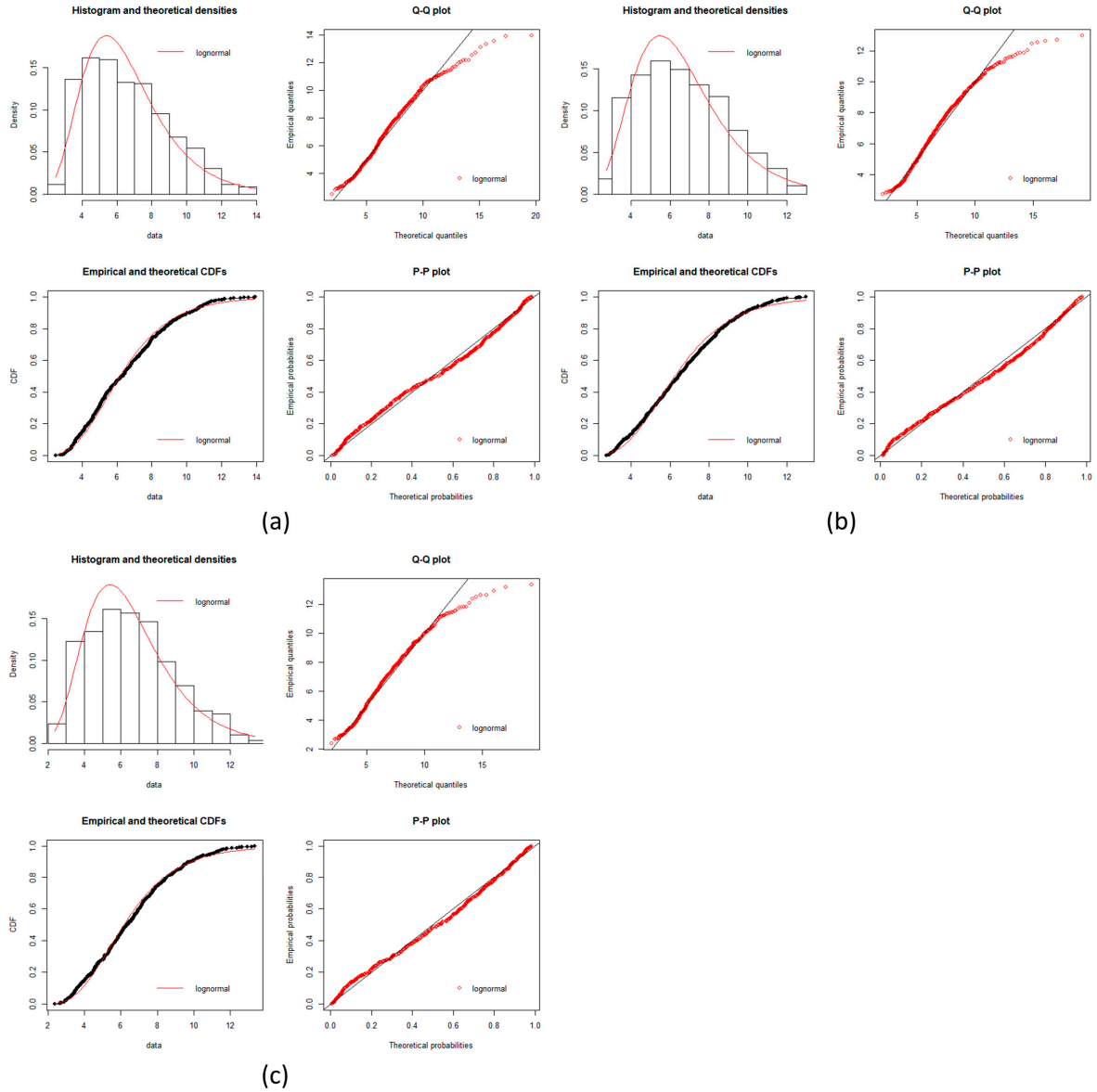

Figure A6. Hepatocellular carcinoma goodness of fit test plot part 2. Plots of goodness fits test for a log-normal distribution is showed in each incise for gene expression data. Histogram, Q-Q plot, CDFs and P-P plots are common plots to test a fitness of distribution. It can be seen that each stage fit well a log-normal distribution (a) EHCH, (b) AHC, (c) VAHC.

Figure A6. Hepatocellular carcinoma goodness of fit test plot part 2. Plots of goodness fits test for a log-normal distribution is showed in each incise for gene expression data. Histogram, Q-Q plot, CDFs and P-P plots are common plots to test a fitness of distribution. It can be seen that each stage fit well a log-normal distribution (a) EHCH, (b) AHC, (c) VAHC.

$$D_{n,\alpha} = \frac{1.358}{44.78} = 0.03032$$

$$D_{n,\alpha} = \frac{1.358}{44.78} = 0.03032$$

Table A1. Kolmogorov-Smirnov test for Melanoma.

Cancer stage	Kolmogorov-Smirnov Statistic (Calculated D_n Value)
NS	0.02327638
BN	0.03602832
AN2	0.02572476
INS	0.02299426
VGP	0.02741794
MGP	0.03271799
LN	0.03189103

Since $D_n < D_{n,\alpha}$ we conclude that the data is a good fit with the normal distribution except for BN which is a good approximation.

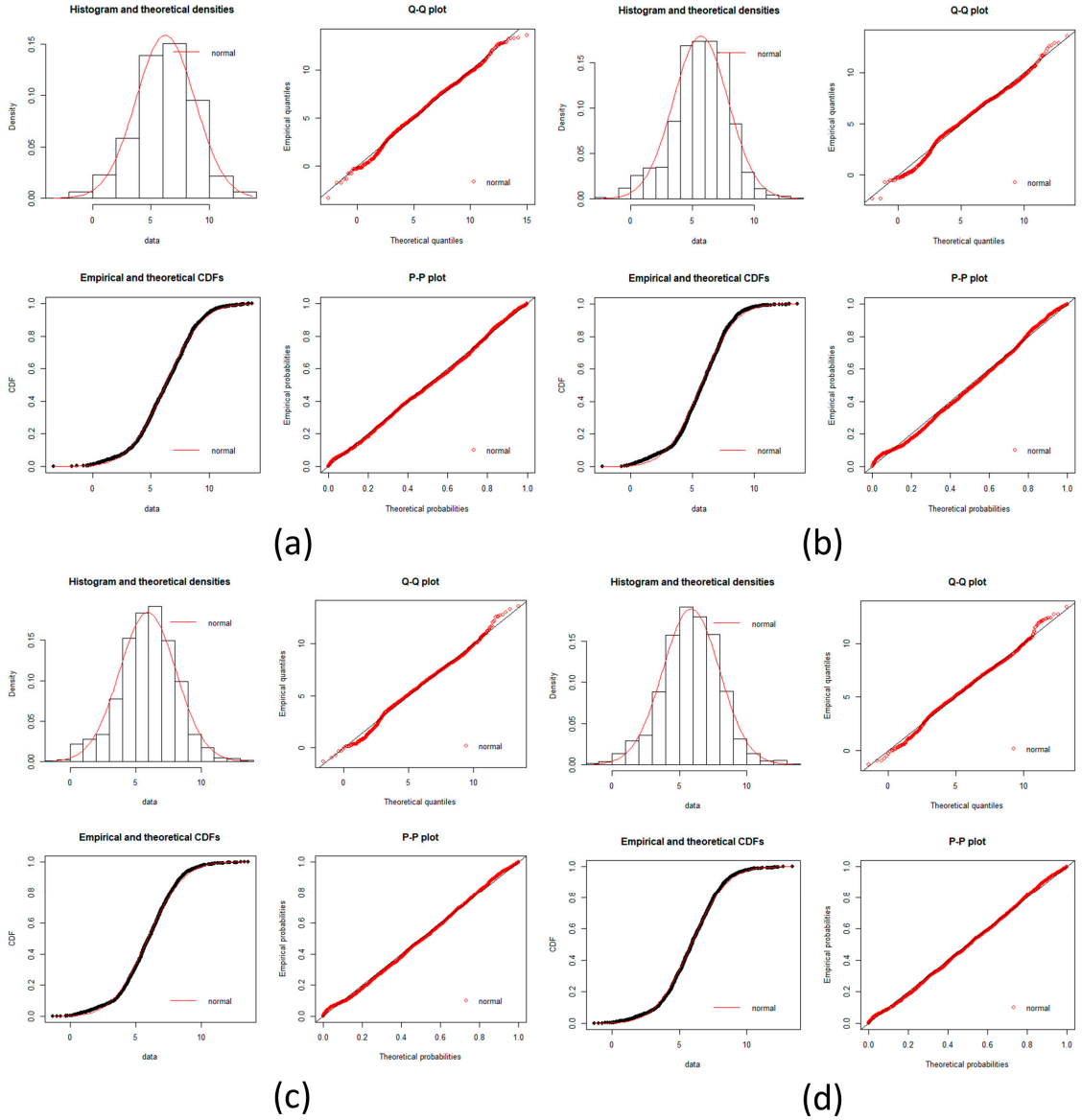

Figure A7. Melanoma goodness of fit test plot part 1. Plots of goodness fits test for a log-normal distribution is showed in each incise for gene expression data. Histogram, Q-Q plot, CDFs and P-P plots are common plots to test a fitness of distribution. It can be seen that each stage fit well for a log normal distribution. (**a**) Normal stage; (**b**) BN stage; (**c**) AN2 stage; (**d**) INS stage.

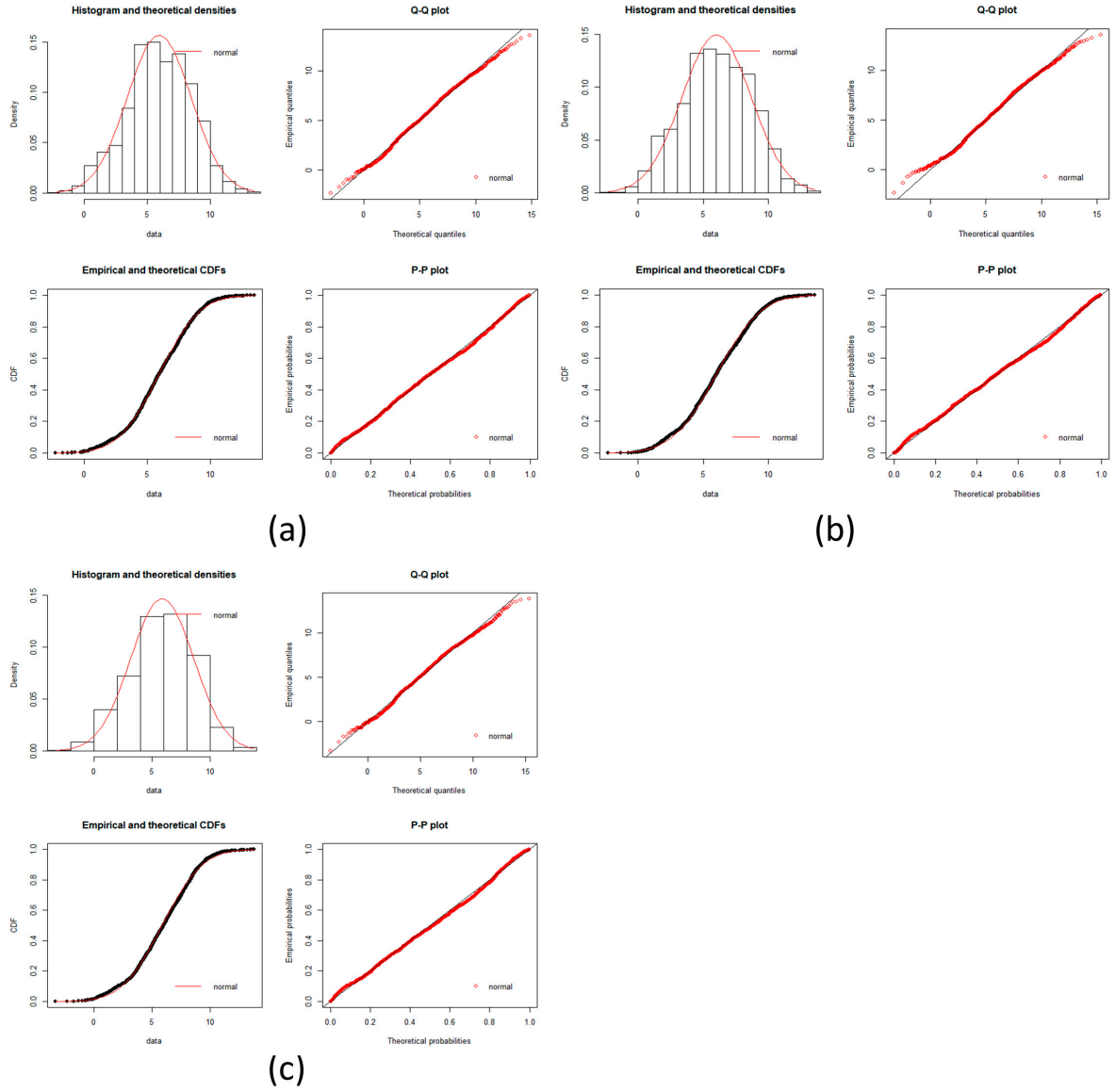

Figure A8. Melanoma goodness of fit test plot part 2. Plots of goodness fits test for a log-normal distribution is showed in each incise for gene expression data. Histogram, Q-Q plot, CDFs and P-P plots are common plots to test a fitness of distribution. It can be seen that each stage fit well for a log-normal distribution. (**a**) VGP stage; (**b**) MGP stage; (**c**) LN stage.

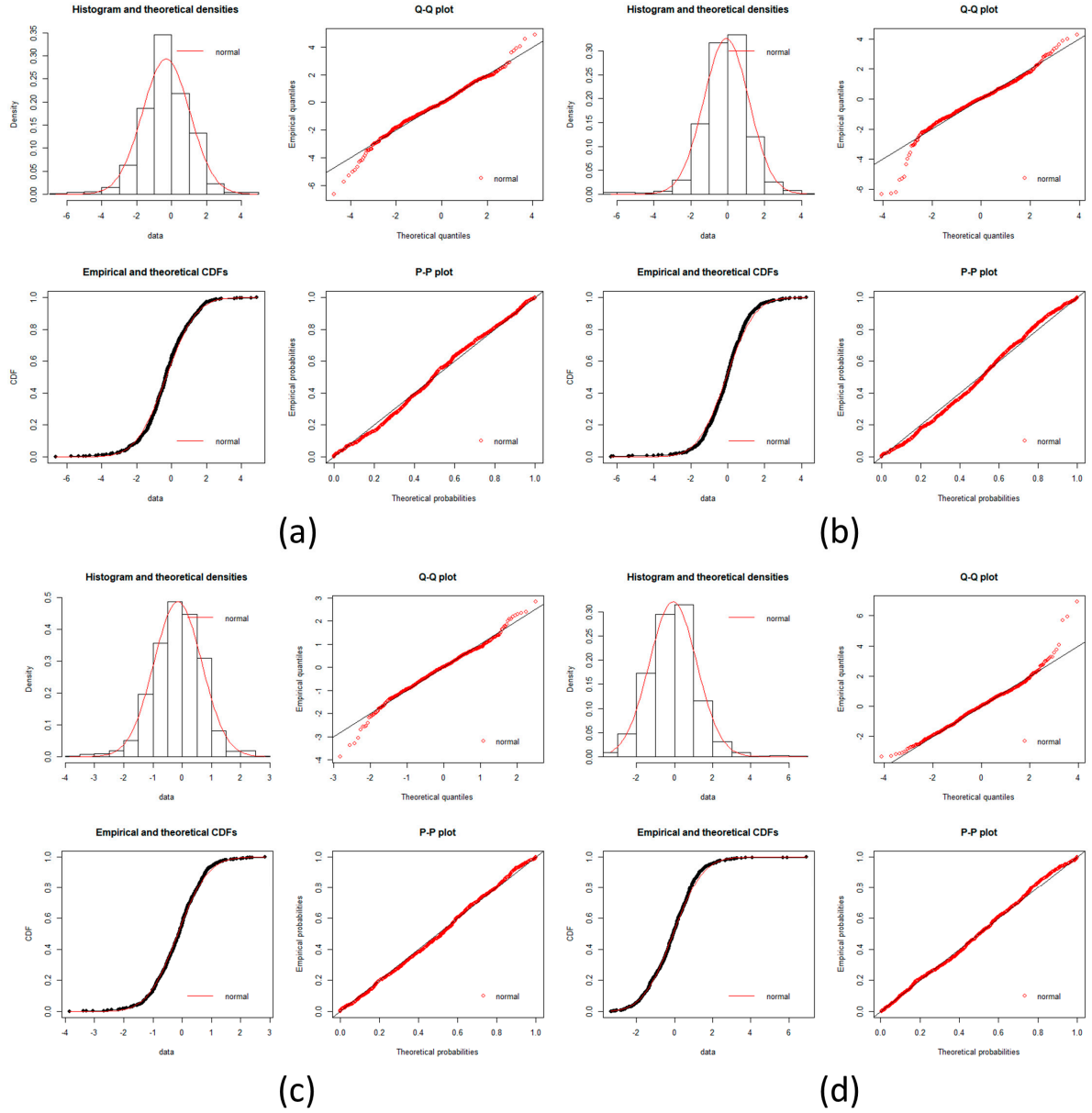

Figure A9. Squamous cell carcinoma of the lung goodness of fit test plot part 1. Plots of goodness fits test for a log-normal distribution is showed in each incise for gene expression data. Histogram, Q-Q plot, CDFs and P-P plots are common plots to test a fitness of distribution. It can be seen that each stage fit well for a log normal distribution. (**a**) Normal stage; (**b**) HYP stage; (**c**) META stage; (**d**) MILD DYS stage.

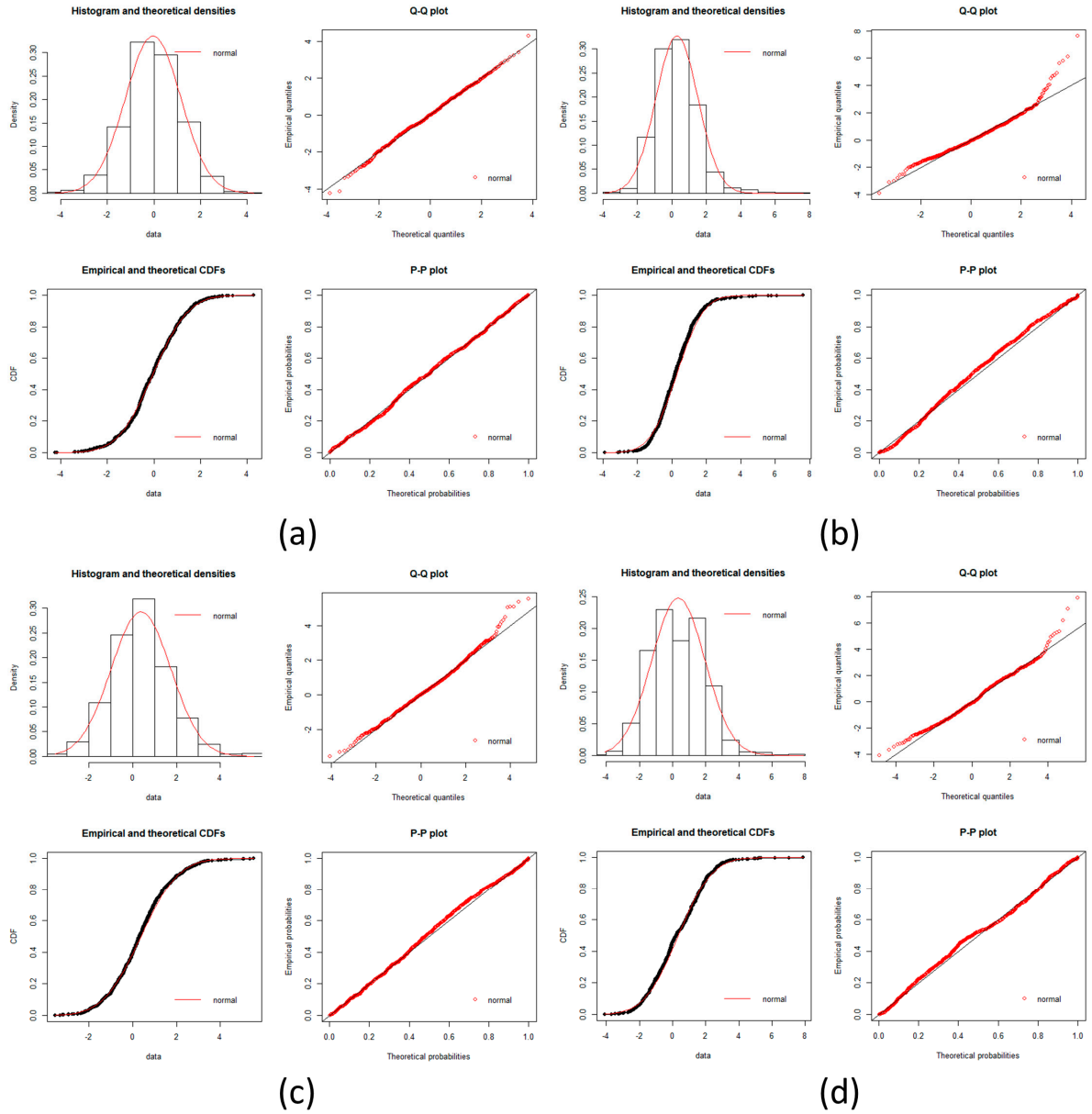

Figure A10. Squamous cell carcinoma of the lung goodness of fit test plot part 2. Plots of goodness fits test for a log-normal distribution is showed in each incise for gene expression data. Histogram, Q-Q plot, CDFs and PP plots are common plots to test a fitness of distribution. It can be seen that each stage fit well for a log normal distribution: (a) MOD DYS stage; (b) SEV DYS stage; (c) CIN STU stage; (d) SQ stage.

$$D_{n,\alpha} = \frac{1.358}{29.32} = 0.04631$$

Table A12. Kolmogorov-Smirnov test for Squamous cell carcinoma.

Cancer Stage	Kolmogorov-Smirnov Statistic (Calculated D_n Value)
NS	0.0451448
HYP	0.04257946
META	0.03065325
MILD DYS	0.03673821
MOD DYS	0.02808452
SEV DYS	0.04279729
C IN SITU	0.03382777
SQ	0.04313791

Since $D_n < D_{n\alpha}$ we conclude that the data is a good fit with the normal distribution.

Appendix B. Method to Calculate the Determinant of a Matrix, Tables about the Statistics of Local Networks for Each Cancer, and a Glossary of Terms

Appendix B.1. Calculation of the Determinant of the Covariance Matrix

A covariance matrix of X is a square $k \times k$ matrix whose generic (i,j)-th entry is equal to the covariance between x_i and x_j The diagonal entries are equal to the variances of the individual components of X (see equation below). Assuming two samples, the 2×2 covariance matrix is given by:

$$k = \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] \\ Cov[X_2, X_1] & Var[X_2] \end{bmatrix}$$

To calculate the determinant of the covariance matrix, we have:

$$|\kappa| = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

$$|\kappa| = (1 \times 1) - (0.3 \times 0.3)$$

$$|\kappa| = 0.91$$

where $||$ denotes the determinant of the matrix.

The following tables summarize some descriptors of the local networks for each group in each cancer.

Table A13. Melanoma local networks and size distribution.

Group *	Number of Local Networks	Size Distribution of Local Networks **
1	8	8:2
2	61	61:2
3	258	258:2
4	181	181:2
5	3	5:3
6	30	29:2, 1:4
7	546	201:3, 130: 4, 65:5, 41:6, 30:8, 27: 7, 13:10, 9:9, 9:11, 7:12, 2:13, 2:14, 2;15, 2:16, 2:19, 1:17, 1:20, 1:21, 1:22
8	925	106:3, 100:4, 104:5, 85:6, 73:7, 63:8, 46: 9, 51:10, 34:11, 25:12, 15:13, 18:16, 14:17, 15:18, 11: 19, 8:20, 7:21, 13:22, 12:23, 6:24, 4:25, 5: 26, 4:27, 2:28, 4:29, 4:30, 4:31, 3:32, 7:33, 4:35, 2: 36, 3:37, 4:38, 3:39, 1:40, 4:41, 2:42, 2:43, 2:44, 1:46, 1:48, 2:50, 2:52, 1:55, 1:62, 2:63, 2:35, 1:66, 1:70, 1: 73, 1:75, 1:90, 1:110, 1:118, 1:162, 1:171

* Group count start from the group with the minimum entropy values. ** First number correspond to the number of local networks; second number corresponds to the size of the local networks, e.g., Group 6 has 29 local networks with 2 nodes and 1 local network with 4 nodes.

Table A14. HCC local networks and size distribution.

Group *	Number of Local Networks	Size Distribution of Local Networks **
1	9	9:2
2	68	68:2
3	102	11:3, 91:2
4	91	91:2
5	42	31:4, 11:3
6	43	1:5, 42:4
7	47	47:5
8	31	31:6
9	47	16:7, 31:6
10	34	4:9, 16:8, 14:7
11	106	1:83, 1:63, 1:55, 1:54, 1:50, 1:43, 1:40, 1:38, 1:32, 1:30, 1:29, 1:26, 2:25, 3:23, 1:22, 2:21, 3:20, 3:19, 4:18, 7:17, 5:16, 3:15, 6:14, 4:13, 8:12, 7:11, 12:10, 8:9, 16:8

* Group counts start from the group with the minimum entropy value. ** First number corresponds to the number of local networks; second number corresponds to the size of the local networks, e.g., Group 1 has 9 local networks with 2 nodes.

Table A15. Pancreatic local networks and size distribution.

Group *	Number of Local Networks	Size Distribution of Local Networks **
1	87	87:2
2	341	341:2
3	26	1:3, 25:2
4	17	17:3
5	267	267:3
6	10	10:2
7	22	2:5, 20:4
8	1170	203:4, 150:5, 138:6, 104:7, 82:8, 72:9, 63:10, 45:11, 38:12, 33:13, 32:14, 19:15, 17:19, 13:20, 12:17, 9:26, 9:25, 9:21, 9:18, 8:29, 8:24, 8:23, 6:22, 5:33, 4:30, 4:28, 4:27, 3:51, 3:47, 3:43, 3:36, 3:32, 2:79, 2:62, 2:55, 2:53, 2:49, 2:41, 2:37, 2:35, 2:34, 1:168, 1:128, 1:89, 1:83, 1:78, 1:77, 1:69, 1:66, 1: 63, 1:61, 1:54, 1:42, 1:39, 1:38, 1:31
9	5	1:40, 1:19, 1:10, 1:7, 1:4

* Group counts start from the group with the minimum entropy value. ** First number corresponds to the number of local networks; second number corresponds to the size of the local networks, e.g., Group 3 has 1 local network with 3 nodes and 25 local networks with 2 nodes.

Table A16. Squamous cell carcinoma local networks and size distribution.

Group *	Number of Local Networks	Size Distribution of Local Networks **
1	16	16:2
2	117	2:3, 115:2
3	174	16:3, 158:2
4	143	126:3, 17:2
5	77	36:4, 41:3
6	67	1:5, 66:4
7	49	48:5, 1:4
8	33	24:6, 9:5
9	33	2:7, 26:6
10	32	31:7, 1:6
11	29	11: 8, 2:7
12	13	1: 9, 12:8
13	11	11:9
14	11	11:10
15	7	7:11
16	4	1:11, 3:12
17	3	3:12
18	28	1: 24, 2:23, 1:21, 1:18, 5:17, 1:16, 4:15, 6:14, 7:13
19	12	1: 57, 1:56, 1:37, 1:36, 1:35, 1:33, 1:29, 2:24, 1:22, 1:18, 1:16, 1:14

* Group count start from the group with the minimum entropy values. ** First number corresponds to the number of local networks. Second number corresponds to size of the local networks, e.g., Group 12 has 1 local network with 9 nodes and 12 local networks with 8 nodes.

Appendix B.2. Glossary

Sample—It is the gene expression levels coming from one individual.
Network—All the nodes from the APID PPI data that represents the selected genes from the differential expression analysis for each stage in each type of cancer.
Subnetwork—A group of local networks binned together based on similarity of Multivariate Normal entropy in the normal group.
Local network—It is one node of the network plus its immediate neighbors based on the APID PPI data.

References

1. WHO. Cancer. Available online: http://www.who.int/mediacentre/factsheets/fs297/en/ (accessed on 26 October 2017).
2. What Is Cancer? Available online: https://www.cancer.gov/about-cancer/understanding/what-is-cancer (accessed on 26 October 2017).
3. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006; ISBN 978-0-471-24195-9.
4. Chen, B.; Wang, J.; Zhao, H.; Principe, J. Insights into Entropy as a Measure of Multivariate Variability. *Entropy* **2016**, *18*, 196. [CrossRef]
5. Torre, L.A.; Bray, F.; Siegel, R.L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A. Global cancer statistics, 2012: Global Cancer Statistics, 2012. CA. *Cancer J. Clin.* **2015**, *65*, 87–108. [CrossRef] [PubMed]
6. Waller, L.P.; Deshpande, V.; Pyrsopoulos, N. Hepatocellular carcinoma: A comprehensive review. *World J. Hepatol.* **2015**, *7*, 2648–2663. [CrossRef] [PubMed]
7. Raza, A. Hepatocellular carcinoma review: Current treatment, and evidence-based medicine. *World J. Gastroenterol.* **2014**, *20*, 4115. [CrossRef] [PubMed]
8. Balogh, J.; Victor, D.; Asham, E.H.; Burroughs, S.G.; Boktour, M.; Saharia, A.; Li, X.; Ghobrial, M.; Monsour, H. Hepatocellular carcinoma: A review. *J. Hepatocell. Carcinoma* **2016**, *3*, 41–53. [CrossRef] [PubMed]
9. Liu, R.; Li, M.; Liu, Z.-P.; Wu, J.; Chen, L.; Aihara, K. Identifying critical transitions and their leading biomolecular networks in complex diseases. *Sci. Rep.* **2012**, *2*, 813. [CrossRef] [PubMed]
10. Yabar, C.S.; Winter, J.M. Pancreatic Cancer. *Gastroenterol. Clin. N. Am.* **2016**, *45*, 429–445. [CrossRef] [PubMed]
11. Makohon-Moore, A.; Iacobuzio-Donahue, C.A. Pancreatic cancer biology and genetics from an evolutionary perspective. *Nat. Rev. Cancer* **2016**, *16*, 553–565. [CrossRef] [PubMed]
12. Kleeff, J.; Korc, M.; Apte, M.; La Vecchia, C.; Johnson, C.D.; Biankin, A.V.; Neale, R.E.; Tempero, M.; Tuveson, D.A.; Hruban, R.H.; et al. Pancreatic cancer. *Nat. Rev. Dis. Primer* **2016**, *2*, 16022. [CrossRef] [PubMed]
13. Heist, R.S.; Sequist, L.V.; Engelman, J.A. Genetic Changes in Squamous Cell Lung Cancer: A Review. *J. Thorac. Oncol.* **2012**, *7*, 924–933. [CrossRef] [PubMed]
14. Derman, B.A.; Mileham, K.F.; Bonomi, P.D.; Batus, M.; Fidler, M.J. Treatment of advanced squamous cell carcinoma of the lung: A review. *Transl. Lung Cancer Res.* **2015**, *4*, 524–532. [CrossRef] [PubMed]
15. Drilon, A.; Rekhtman, N.; Ladanyi, M.; Paik, P. Squamous-cell carcinomas of the lung: Emerging biology, controversies, and the promise of targeted therapy. *Lancet Oncol.* **2012**, *13*, e418–e426. [CrossRef]
16. Goodwin, J.; Neugent, M.L.; Lee, S.Y.; Choe, J.H.; Choi, H.; Jenkins, D.M.R.; Ruthenborg, R.J.; Robinson, M.W.; Jeong, J.Y.; Wake, M.; et al. The distinct metabolic phenotype of lung squamous cell carcinoma defines selective vulnerability to glycolytic inhibition. *Nat. Commun.* **2017**, *8*, 15503. [CrossRef] [PubMed]
17. Gandara, D.R.; Hammerman, P.S.; Sos, M.L.; Lara, P.N.; Hirsch, F.R. Squamous Cell Lung Cancer: From Tumor Genomics to Cancer Therapeutics. *Clin. Cancer Res.* **2015**, *21*, 2236–2243. [CrossRef] [PubMed]
18. Miller, A.J.; Mihm, M.C. Melanoma. *N. Engl. J. Med.* **2006**, *355*, 51–65. [CrossRef] [PubMed]
19. Schadendorf, D.; Fisher, D.E.; Garbe, C.; Gershenwald, J.E.; Grob, J.-J.; Halpern, A.; Herlyn, M.; Marchetti, M.A.; McArthur, G.; Ribas, A.; et al. Melanoma. *Nat. Rev. Dis. Primer* **2015**, 15003. [CrossRef] [PubMed]
20. Shain, A.H.; Bastian, B.C. From melanocytes to melanomas. *Nat. Rev. Cancer* **2016**, *16*, 345–358. [CrossRef] [PubMed]
21. Teschendorff, A.E.; Sollich, P.; Kuehn, R. Signalling entropy: A novel network-theoretical framework for systems analysis and interpretation of functional omic data. *Methods* **2014**, *67*, 282–293. [CrossRef] [PubMed]

22. West, J.; Bianconi, G.; Severini, S.; Teschendorff, A.E. Differential network entropy reveals cancer system hallmarks. *Sci. Rep.* **2012**, *2*. [CrossRef] [PubMed]
23. Brehme, M.; Koschmieder, S.; Montazeri, M.; Copland, M.; Oehler, V.G.; Radich, J.P.; Brümmendorf, T.H.; Schuppert, A. Combined Population Dynamics and Entropy Modelling Supports Patient Stratification in Chronic Myeloid Leukemia. *Sci. Rep.* **2016**, *6*. [CrossRef] [PubMed]
24. Park, Y.; Lim, S.; Nam, J.-W.; Kim, S. Measuring intratumor heterogeneity by network entropy using RNA-seq data. *Sci. Rep.* **2016**, *6*. [CrossRef] [PubMed]
25. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
26. Khalil, D.N.; Smith, E.L.; Brentjens, R.J.; Wolchok, J.D. The future of cancer treatment: Immunomodulation, CARs and combination immunotherapy. *Nat. Rev. Clin. Oncol.* **2016**, *13*, 273–290. [CrossRef] [PubMed]
27. Sridharan, K.; Gogtay, N.J. Therapeutic nucleic acids: Current clinical status: Therapeutic nucleic acids. *Br. J. Clin. Pharmacol.* **2016**, *82*, 659–672. [CrossRef] [PubMed]
28. Wasmer, M.-H.; Krebs, P. The Role of IL-33-Dependent Inflammation in the Tumor Microenvironment. *Front. Immunol.* **2017**, *7*. [CrossRef] [PubMed]
29. Wurmbach, E.; Chen, Y.; Khitrov, G.; Zhang, W.; Roayaie, S.; Schwartz, M.; Fiel, I.; Thung, S.; Mazzaferro, V.; Bruix, J.; et al. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. *Hepatology* **2007**, *45*, 938–947. [CrossRef] [PubMed]
30. Smith, A.P.; Hoek, K.; Becker, D. Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanomas. *Cancer Biol. Ther.* **2005**, *4*, 1018–1029. [CrossRef] [PubMed]
31. Hiraoka, N.; Yamazaki–Itoh, R.; Ino, Y.; Mizuguchi, Y.; Yamada, T.; Hirohashi, S.; Kanai, Y. CXCL17 and ICAM2 Are Associated With a Potential Anti-Tumor Immune Response in Early Intraepithelial Stages of Human Pancreatic Carcinogenesis. *Gastroenterology* **2011**, *140*, 310–321.e4. [CrossRef] [PubMed]
32. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]
33. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300.
34. Wright, S.P. Adjusted P-Values for Simultaneous Inference. *Biometrics* **1992**, *48*, 1005. [CrossRef]
35. R & Bioconductor—Manuals. Available online: http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual (accessed on 29 October 2017).
36. Alonso-López, D.; Gutiérrez, M.A.; Lopes, K.P.; Prieto, C.; Santamaría, R.; De Las Rivas, J. APID interactomes: Providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res.* **2016**, *44*, W529–W535. [CrossRef] [PubMed]
37. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]
38. Shannon, P.T.; Grimes, M.; Kutlu, B.; Bot, J.J.; Galas, D.J. RCytoscape: Tools for exploratory network analysis. *BMC Bioinform.* **2013**, *14*, 217. [CrossRef] [PubMed]
39. Delignette-Muller, M.L.; Dutang, C. fitdistrplus: An *R* Package for Fitting Distributions. *J. Stat. Softw.* **2015**, *64*. [CrossRef]
40. Ahmed, N.A.; Gokhale, D.V. Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. Inf. Theory* **1989**, *35*, 688–692. [CrossRef]
41. UniProt. Available online: http://www.uniprot.org/ (accessed on 29 October 2017).
42. Martínez, C.; García-Martín, E.; Ladero, J.M.; Sastre, J.; Garcia-Gamito, F.; Diaz-Rubio, M.; Agúndez, J.A. Association of CYP2C9 genotypes leading to high enzyme activity and colorectal cancer risk. *Carcinogenesis* **2001**, *22*, 1323–1326. [CrossRef] [PubMed]
43. Jernström, H.; Bågeman, E.; Rose, C.; Jönsson, P.-E.; Ingvar, C. CYP2C8 and CYP2C9 polymorphisms in relation to tumour characteristics and early breast cancer related events among 652 breast cancer patients. *Br. J. Cancer* **2009**, *101*, 1817–1823. [CrossRef] [PubMed]
44. Kiyohara, M.H.; Dillard, C.; Tsui, J.; Kim, S.R.; Lu, J.; Sachdev, D.; Goodglick, L.; Tong, M.; Torous, V.F.; Aryasomayajula, C.; et al. EMP2 is a novel therapeutic target for endometrial cancer stem cells. *Oncogene* **2017**, *36*, 5793–5807. [CrossRef] [PubMed]

45. Guo, X.; Yang, M.; Gu, H.; Zhao, J.; Zou, L. Decreased expression of SOX6 confers a poor prognosis in hepatocellular carcinoma. *Cancer Epidemiol.* **2013**, *37*, 732–736. [CrossRef] [PubMed]
46. Ye, T.; Xu, J.; Du, L.; Mo, W.; Liang, Y.; Xia, J. Downregulation of UBAP2L inhibits the epithelial-mesenchymal transition via SNAIL1 regulation in hepatocellular carcinoma cells. *Cell. Physiol. Biochem.* **2017**, *41*, 1584–1595. [CrossRef] [PubMed]
47. Sturla, L.-M.; Tong, M.; Hebda, N.; Gao, J.; Thomas, J.-M.; Olsen, M.; de la Monte, S.M. Aspartate-β-hydroxylase (ASPH): A potential therapeutic target in human malignant gliomas. *Heliyon* **2016**, *2*, e00203. [CrossRef] [PubMed]
48. Fiori, V.; Magnani, M.; Cianfriglia, M. The expression and modulation of CEACAM1 and tumor cell transformation. *Ann. DellIstituto Super. Sanità* **2012**, *48*, 161–171. [CrossRef]
49. Yang, C.; He, P.; Liu, Y.; He, Y.; Yang, C.; Du, Y.; Zhou, M.; Wang, W.; Zhang, G.; Wu, M.; Gao, F. Assay of serum CEACAM1 as a potential biomarker for breast cancer. *Clin. Chim. Acta* **2015**, *450*, 277–281. [CrossRef] [PubMed]
50. Kato, T.; Wada, H.; Patel, P.; Hu, H.; Lee, D.; Ujiie, H.; Hirohashi, K.; Nakajima, T.; Sato, M.; Kaji, M.; et al. Overexpression of KIF23 predicts clinical outcome in primary lung cancer patients. *Lung Cancer* **2016**, *92*, 53–61. [CrossRef] [PubMed]
51. Luo, C.; Yao, Y.; Yu, Z.; Zhou, H.; Guo, L.; Zhang, J.; Cao, H.; Zhang, G.; Li, Y.; Jiao, Z. UBE2T knockdown inhibits gastric cancer progression. *Oncotarget* **2017**. [CrossRef] [PubMed]
52. Gregory, T.R. Synergy between sequence and size in Large-scale genomics. *Nat. Rev. Genet.* **2005**, *6*, 699–708. [CrossRef] [PubMed]
53. Lemon, J.; Bolker, B.; Oom, S.; Klein, E.; Rowlingson, B.; Wickham, H.; Tyagi, A.; Eterradossi, O.; Grothendieck, G.; Toews, M.; et al. plotrix: Various Plotting Functions. 2017. Available online: https://cran.r-project.org/web/packages/plotrix/plotrix.pdf (accessed on 24 February 2017).
54. Wickham, H.; Francois, R.; Henry, L.; Müller, K. RStudio dplyr: A Grammar of Data Manipulation. 2017. Available online: https://cran.r-project.org/web/packages/dplyr/index.html (accessed on 24 February 2017).
55. Tibshirani, R.; Leisch, F. Bootstrap: Functions for the Book "An Introduction to the Bootstrap". 2017. Available online: https://cran.r-project.org/web/packages/bootstrap/bootstrap.pdf (accessed on 24 February 2017).

CAPÍTULO 2

Juarez-Flores, A., & José, M. (2019). Original Article Squamous cell carcinoma of the lung: Gene expression and network analysis during carcinogenesis. International Journal of Clinical and Experimental Medicine, 12, 6671–6683.

Int J Clin Exp Med 2019;12(6):6671-6683
www.ijcem.com /ISSN:1940-5901/IJCEM0088518

Original Article

Squamous cell carcinoma of the lung: gene expression and network analysis during carcinogenesis

Angel Juarez-Flores[1,2], Marco V José[2]

[1]*Posgrado en Ciencias Biológicas, Unidad de Posgrado, Circuito de Posgrados, Ciudad Universitaria, Universidad Nacional Autónoma de México, CP 04510, Mexico City, Mexico;* [2]*Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, CP 04510, Mexico City, Mexico*

Received October 31, 2019; Accepted March 12, 2019; Epub June 15, 2019; Published June 30, 2019

Abstract: Lung cancer is one of the most common and deadliest types of cancer. Most often, diagnosis is made in the later stages of the disease, with few treatment options available. Squamous cell carcinoma of the lung (SCCL) is one of the most common types of lung cancer. Knowledge concerning its carcinogenic process lags behind that of other cancers of the lungs. Aiming to understand the biological phenomena underlying each stage of the disease and unveil the most significant genes, the current study carried out bioinformatic analysis of different samples that corresponded to the carcinogenic process. New relevant genes for early diagnosis and treatment are proposed and expression profiles for each stage are presented. Based on Protein-Protein interaction networks of these genes, this study proposes that they function as gatekeepers for a wide variety of processes. MYC, MCM2, AURKA, CUL3, and DDIT4L are proposed as a possible group for treatment of SCCL. This work provides a general panorama of the transcriptome profile of SCCL, with a plethora of information regarding its carcinogenesis. Results obtained by this *in silico* approach constitute a guide for further experimental works necessary for corroborating and validating their potential application for diagnosis and treatment of the disease.

Keywords: Squamous lung cancer, carcinogenesis, microarray, carcinoma, transcriptome

Introduction

In 2018, statistical epidemiological estimates showed that lung cancer is one of the most frequent and deadliest types of cancer, worldwide, even in developed countries. For both sexes, statistics showed 2,093,876 new cases and 1,761,007 deaths [1, 2]. Lung cancer is commonly divided into two types, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC comprises approximately 85% of all lung cancers [3, 4]. NSCLC adenocarcinoma and squamous cell carcinoma (SCC) of the lungs are the most prevalent [5, 6]. However, cancer therapies mostly focus on treatment of adenocarcinoma. Incidence of adenocarcinoma has increased in recent years [5]. In most patients, cancer diagnosis is made in late stages when metastasis is often present and available treatments are not as effective [7]. Surgery is the best treatment option, but only in the earlier stages of the disease. Available therapies for advanced stages are limited. There has been little improvement observed in patient survival. Standard treatment consists of platinum-based doublet therapy [7, 8]. Squamous cell carcinoma of the lungs is mainly associated with cigarette smoking [5]. Its carcinogenic process is described by different stages. The first step is hyperplasia, followed by metaplasia, varying degrees of dysplasia, and carcinoma *in situ*, finally reaching squamous cell carcinoma [6, 9]. Histology is the main criterion in classifying stages of carcinogenesis. Some chromosomal losses appear in the process [6]. A better understanding of pre-disease and early events that occur during the carcinogenic process will assist in improving early diagnosis and achieving better outcomes. The current study consists of transcriptome profiling, with enrichment and network analysis as an aid, aiming to better understand the carcinogenic process of SCC. This study highlights genes with the most statistically significant expression. Underlying

biological processes for each stage are described with the help of gene ontology. This study also used a Protein-Protein interaction (PPI) network to highlight some proteins with a high number of connections.

Materials and methods

Data retrieval

A search was conducted in the National Center for Biotechnology Information Gene Expression Omnibus database, retrieving the following datasets: Squamous cell carcinoma, GEO accession: GSE33479, with 122 samples which represent carcinogenic stages. The samples were divided, including 13 with normal histology and normofluorescence and 14 with normal histology and hypofluorescence. They were selected as the control group, including 15 for metaplasia, 13 for mild dysplasia, 13 for moderate dysplasia, 12 for severe dysplasia, 13 for carcinoma *in situ*, and 14 for squamous cell carcinoma of the lungs. The platform was: Agilent-014850 Whole Human Genome, Microarray 4x44K G4112F. All analyses were performed using *R* v3.4.3 software (http://www.R-project.org).

Differential gene expression analysis

First, this study obtained preprocessed data from the SOFT file, according to the manual of the *GEOquery* R package and *hgug4112a.db* R package was used to annotate each gene ID to the data [10]. Differential gene expression (DGE) analysis was performed, making a comparison of each stage against the control using the *limma* package. It fits a generalized linear model before comparisons and calculates a moderate t-statistic for each contrast [11, 12]. A *p*-value was obtained. It was adjusted based on the Benjamini and Hochberg False Discovery Rate correction using *limma* [11, 13]. Volcano plots were constructed for each cancer stage, in which upper left and upper right points in the plot were selected genes that fulfilled a double criterion: False Discovery Rate (FDR) adjusted p-value < 0.05 and a Fold Change value > 1.5 or < -1.5.

Gene ontology enrichment analysis

Gene Ontology (GO) enrichment analysis was performed with the topGO package (http://bioconductor.org/packages/topGO/) as follows: As the gene universe, all of the genes contained in the Agilent chip and *hgug4112a.db* library were used for annotation. Relevant genes were the ones showing an FDR adjusted p-value < 0.05. Statistical testing for *topGO* classic Fisher, classic Kolmogorov-Smirnov (KS), and a variation of KS (elimKS) available in the *topGO* program, was conducted. Classic Fisher is based on gene counts. It is related to hypergeometric probability, rendering a p-value of the overlap between two independent sets. In the topGO version, it uses modular enrichment analysis (MEA), which considers inter-relationships of GO terms. The list of relevant genes was input [14, 15]. Classic K-S and elimKS are based on gene scores, providing Gene set enrichment analysis which reduces arbitrary factors and uses all information of the microarray. They use scores (p-values) obtained from DGE analysis as input. Moreover, elimKS is a more conservative KS variation [14, 15]. More enrichment analysis was made using *PANTHER* v13.1 (PANTHER Enrichment Test Released 201704-13). Input data were the list of selected genes by the volcano plots (double criteria). Gene Universe was the database default for *Homo sapiens*, using the complete GO annotation dataset provided by the software (GO Ontology Database Released 2018-09-06) [16]. Selected analysis for *PANTHER* was statistical overrepresentation enrichment testing. The test type was Fisher's exact test, with FDR multiple test correction. This study compared the results of *topGO* and *PANTHER*, finding common *GO* enriched terms in each stage, except for hyperplasia. No statistical differences were found with *limma* and the normal stage (used as control), according to differential gene expression analysis.

Network construction and analysis

For this study, researchers downloaded the full interactome of *Homo sapiens* from Mentha database [17]. Retrieval, construction, management, and network analysis of the PPI were made with Cytoscape software. Retrieved interactions were curated by the database, which only integrates data from experimentally determined direct protein interactions [18]. For network construction, duplicated edges (lines that bind nodes), nodes (ellipses that represent a protein), and self-loops were deleted to create an interactome network, functioning as a basic template. Network Analyzer was used to calculate the number of connections for each node

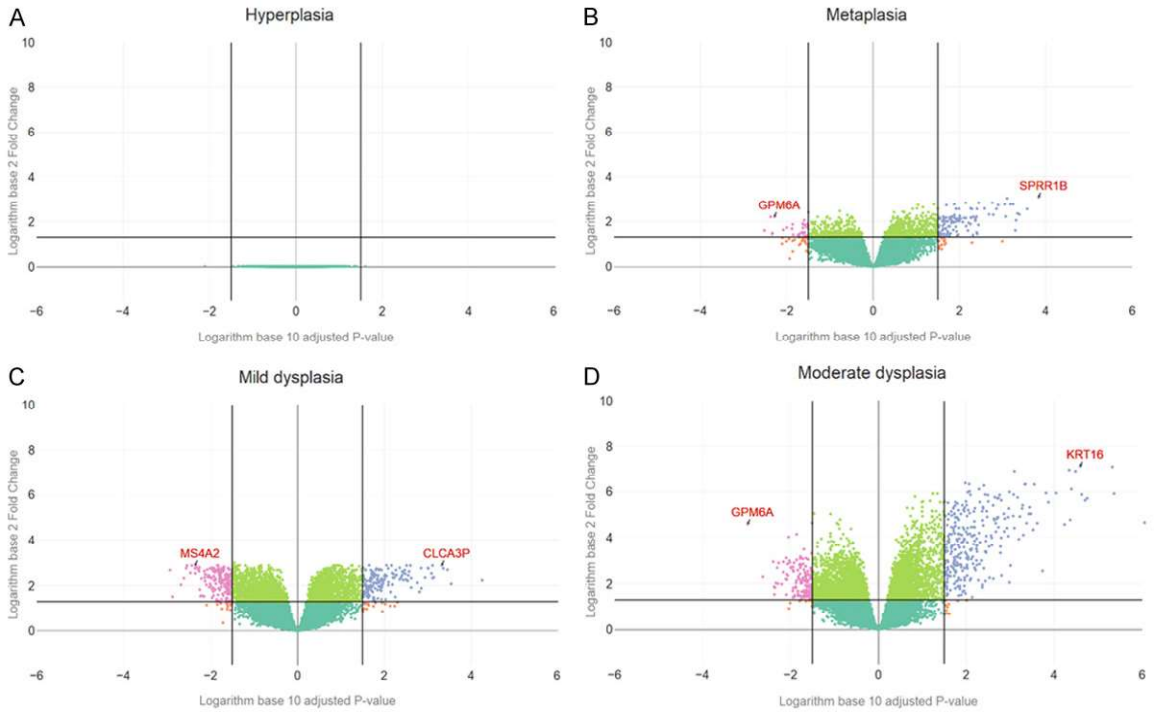

Figure 1. Volcano plots. (A) Hyperplasia; (B) Metaplasia; (C) Mild dysplasia; (D) Moderate dysplasia. Two genes with FC > 1.5 and the minor *p*-value are selected were tagged with their respective names in each cancer stage. Vertical lines divided the genes by FC criteria > 1.5 to the right (over expressed) or < -1.5 to the left (down expressed). The black horizontal line is the *p*-value cutoff: above the line < 0.5, in the line = 0.05, down the line > 0.05.

[19]. Networks with the first neighbors of some of the highlighted genes in the Volcano plots were retrieved from the basic template network, creating a graphical network output for each node. The size represents the number of connections they had in the original template. Next, generated lists of all the genes selected by the double filter criteria in all stages were merged. Duplicated genes were deleted, creating a single list with 2,199 genes. This list was used to find matches with the created interactome network, with 1,793 genes identified. Nodes that matched the previous list were selected to create a new network. Some nodes had no connections. Thus, it was processed to eliminate these nodes, creating a unique network in which every node was connected. This study obtained a final network with 1,152 nodes. For every node in the final network, Network Analyzer was used to calculate the degree, number of connections of each node, and betweenness. This is a measure based on the shortest paths (the short path from a point a to a point b), indicating that it is a measure of highest number of times that a node is passed by the shortest paths [20].

Results

First, differential gene expression analysis was conducted, using double criteria calculated from the volcano plots for every stage of the carcinogenic process. **Figures 1A-D** and **2A**, **2B** show the so-called pre-disease stages. In **Figure 2C**, the cancer stages are shown. As stages become more advanced, the number of significant genes increases (both by fold change and statistical *p*-value (see **Figure 2D** for a better understanding), indicating greater alterations in expression profiles patterns, according to the advancement of the disease. This could indicate a necessity of the cells for new processes to be upregulated or downregulated, depending on the requirements for its development and potential outcomes of cancer stage and metastasis. An example is the case of metalloprotease proteins, which were detected in the moderate dysplasia stage (See Table S1 in Supplementary Information at the end of the paper), upregulated (data not shown) with the highest number of them in the carcinoma *in situ* and squamous cell carcinoma stages. Expression profiles for each stage were found

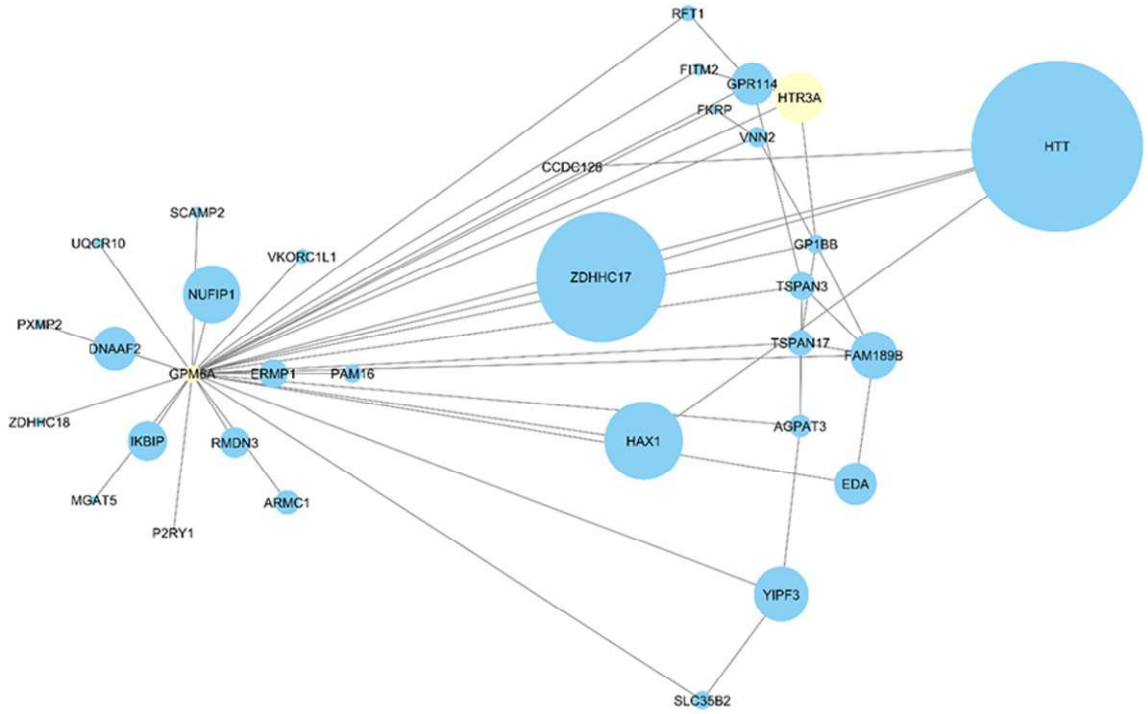

Figure 2. GPM6A network. This graph shows the first neighbors nodes that are connected to GPM6A. The size of the nodes is related to the number of connections they had in the base network. Nodes in yellow color are the ones that are in the list created using the double filter criteria. HTT is the most connected node that interact directly with GPM6A.

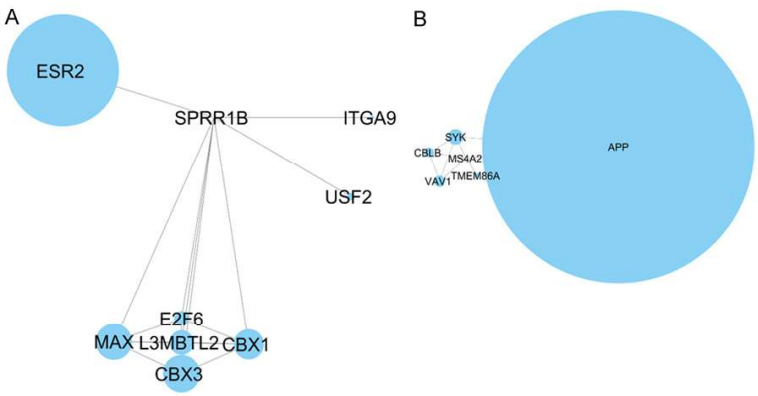

Figure 3. SPRR1B and MS4A2 networks. A: Shows the first neighbors nodes that are connected to SPRR1B; ESR2 is the most connected node. B: Shows the first neighbors nodes that are connected to MS4A2; APP is the most connected node. The size of the nodes is related to the number of connections they had in the base network.

and presented in Table S1, which corresponds to the genes with a differential expression. For each stage, two genes are pointed out using an arrow with their respective names, except for the hyperplasia stage, which shows no statistical significance.

In **Figure 1**, selected genes for each stage (except for hyperplasia) are: a) Metaplasia: GPM6A-This is a protein involved with neural differentiation. It may be involved in regulation of endocytosis and intracellular trafficking of G-protein-coupled receptors. Thus far, it has not been associated with cancer [21]; SPRR1B-This is an envelope protein with transglutaminase cross-linking properties of keratinocytes. It can function as an amine donor and acceptor in transglutaminase-mediated cross-linkage. It has been proposed as a biomarker for squamous metaplasia and oral squamous cell cancer stem-like cells. It has also been suggested that its upregulation suppresses RASSF4, a tumor suppressor [22-25]; b) Mild dysplasia: MS4A2-This is a high affinity receptor that binds to the Fc region of immunoglobulins

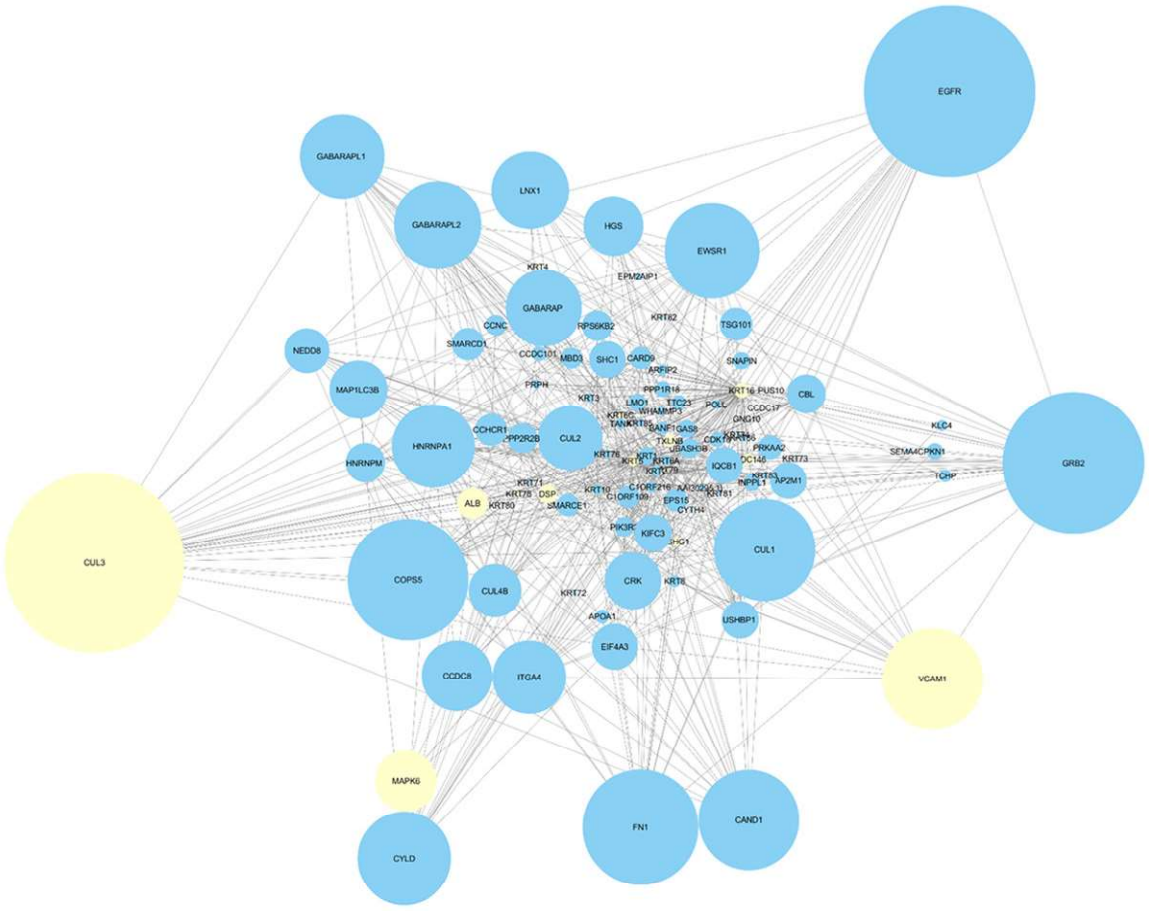

Figure 4. KRT16 network. Shows the first neighbors nodes that are connected to KRT16. The size of the nodes is related to the number of connections they had in the base network. Nodes in yellow color are the ones that are in the list created using the double filter criteria. CUL3 and EFG3 are the most connected nodes that interact directly with KRT16.

epsilon. It can mediate the secretion of some lymphokines. In adenocarcinoma, it has been suggested that its expression is an independent prognostic marker for patient survival [26-28]; CLCA3P-This is a pseudogene. When it is cloned and expressed, it produces a protein that is secreted into the culture supernatant, but no function has yet been ascribed [29]; c) Moderate dysplasia: GPM6A, KRT16-This is a keratin protein that could act as a regulator of innate immunity in response to skin barrier breach. In oral squamous cell carcinoma, it is overexpressed. In metastatic breast cancer, it is associated with a shorter relapse-free survival [30-32]. Although **Figures 3-5** shows that GPM6A, SPRR1B, and MS4A2 had few interactions, proteins that were connected to them had plenty of interactions. Thus, they seem to function as "gatekeepers" for a wide variety of processes. KRT16 had more direct interactions than others. It also had interactions with highly connected proteins, as seen in **Figure 4**.

In **Figure 2**, selected genes included: a) Severe dysplasia: HSD17B13-This is a member of the hydroxysteroid 17-beta family. Its function is unknown. Overexpression in mice and cultured hepatocyte lines increases lipogenesis. In hepatocellular carcinoma, its downregulation has been associated with worse survival in patients [33, 34]. KRT6B-This is a member of the keratin gene family and a type II cytokeratin. It has been suggested that its upregulation might contribute to renal cell carcinoma progression [35, 36]; b) Carcinoma *in situ*: ASXL3-This is a putative polycomb group protein. It acts by forming multiprotein complexes to maintain a repressive state in lung induced-pluripotent stem cells. Its silencing has inhibited proliferation and diminished malignant

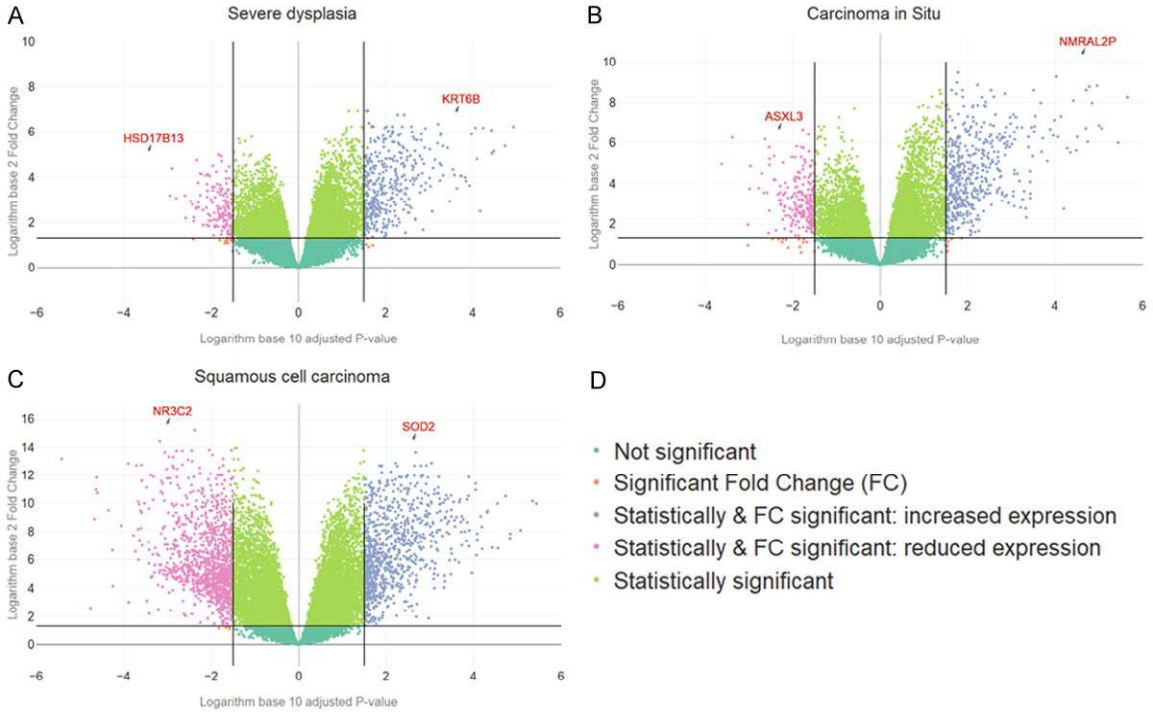

Figure 5. Volcano plots (continuation). (A) Severe dysplasia; (B) Carcinoma *in situ*; (C) Squamous cell carcinoma of the lungs; (D) Color codes. The two genes with FC > 1.5 and the minor *p*-value are selected and tagged with their respective names in each cancer stage. Vertical lines divided the genes by FC criteria > 1.5 to the right (over expressed) or < -1.5 to the left (down expressed). The black horizontal line is the *p*-value cutoff: above the line < 0.5, in the line = 0.05, down the line > 0.05.

growth in small cell lung cancer cells. Thus, it was proposed to be a novel candidate target for therapy [37, 38]; NMRAL2P-This is a pseudogene. It has been identified as a direct transcriptional target of Nrf2 in colon cancer cell lines [39]; c) Squamous cell carcinoma of the lungs: NR3C2-This is a mineralocorticoid receptor. It binds to mineralocorticoid response elements and transactivates target genes. Reduced levels of expression have correlated with poorer survival in pancreatic ductal adenocarcinoma [40, 41]; SOD2-This is a superoxide dismutase which destroys superoxide anion radicals. It was observed to have tumor suppressive and promoting functions [42, 43]. **Figures 6-8** show connections at the first neighbor level for highlighted proteins (**Figure 5**). Except for HSD17B13 (**Figure 6**), the proteins had interactions with highly connected proteins. KRT6B and KRT16 had the most direct interactions.

To gain further insight concerning the underlying processes, this study conducted enrichment analysis with different methods, using two different software platforms. Results were compared to found matches, as summarized in **Table 1**. **Table 1** only shows the results that matched between topGO and PANTHER. In **Table 2**, some cancer stages for GO molecular function and cellular component did not have a match between topGO and PANTHER. Thus, only topGO principal results are used in **Tables 2** and **3** (see caption). In GO biological process (BP) (**Table 1**), cell division was found to be important in four stages between metaplasia to carcinoma *in situ* (except mild dysplasia). Cornification was also found in the four stages, which could mean that this process starts early in the carcinogenic process. DNA replication initiation was found in two stages (moderate and SCC) and sister chromatid cohesion was found in two stages (moderate and severe dysplasia). **Table 2** shows GO molecular function (MF). Protein binding was found in the four stages from moderate dysplasia to squamous cell carcinoma. **Table 3** shows GO cellular components (CC). Cytosol and condensed chromosome outer kinetochore were found in all stages, except for mild dysplasia. Extracellular exosome was encountered in mild dysplasia and squamous cell carcinoma.

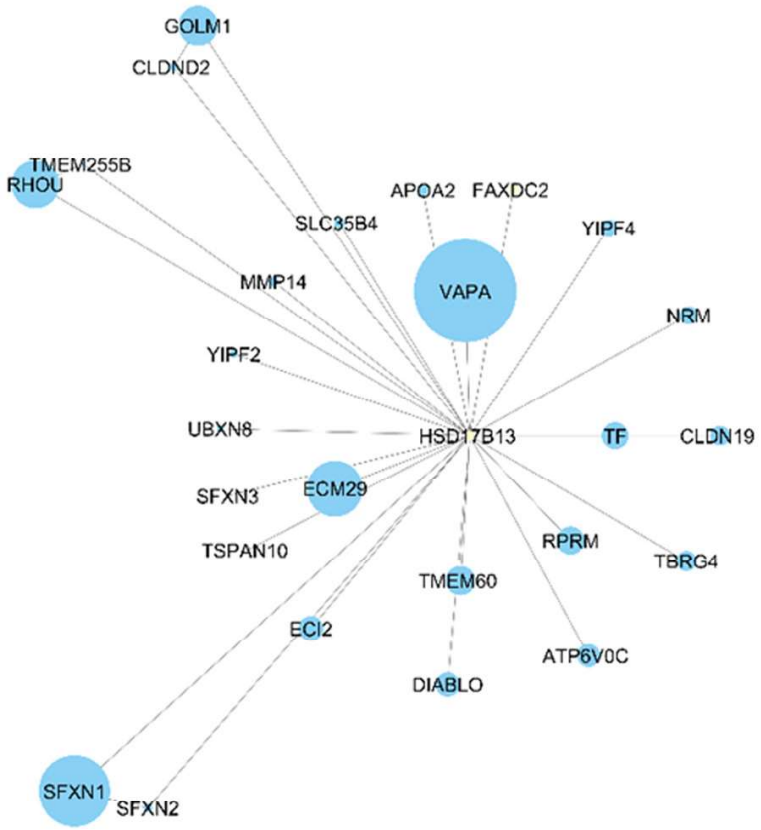

Figure 6. HSD17B13 network. The first neighbors nodes that are connected to HSD17B13 are shown. The size of the nodes is related to the number of connections they had in the base network. VAPA is the most connected node that interacts directly with HSD17B13.

A network was created. Components were the protein products of genes that were selected by the double filter criteria. This study calculated degrees (number of connection) and betweenness using the Network Analyzer. In **Table 4**, the top five nodes for each measure, in descending order, are shown. They may play important roles in SCCL development: MYC (found upregulated in this work) is a transcription factor that promotes angiogenesis trough VEGFA. Its upregulation could promote esophageal squamous cell carcinoma [44, 45]; MCM2 (upregulated) and AURKA (upregulated) participate in cell cycle regulation [46, 47]. MCM2 deregulation was found to be involved in lung cancer cell proliferation [48]; AURKA has been found to be overexpressed in oral squamous cell carcinoma cells. Its suppression inhibits cancer growth [49]; CUL3 (downregulated) participates in ubiquitination and proteasomal degradation. When silent in breast cancer, it makes cell lines more resistant to treatments, such as doxorubicin and paclitaxel [50, 51]; DDIT4L (downregulated) acts as an inhibitor of mTOR signaling. It was found that its downregulation promotes cutaneous SCCL proliferation [52-54]. These proteins may also function as an axe of regulation, due to high connectivity and the short paths that pass through them. However, the genes previously highlighted are not in the list of top nodes, but they are biologically relevant.

Discussion

During analysis, a wide variety of genes with a significant change of expression were detected. This study points out some of them, denoting the possible importance of their expression during the carcinogenesis process of squamous cell carcinoma of the lungs. Most of the highlighted genes have been reported in other types of cancer [22, 26, 30, 31, 35, 39, 43]. Only CLCA3P, which is a pseudogene, had not been reported in cancer. Its function remains unknown [29]. Occurring with other types of cancers, genes SPRR1B, KRT16, KRT6B, and NMRAL2P were found via the double filter criteria to be significant in the carcinogenesis process (except hyperplasia). Upregulation of these genes has been associated with cancer promotion or patient poor survival [22, 30, 31, 35, 39]. GP6MA was found to be significant in the carcinogenesis process. It is downregulated, but it has not yet been associated with cancer. Since GP6MA is related with G-protein-coupled receptors, it could play an important role in SCC promotion. The latter, together with the above-mentioned genes, may play a role as early biomarkers for SCC. Expression patterns of the highlighted genes matched those observed in other types of cancers. Therefore, they should participate as cancer promoter-genes beginning at precancerous stages, making them potentially useful as therapeutics targets. In this list of highlighted genes, there were

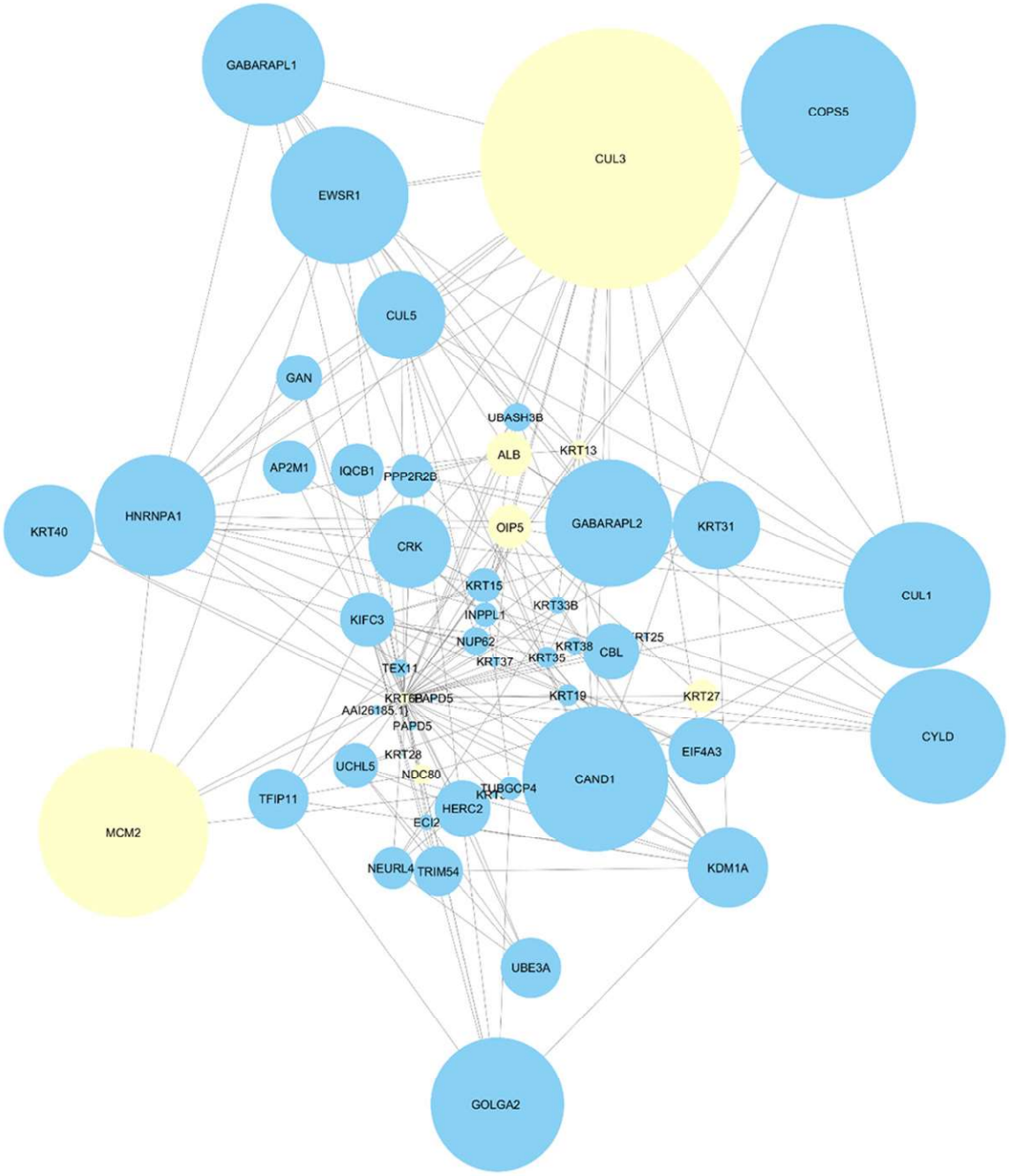

Figure 7. KRT6B network. The first neighbors nodes that are connected to KRT6B are displayed. The size of the nodes is related to the number of connections they had in the base network. Nodes in yellow color are the ones that are in the list created using the double filter criteria. CUL3 is the most connected node that interact directly with KRT6B.

others, like HSD17B13 and CLCA3P, with expression levels found to be significant since the mild dysplasia stage, including SOD2 since the carcinoma *in situ* stage, ASXL3 since the severe dysplasia stage, and NR3C2, only in the cancer stage. Only MS4A2 was found to be significant both in mild dysplasia and moderate dysplasia stages. Some upregulated genes were identified as part of the keratin family. This is a family of structural proteins. In preinvasive lesions, this study found two types of keratins that were upregulated (**Figures 1**, **2**).

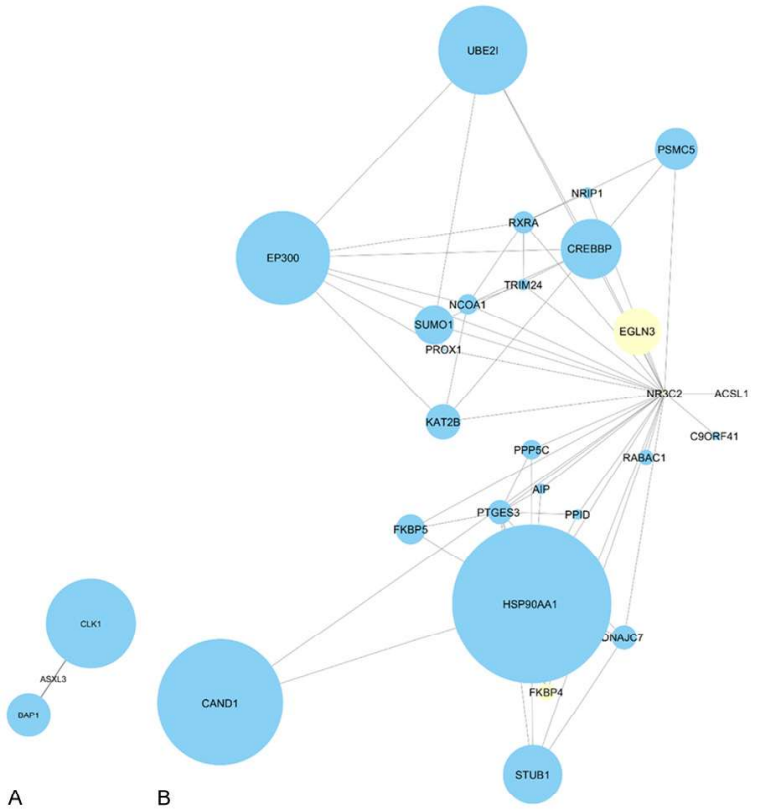

Figure 8. ASXL3 and NR3C2 network. A: Shows the first neighbors nodes that are connected to ASXL3. The size of the nodes is related to the number of connections they had in the base network. CLK1 is the most connected node that interact directly with ASXL3; B: The first neighbors nodes that are connected to NR3C2 are shown. The size of the nodes is related to the number of connections they had in the base network. Nodes in yellow color are the ones that are in the list created using the double filter criteria. HSP90AA1 is the most connected node that interact directly with NR3C2.

As mentioned before, these proteins do not have much interaction, compared to other proteins (**Figures 3**-**8**). However, since they are connected to highly connected ones, they are likely to function as gatekeepers in a wide variety of signaling process. Networks are useful in supporting and visualizing results of volcano plots. These plots assist in observing details of how the highlighted genes are connected. Centrality measures, such as degrees and betweenness of the network, are useful in finding some proteins that had significant importance. However, this analysis only provides a glimpse of the full panorama of what is happening biologically. The full panorama is obtained with network analysis, combined with volcano plots. According to GO analysis, it was found that the cornification process, which is a type of cell death, was significant in some stages [9, 55]. Also, a portion of the genes related to the cornification process was found to have a significant change in expression values, beginning at the earliest moment of the carcinogenesis process (data not shown). This suggests that they could be necessary in reaching the cancer stage. Subsequent development may be in a way, similar to that which occurs with KRT6B and KRT16 in renal cell carcinoma and breast cancer, respectively [30, 35]. Another process that was pointed out by enrichment analysis was cell division. This was expected due to the constant cell division in cancer cells. A very important portion in the number of genes was found to be significant. However, there were none in the ones selected for volcano plots. MYC, MCM2, AURKA, CUL3, and DDIT4L were found by DEG analysis. GO enrichment analysis showed that the processes in which they participate are relevant in various stages. Network analysis also found them as important participant proteins in the networks of the DEG. Thus, it is proposed that they should be studied as a group of proteins for possible combined multi-target treatment of SCCL. Some processes and genes that are upregulated or downregulated were detected, beginning with the earliest moments of the carcinogenic process. They continued to the last stages of the process, suggesting importance not only in cancer *per se* but also in pre-disease stages. This could also be the case for KRT16. KRT16 expression was different and statistically significant, beginning at metaplasia all the way to SCCL. Further experimentation is necessary to validate the genes proposed in this study, determining whether they are also useful in finding unknown functions of the carcinogenic process.

Conclusion

The current study identified genes that may herald the development of squamous cell carcinoma of the lungs. These genes were shown

Table 1. GO Biological process similarities between topGO and PANTHER for each stage of squamous lung cancer

Biological Process					
Metaplasia	Mild dysplasia	Moderate dysplasia	Severe dysplasia	Carcinoma in situ	Squamous cell carcinoma
Cell division (GO: 0051301)	Cornification (GO: 0070268)	Cell division (GO: 0051301)	Sister chromatid cohesion (GO: 0007062)	Anaphase-promoting complex-dependent catabolic process (GO: 0031145)	Neutrophil degranulation (GO: 0043312)
Cornification (GO: 0070268)		Cornification (GO: 0070268)	Cell division (GO: 0051301)	Cell division (GO: 0051301)	DNA replication initiation (GO: 0006270)
		Sister chromatid cohesion (GO: 0007062)	Cornification (GO: 0070268)		Anaphase-promoting complex-dependent catabolic process (GO: 0031145)
		DNA replication initiation (GO: 0006270)	Anaphase-promoting complex-dependent catabolic process (GO: 0031145)		

Table 2. GO Molecular function similarities between topGO and PANTHER for each stage of squamous lung cancer

Molecular function*					
Alditol: NADP+ 1-oxidoreductase activity (GO: 0004032)	Extracellular matrix structural constituent (GO: 0005201)	Structural constituent of muscle (GO: 0008307)	Phosphatidic acid transporter activity (GO: 1990050)	Protein binding (GO: 0005515)	Protein binding (GO: 0005515)
	Platelet-derived growth factor binding (GO: 0048407)	Protein binding (GO: 0005515)	ATP binding (GO: 0005524)		RNA binding (GO: 0003723)
		Cadherin binding (GO: 0045296)	Protein binding (GO: 0005515)		Dynein light chain binding (GO: 0045503)
					Microtubule motor activity (GO: 0003777)

*In molecular function, there were no matches between topGO and PANTHER but they were found in Metaplasia, Moderate dysplasia, and Severe dysplasia.

Table 3. GO Cellular component similarities between topGO and PANTHER for each stage of squamous lung cancer

Cellular component*					
Cytoplasm (GO: 0005737)	Extracellular exosome (GO: 0070062)	Cytosol (GO: 0005829)	Cytosol (GO: 0005829)	Cytosol (GO: 0005829)	Extracellular exosome (GO: 0070062)
Condensed chromosome outer kinetochore (GO: 0000940)		Condensed chromosome outer kinetochore (GO: 0000940)	Condensed chromosome outer kinetochore (GO: 0000940)	Condensed chromosome outer kinetochore (GO: 0000940)	Condensed chromosome outer kinetochore (GO: 0000940)
Cytosol (GO: 0005829)		Centrosome (GO: 0005813)	Condensed chromosome kinetochore (GO: 0000777)	Cornified envelope (GO: 0001533)	Spindle microtubule (GO: 0005876)
					Nucleoplasm (GO: 0005654)
					Tertiary granule membrane (GO: 0070821)

*In Cellular components, no matches found between topGO and PANTHER with Metaplasia.

Table 4. Top nodes in the final PPI network

Top nodes for Betweenness	Top nodes for Degree*
MYC	MYC
DDIT4L	DDIT4L
MCM2	MCM2
CUL3	CUL3
AURKA	CDK1

*Most of the top nodes were shared between the degree and betweenness measures.

to be involved in keratinization and cornification processes. This study also identified individual genes, previously reported in many types of cancer, that are relevant in the processes of growth, metastasis, and cell division. Network analysis results suggest that these genes could function as gatekeepers for a wide variety of signaling processes. Likewise, some processes were expected to be relevant, including cell division or DNA replication initiation. This work provides a general panorama of the transcriptome profile of squamous cell carcinoma of the lungs, contributing a wealth of information concerning its carcinogenesis. Results obtained from this *in silico* approach constitute a guide for further experimental works. Future research is necessary, corroborating the potential application of these in diagnosis and treatment of the disease.

Acknowledgements

Angel Juarez-Flores is a doctoral student from Programa de Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México (UNAM), and a fellowship recipient from Consejo Nacional de Ciencia y Tecnología (CONACYT) (number: 775924). This paper constitutes a partial fulfilment of the Graduate Program in Biological Sciences of the UNAM. MVJ was financially supported by DGAPA PAPIIT-IN-224015, UNAM, México. We would like to thank Martha Cariño Aguilar for bibliographical support.

Disclosure of conflict of interest

None.

Address correspondence to: Marco V José, Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, CP 04510, Mexico City, Mexico. E-mail: marcojose@biomedicas.unam.mx

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394-424.

[2] Siegel RL, Miller KD and Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018; 68: 7-30.

[3] Gridelli C, Rossi A, Carbone DP, Guarize J, Karachaliou N, Mok T, Petrella F, Spaggiari L and Rosell R. Non-small-cell lung cancer. Nat Rev Dis Primers 2015; 1: 15009.

[4] Inamura K. Lung cancer: understanding its molecular pathology and the 2015 WHO classification. Front Oncol 2017; 7: 193.

[5] Gandara DR, Hammerman PS, Sos ML, Lara PN Jr and Hirsch FR. Squamous cell lung cancer: from tumor genomics to cancer therapeutics. Clin Cancer Res 2015; 21: 2236-2243.

[6] Drilon A, Rekhtman N, Ladanyi M and Paik P. Squamous-cell carcinomas of the lung: emerging biology, controversies, and the promise of targeted therapy. Lancet Oncol 2012; 13: e418-426.

[7] Herbst RS, Morgensztern D and Boshoff C. The biology and management of non-small cell lung cancer. Nature 2018; 553: 446-454.

[8] Hirsch FR, Scagliotti GV, Mulshine JL, Kwon R, Curran WJ Jr, Wu YL and Paz-Ares L. Lung cancer: current therapies and new targeted treatments. Lancet 2017; 389: 299-311.

[9] Travis WD, World Health Organization. International agency for research on cancer, international academy of pathology. And international association for the study of lung cancer. Pathology and genetics of tumours of the lung, pleura, thymus and heart. Lyon: IARC Press, 2004.

[10] Davis S and Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and bioconductor. Bioinformatics 2007; 23: 1846-1847.

[11] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.

[12] Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M and Barton GJ. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 2016; 22: 839-851.

[13] Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the

Royal Statistical Society. Series B (Methodological) 1995; 57: 289-300.
[14] Glass K and Girvan M. Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets. Sci Rep 2014; 4: 4191.
[15] Ackermann M and Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinformatics 2009; 10: 47.
[16] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A and Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003; 13: 2129-2141.
[17] Calderone A, Castagnoli L and Cesareni G. Mentha: a resource for browsing integrated protein-interaction networks. Nat Methods 2013; 10: 690-691.
[18] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13: 2498-2504.
[19] Assenov Y, Ramirez F, Schelhorn SE, Lengauer T and Albrecht M. Computing topological parameters of biological networks. Bioinformatics 2008; 24: 282-284.
[20] Dehmer M. Structural analysis of complex networks. Dordrecht; New York: Birkhäuser; 2011.
[21] Michibata H, Okuno T, Konishi N, Kyono K, Wakimoto K, Aoki K, Kondo Y, Takata K, Kitamura Y and Taniguchi T. Human GPM6A is associated with differentiation and neuronal migration of neurons derived from human embryonic stem cells. Stem Cells Dev 2009; 18: 629-639.
[22] Li S, Nikulina K, DeVoss J, Wu AJ, Strauss EC, Anderson MS and McNamara NA. Small proline-rich protein 1B (SPRR1B) is a biomarker for squamous metaplasia in dry eye disease. Invest Ophthalmol Vis Sci 2008; 49: 34-41.
[23] Michifuri Y, Hirohashi Y, Torigoe T, Miyazaki A, Fujino J, Tamura Y, Tsukahara T, Kanaseki T, Kobayashi J, Sasaki T, Takahashi A, Nakamori K, Yamaguchi A, Hiratsuka H and Sato N. Small proline-rich protein-1B is overexpressed in human oral squamous cell cancer stem-like cells and is related to their growth through activation of MAP kinase signal. Biochem Biophys Res Commun 2013; 439: 96-102.
[24] An G, Tesfaigzi J, Chuu YJ and Wu R. Isolation and characterization of the human spr1 gene and its regulation of expression by phorbol ester and cyclic AMP. J Biol Chem 1993; 268: 10977-10982.
[25] Candi E, Tarcsa E, Idler WW, Kartasova T, Marekov LN and Steinert PM. Transglutaminase cross-linking properties of the small proline-rich 1 family of cornified cell envelope proteins. Integration with loricrin. J Biol Chem 1999; 274: 7226-7237.
[26] Ly D, Zhu CQ, Cabanero M, Tsao MS and Zhang L. Role for high-affinity IgE receptor in prognosis of lung adenocarcinoma patients. Cancer Immunol Res 2017; 5: 821-829.
[27] Ra C, Nunomura S and Okayama Y. Fine-tuning of mast cell activation by fcepsilonribeta chain. Front Immunol 2012; 3: 112.
[28] Kuster H, Zhang L, Brini AT, MacGlashan DW and Kinet JP. The gene and cDNA for the human high affinity immunoglobulin E receptor beta chain and expression of the complete human receptor. J Biol Chem 1992; 267: 12782-12787.
[29] Gruber AD and Pauli BU. Molecular cloning and biochemical characterization of a truncated, secreted member of the human family of Ca2+-activated Cl- channels. Biochim Biophys Acta 1999; 1444: 418-423.
[30] Joosse SA, Hannemann J, Spotter J, Bauche A, Andreas A, Muller V and Pantel K. Changes in keratin expression during metastatic progression of breast cancer: impact on the detection of circulating tumor cells. Clin Cancer Res 2012; 18: 993-1003.
[31] Khanom R, Nguyen CT, Kayamori K, Zhao X, Morita K, Miki Y, Katsube K, Yamaguchi A and Sakamoto K. Keratin 17 Is induced in oral cancer and facilitates tumor growth. PLoS One 2016; 11: e0161163.
[32] Lessard JC, Pina-Paz S, Rotty JD, Hickerson RP, Kaspar RL, Balmain A and Coulombe PA. Keratin 16 regulates innate immunity in response to epidermal barrier breach. Proc Natl Acad Sci U S A 2013; 110: 19537-19542.
[33] Su W, Wang Y, Jia X, Wu W, Li L, Tian X, Li S, Wang C, Xu H, Cao J, Han Q, Xu S, Chen Y, Zhong Y, Zhang X, Liu P, Gustafsson JA and Guan Y. Comparative proteomic study reveals 17beta-HSD13 as a pathogenic protein in nonalcoholic fatty liver disease. Proc Natl Acad Sci U S A 2014; 111: 11437-11442.
[34] Chen J, Zhuo JY, Yang F, Liu ZK, Zhou L, Xie HY, Xu X and Zheng SS. 17-beta-hydroxysteroid dehydrogenase 13 inhibits the progression and recurrence of hepatocellular carcinoma. Hepatobiliary Pancreat Dis Int 2018; 17: 220-226.
[35] Hu J, Zhang LC, Song X, Lu JR and Jin Z. KRT6 interacting with notch1 contributes to progression of renal cell carcinoma, and aliskiren inhibits renal carcinoma cell lines proliferation in vitro. Int J Clin Exp Pathol 2015; 8: 9182-9188.
[36] Takahashi K, Paladini RD and Coulombe PA. Cloning and characterization of multiple hu-

man genes and cDNAs encoding highly related type II keratin 6 isoforms. J Biol Chem 1995; 270: 18581-18592.
[37] Katoh M and Katoh M. Identification and characterization of ASXL3 gene in silico. Int J Oncol 2004; 24: 1617-1622.
[38] Shukla V, Rao M, Zhang H, Beers J, Wangsa D, Wangsa D, Buishand FO, Wang Y, Yu Z, Stevenson HS, Reardon ES, McLoughlin KC, Kaufman AS, Payabyab EC, Hong JA, Zhang M, Davis S, Edelman D, Chen G, Miettinen MM, Restifo NP, Ried T, Meltzer PA and Schrump DS. ASXL3 is a novel pluripotency factor in human respiratory epithelial cells and a potential therapeutic target in small cell lung cancer. Cancer Res 2017; 77: 6267-6281.
[39] Johnson GS, Li J, Beaver LM, Dashwood WM, Sun D, Rajendran P, Williams DE, Ho E and Dashwood RH. A functional pseudogene, NMRAL2P, is regulated by Nrf2 and serves as a coactivator of NQO1 in sulforaphane-treated colon cancer cells. Mol Nutr Food Res 2017; 61.
[40] Yang S, He P, Wang J, Schetter A, Tang W, Funamizu N, Yanaga K, Uwagawa T, Satoskar AR, Gaedcke J, Bernhardt M, Ghadimi BM, Gaida MM, Bergmann F, Werner J, Ried T, Hanna N, Alexander HR and Hussain SP. A novel mif signaling pathway drives the malignant character of pancreatic cancer by targeting NR3C2. Cancer Res 2016; 76: 3838-3850.
[41] Arriza JL, Weinberger C, Cerelli G, Glaser TM, Handelin BL, Housman DE and Evans RM. Cloning of human mineralocorticoid receptor complementary DNA: structural and functional kinship with the glucocorticoid receptor. Science 1987; 237: 268-275.
[42] MacMillan-Crow LA and Thompson JA. Tyrosine modifications and inactivation of active site manganese superoxide dismutase mutant (Y34F) by peroxynitrite. Arch Biochem Biophys 1999; 366: 82-88.
[43] Kim YS, Gupta Vallur P, Phaeton R, Mythreye K and Hempel N. Insights into the dichotomous regulation of SOD2 in cancer. Antioxidants (Basel) 2017; 6.
[44] Shi Y, Xu X, Zhang Q, Fu G, Mo Z, Wang GS, Kishi S and Yang XL. tRNA synthetase counteracts c-Myc to develop functional vasculature. Elife 2014; 3: e02349.
[45] Lian Y, Niu X, Cai H, Yang X, Ma H, Ma S, Zhang Y and Chen Y. Clinicopathological significance of c-MYC in esophageal squamous cell carcinoma. Tumour Biol 2017; 39: 1010428-317715804.
[46] Todorov IT, Pepperkok R, Philipova RN, Kearsey SE, Ansorge W and Werner D. A human nuclear protein with sequence homology to a family of early S phase proteins is required for entry into S phase and for cell division. J Cell Sci 1994; 107: 253-265.
[47] Katayama H, Zhou H, Li Q, Tatsuka M and Sen S. Interaction and feedback regulation between STK15/BTAK/Aurora-A kinase and protein phosphatase 1 through mitotic cell division cycle. J Biol Chem 2001; 276: 46219-46224.
[48] Cheung CHY, Hsu CL, Chen KP, Chong ST, Wu CH, Huang HC and Juan HF. MCM2-regulated functional networks in lung cancer by multi-dimensional proteomic approach. Sci Rep 2017; 7: 13302.
[49] Tanaka H, Nakashiro K, Iwamoto K, Tokuzen N, Fujita Y, Shirakawa R, Oka R, Goda H and Hamakawa H. Targeting Aurora kinase A suppresses the growth of human oral squamous cell carcinoma cells in vitro and in vivo. Oral Oncol 2013; 49: 551-559.
[50] Singer JD, Gurian-West M, Clurman B and Roberts JM. Cullin-3 targets cyclin E for ubiquitination and controls S phase in mammalian cells. Genes Dev 1999; 13: 2375-2387.
[51] Loignon M, Miao W, Hu L, Bier A, Bismar TA, Scrivens PJ, Mann K, Basik M, Bouchard A, Fiset PO, Batist Z and Batist G. Cul3 overexpression depletes Nrf2 in breast cancer and is associated with sensitivity to carcinogens, to oxidative stress, and to chemotherapy. Mol Cancer Ther 2009; 8: 2432-2440.
[52] Miyazaki M and Esser KA. REDD2 is enriched in skeletal muscle and inhibits mTOR signaling in response to leucine and stretch. Am J Physiol Cell Physiol 2009; 296: C583-592.
[53] Morquette B, Morquette P, Agostinone J, Feinstein E, McKinney RA, Kolta A and Di Polo A. REDD2-mediated inhibition of mTOR promotes dendrite retraction induced by axonal injury. Cell Death Differ 2015; 22: 612-625.
[54] Zhang X, Luo F, Li J, Wan J, Zhang L, Li H, Chen A, Chen J, Cai T, He X, Lisse TS and Zhao H. DNA Damage-Inducible Transcript 4 is an innate guardian for human squamous cell carcinoma and an molecular vector for anti-carcinoma effect of 1,25(OH)2 D3. Exp Dermatol 2019; 28: 45-52.
[55] Eckhart L, Lippens S, Tschachler E and Declercq W. Cell death by cornification. Biochim Biophys Acta 2013; 1833: 3471-3480.

DISCUSIÓN GENERAL

La presente discusión está dividida en tres partes, la primera y la segunda, tratan sobre lo encontrado en la primera y segunda publicación respectivamente, la tercera parte trata sobre un cierre general a lo mostrado por ambos artículos. El primer artículo versa sobre el uso de una métrica obtenida de la teoría de la información la cual es la entropía. Esta medida se escogió dados los antecedentes, algunas de las definiciones de esta son: 1) el número de bits en promedio que se requieren para describir a una variable aleatoria, 2) el número de preguntas necesarias para identificar a una variable y 3) una medida de la incertidumbre promedio de una sola variable en bits (Cover & Thomas, 2006). Tomando en consideración los antecedentes que se tenían hasta ese momento, de la relación de esta métrica para el estudio de redes de cáncer y de los patrones de expresión genética, se realizó una unificación de todos estos con la finalidad de obtener nuevos resultados, cuyo uso pudiera servir para el tratamiento y detección de cuatro tipos de cáncer, que fueron seleccionados dada su importancia, disponibilidad de datos de pacientes y de datos acerca del proceso carcinogénico. El carcinoma de células escamosas de pulmón tiene una alta proporción en el número de casos de cáncer de pulmón, así como el hepatocarcinoma dentro del hepático; además estos dos forman parte de los principales tipos de cáncer con más fallecimientos y nuevos casos reportados en el mundo. A su vez, el cáncer de páncreas y el melanoma a pesar de no tener tantos casos como los mencionados anteriormente, estos tienen un importante papel dado que tienen altas tasas de mortalidad. En el caso del melanoma es la principal causa de muerte por cáncer de piel. La entropía funcionó como un filtro adicional al tratamiento tradicional de datos para análisis de expresión diferencial de genes, el cual incluye el cálculo del *Fold Change*, el valor p ajustado por *FDR* y la visualización de estos mediante gráficas *volcano*. En el trabajo se utilizaron los valores de expresión normalizados, integrados a las redes, y el concepto de redes locales para la creación de matrices, que permitieron el cálculo de entropía, lo que permitió observar un patrón generalizado en estos cuatro tipos de cáncer, el cual fue un cambio en el valor de entropía durante la transición pre-cáncer/cáncer. El patrón observado consistió en un súbito cambio entre los estadios pre-cancerígeno al cancerígeno: un aumento en la entropía en aquellos datos que provenían únicamente de la expresión de genes, un descenso en aquellos donde los datos provenían de las redes locales y la expresión de los genes de cada uno de los nodos que componían dichas redes. En estos últimos se obtuvo que dichos cambios son estadísticamente significativos en tres de los cuatro tipos de cáncer analizados. Este cambio en la entropía muestra una similitud como uno de los detectados en algunos sistemas complejos el cual es conocido como desaceleración crítica. Este fenómeno, genera cambios característicos en el comportamiento de un sistema y ocurre cuando se encuentra cerca de una transición crítica. Este fenómeno se ha observado de diferentes maneras como el cambio en la varianza y la autocorrelación de diferentes parámetros medidos, por ejemplo, lo que ocurre en el cerebro antes de que exista una crisis epiléptica, donde hay un incremento súbito en la varianza de la señal segundos antes de que se dé la crisis. En el caso de la entropía este cambio súbito esta principalmente influenciado por los patrones de expresión que se están teniendo en cada uno de los estadios, muestra la incertidumbre que existe entre un estadio y el otro respecto a la señal detectada; los patrones de expresión diferencial cuando se encuentra en el estadio del cáncer son muy diferentes respecto a lo que se encuentra normalmente. El punto en donde ocurre la transición crítica en este caso es cuando se pasa al estadio de cáncer. Además de este análisis general, se utilizó a la entropía, como una medida adicional a los filtros de *Fold Change* y el *FDR*, con el principal motivo de poder remarcar de manera puntual a genes cuyos cambios en su expresión y en las funciones biológicas de sus productos proteicos estuvieran promoviendo el desarrollo tumoral en cada uno de los estadios. Una vez realizado esto, se obtuvieron diferentes grupos de genes en cada uno de los estadios. Estos grupos de genes, presentaban cambios en su entropía, que fueron visiblemente detectables aplicando una escala para el análisis de cada una de las redes de cáncer.

Algunos de estos genes, habían sido señalados previamente en otras investigaciones, debido a su función en el apoyo al desarrollo carcinogénico en el mismo tipo de cáncer o en otros tipos. Varios de estos genes fueron resaltados como importantes para un futuro estudio, en donde se analice el impacto que tienen estos, en el progreso de los tipos de cáncer en los cuales estaban teniendo el cambio, la selección se basó en aquellos que no tenían estudios previos específicos para el tipo de cáncer en el que fueron resaltados, y cuya función biológica que normalmente desempeñan podría llegar a tener una influencia en el desarrollo del cáncer. Los genes seleccionados como de importancia para la transición pre-cáncer a cáncer en cada uno de los cuatros tipos de cáncer fueron los siguientes: para melanoma, FDX1, MUT, VAMP4, DENND4A, para hepatocarcinoma, CEP41, para el cáncer pancreático, FAR1, HCCS y para el carcinoma de células escamosas de pulmón, PIH1D2. Todos estos participan en procesos muy variados, como es el caso de FAR1 el cual cataliza la reducción de acil-CoA graso de C16 o C18 saturado e insaturado a alcoholes grasos, o el caso de CEP41 que es requerido durante la ciliogénesis para la glutamilación de la tubulina en los cilios (*UniProt*, s/f).

El segundo trabajo se realizó en el carcinoma de células escamosas, específicamente fue un análisis *in silico* de todo el proceso carcinogénico el cual ya se encuentra descrito en la literatura, sin embargo, hasta ese momento y para conocimiento de los autores no existía un estudio de este tipo que englobara a todo el proceso. El proceso consiste de 7 estadios, dentro de los cuales se encuentra el estadio donde el cáncer ya se encuentra establecido. Adicionalmente se utilizó el estadio normal el cual fungió como control. En este trabajo, se lograron identificar los procesos biológicos y la función molecular que principalmente se encontraban afectados y los cuales tendrían una gran relevancia para el desarrollo de este cáncer, para esto se utilizaron dos programas de enriquecimiento de datos basados en la ontología génica. Un ejemplo de un proceso biológico señalado es la cornificación, el cual es considerado como un tipo de muerte celular, en el cual va aumentando la queratinización. La queratinización es una de las principales características citológicas que se presentan en este cáncer. Este proceso se observó muy enriquecido durante los estadios tempranos que aún no son cancerígenos. Adicionalmente se realizó la construcción de redes con los resultados del análisis de expresión diferencial. Usando las métricas estadísticas para redes conocidas como medidas de centralidad, se realizó la identificación de manera rápida de aquellos componentes dentro de la red que pudieran ser importantes. Cabe destacar, que el uso de estas medidas es muy amplio en diferentes áreas del conocimiento. Se usaron dos, de las cuales los estudios en estudios en levadura han demostrado ser acertadas al identificar componentes funcionales que tienen gran impacto en la estabilidad y desarrollo de las células a nivel biológico y que fueron comprobados experimentalmente. Las medidas fueron las conocidas como intermediación y grado. Este estudio identificó algunos genes, que previamente han sido señalados como *drivers* en algunos tipos de cáncer como por ejemplo MYC y AURKA, que se encontraban sobreexpresados y actúan en la promoción del desarrollo en este tipo de cáncer, este último gen relacionado con el progreso del cáncer de pulmón y por otro lado en el caso de CUL3 la cual se encontraba subexpresada concuerda con lo encontrado en otros estudios, donde se ha visto que actúa como un supresor de tumor para este tipo de cáncer (Gabay et al., 2014; Inamura, 2017, p. 3; Lo Iacono et al., 2011). En este trabajo se realizó el uso de un término el cual no había sido utilizado en este tipo de redes biológicas, el cual es: *gatekeeper,* esto se logró a través del estudio visual de las redes en donde se seleccionaron los nodos cuyos primeros vecinos estaban ampliamente conectados a pesar de que estos nodos seleccionados no lo estaban. El término *gatekeeper* hace alusión a lo que su propio nombre indica: “guardián de la puerta”. En este trabajo se está proponiendo que los productos proteicos, los cuales están representados en las redes, fungen como mediadores del acceso a diferentes procesos biológicos, y esto sucede en cualquier red de interacción proteína-proteína. Por tanto, este concepto puede ser aplicado a cualquier red que sea de este tipo.

El presente trabajo logró obtener diversos resultados, que a grandes rasgos denotan nuevas opciones de grupos de genes, que pudieran ser usados en los distintos tipos de cáncer analizados con diferentes fines ya sean terapéuticos, de monitoreo o para obtener un mejor entendimiento del cambio que ocurre en los sistemas de regulación celular en las distintas etapas del proceso carcinogénico en distintos tipos de cáncer; además, en el segundo artículo se está realizando por primera vez un análisis y descripción integral de manera *in silico* de lo que sucede en este tipo de cáncer así como un aporte de tipo conceptual, dado que se está incluyendo el uso de un concepto nuevo, que hemos propuesto para la interpretación de los resultados obtenidos mediante el análisis de redes. Este nuevo concepto no sólo aplica en la enfermedad del cáncer, sino que es general para todo tipo de redes de interacción proteína-proteína (PPI por sus siglas en inglés) y cuyo origen es de una medida de centralidad que forma parte de la teoría de redes, que ahora se le está asignando una interpretación biológica y con esto se abre un nuevo panorama para una interpretación más funcional cuando se usa esta métrica de redes.

CONCLUSIONES GENERALES

El análisis de distintos tipos de cáncer usando diferentes métricas y métodos permitieron la identificación de un cambio general y súbito, que acontece en cuatro tipos de cáncer durante su proceso carcinogénico, el cual está representado por la entropía de la información. Dicho cambio es representativo de lo que va sucediendo en los distintos estadios, dado sus patrones de expresión genética. Los grupos de genes identificados por cada estadio pueden tener potencial impacto en el desarrollo de cada uno de los cánceres y podrían ser considerados para una evaluación experimental y conocer su viabilidad como candidatos para el desarrollo de nuevas terapías. En el caso específico del carcinoma de células escamosas de pulmón, se encontró de manera teórica a los principales procesos y genes afectados en cada uno de los estadios de su carcinogénesis. Sin embargo, aún se necesitan extensos estudios experimentales primeramente a nivel *in vitro* y subsecuentemente a nivel *in vivo,* para incentivar y conocer la viabilidad de estos en el desarrollo de posibles nuevas terapias. Por último, la introducción de un nuevo término en el análisis de redes biológicas de interacción proteína-proteína permitirá una interpretación más amigable de los resultados obtenidos de estas.

RECOMENDACIONES GENERALES

- Realizar los análisis presentados en estos estudios en diversas poblaciones
- Ampliar el número de muestras usadas y a su vez utilizar *single-cell sequencing* para evaluar distintas poblaciones dentro de un mismo cáncer
- Analizar otros tipos de cáncer con esta metodología para excluir y corroborar aquellos donde el resultado respecto a la entropía sea similar a los mostrados en estos trabajos.
- Evaluar a nivel experimental el impacto de los genes mencionados para el proceso carcinogénico.

REFERENCIAS BIBLIOGRÁFICAS

Balogh, J., Victor, D., Asham, E. H., Burroughs, S. G., Boktour, M., Saharia, A., Li, X., Ghobrial, M., & Monsour, H. (2016). Hepatocellular carcinoma: A review. *Journal of Hepatocellular Carcinoma*, *Volume 3*, 41–53. https://doi.org/10.2147/JHC.S61146

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71. https://doi.org/10.1016/j.socnet.2004.11.008

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: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Brehme, M., Koschmieder, S., Montazeri, M., Copland, M., Oehler, V. G., Radich, J. P., Brümmendorf, T. H., & Schuppert, A. (2016). Combined Population Dynamics and Entropy Modelling Supports Patient Stratification in Chronic Myeloid Leukemia. *Scientific Reports*, *6*(1), 24057. https://doi.org/10.1038/srep24057

Breitkreutz, D., Hlatky, L., Rietman, E., & Tuszynski, J. A. (2012). Molecular signaling network complexity is correlated with cancer patient survivability. *Proceedings of the National Academy of Sciences*, *109*(23), 9209–9212. https://doi.org/10.1073/pnas.1201416109

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed). Wiley-Interscience.

Dehmer, M. (Ed.). (2011). *Structural Analysis of Complex Networks*. Birkhäuser Boston. https://doi.org/10.1007/978-0-8176-4789-6

Derman, B. A., Mileham, K. F., Bonomi, P. D., Batus, M., & Fidler, M. J. (2015). Treatment of advanced squamous cell carcinoma of the lung: A review. *Translational Lung Cancer Research*, *4*(5), 524–532. PMC. https://doi.org/10.3978/j.issn.2218-6751.2015.06.07

Drilon, A., Rekhtman, N., Ladanyi, M., & Paik, P. (2012). Squamous-cell carcinomas of the lung: Emerging biology, controversies, and the promise of targeted therapy. *The Lancet Oncology*, *13*(10), e418–e426. https://doi.org/10.1016/S1470-2045(12)70291-7

El cáncer en el mundo y México. (s/f). INFOCáncer. Recuperado el 13 de mayo de 2020, de http://www.infocancer.org.mx/images/logo_infocancer_reg_mexico.png

Gabay, M., Li, Y., & Felsher, D. W. (2014). MYC Activation Is a Hallmark of Cancer Initiation and Maintenance. *Cold Spring Harbor Perspectives in Medicine*, *4*(6), a014241–a014241. https://doi.org/10.1101/cshperspect.a014241

Gandara, D. R., Hammerman, P. S., Sos, M. L., Lara, P. N., & Hirsch, F. R. (2015). Squamous Cell Lung Cancer: From Tumor Genomics to Cancer Therapeutics. *Clinical Cancer Research*, *21*(10), 2236–2243. https://doi.org/10.1158/1078-0432.CCR-14-3039

Ghasemi, M., Seidkhani, H., Tamimi, F., Rahgozar, M., & Masoudi-Nejad, A. (2014). Centrality Measures in Biological Networks. *Current Bioinformatics*, *9*(4), 426–441. https://doi.org/10.2174/15748936113086660013

Gridelli, C., Rossi, A., Carbone, D. P., Guarize, J., Karachaliou, N., Mok, T., Petrella, F., Spaggiari, L., & Rosell, R. (2015). Non-small-cell lung cancer. *Nature Reviews Disease Primers*, *1*(1), 15009. https://doi.org/10.1038/nrdp.2015.9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Heist, R. S., Sequist, L. V., & Engelman, J. A. (2012). Genetic Changes in Squamous Cell Lung Cancer: A Review. *Journal of Thoracic Oncology*, *7*(5), 924–933. https://doi.org/10.1097/JTO.0b013e31824cc334

Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature*, *553*(7689), 446–454. https://doi.org/10.1038/nature25183

Hirsch, F. R., Scagliotti, G. V., Mulshine, J. L., Kwon, R., Curran, W. J., Wu, Y.-L., & Paz-Ares, L. (2017). Lung cancer: Current therapies and new targeted treatments. *The Lancet*, *389*(10066), 299–311. https://doi.org/10.1016/S0140-6736(16)30958-8

Inamura, K. (2017). Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification. *Frontiers in Oncology*, *7*, 193. https://doi.org/10.3389/fonc.2017.00193

Liu, R., Li, M., Liu, Z.-P., Wu, J., Chen, L., & Aihara, K. (2012). Identifying critical transitions and their leading biomolecular networks in complex diseases. *Scientific Reports*, *2*, 813. https://doi.org/10.1038/srep00813

Lo Iacono, M., Monica, V., Saviozzi, S., Ceppi, P., Bracco, E., Papotti, M., & Scagliotti, G. V. (2011). Aurora Kinase A expression is associated with lung cancer histological-subtypes and with tumor de-differentiation. *Journal of Translational Medicine*, *9*(1), 100. https://doi.org/10.1186/1479-5876-9-100

Marcus, F. B. (2008). *Bioinformatics and Systems Biology*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78353-4

Miller, A. J., & Mihm, M. C. (2006). Melanoma. *New England Journal of Medicine*, *355*(1), 51–65. https://doi.org/10.1056/NEJMra052166

Moore, A., & Donahue, T. (2019). Pancreatic Cancer. *JAMA*, *322*(14), 1426. https://doi.org/10.1001/jama.2019.14699

O'Neill, C. H., & Scoggins, C. R. (2019). Melanoma. *Journal of Surgical Oncology*, *120*(5), 873–881. https://doi.org/10.1002/jso.25604

Prokop, A., & Csukás, B. (Eds.). (2013). *Systems Biology*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6803-1

Raman, K. (2010). Construction and analysis of protein–protein interaction networks. *Automated Experimentation*, *2*(1), 2. https://doi.org/10.1186/1759-4499-2-2

Raza, A. (2014). Hepatocellular carcinoma review: Current treatment, and evidence-based medicine. *World Journal of Gastroenterology*, *20*(15), 4115. https://doi.org/10.3748/wjg.v20.i15.4115

Schadendorf, D., Fisher, D. E., Garbe, C., Gershenwald, J. E., Grob, J.-J., Halpern, A., Herlyn, M., Marchetti, M. A., McArthur, G., Ribas, A., Roesch, A., & Hauschild, A. (2015). Melanoma. *Nature Reviews Disease Primers*, 15003. https://doi.org/10.1038/nrdp.2015.3

Scheffer, M., Carpenter, S. R., Lenton, T. M., Bascompte, J., Brock, W., Dakos, V., van de Koppel, J., van de Leemput, I. A., Levin, S. A., van Nes, E. H., Pascual, M., & Vandermeer, J. (2012). Anticipating Critical Transitions. *Science*, *338*(6105), 344–348. https://doi.org/10.1126/science.1225244

Scheffer, Marten, Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, *461*(7260), 53–59. https://doi.org/10.1038/nature08227

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(1), 7–30. https://doi.org/10.3322/caac.21590

Singh, V., & Dhar, P. K. (Eds.). (2015). *Systems and Synthetic Biology*. Springer Netherlands. https://doi.org/10.1007/978-94-017-9514-2

Teschendorff, A. E., Sollich, P., & Kuehn, R. (2014). Signalling entropy: A novel network-theoretical framework for systems analysis and interpretation of functional omic data. *Methods*, *67*(3), 282–293. https://doi.org/10.1016/j.ymeth.2014.03.013

UniProt. (s/f). Recuperado el 29 de agosto de 2020, de https://www.uniprot.org/

Waller, L. P. (2015). Hepatocellular carcinoma: A comprehensive review. *World Journal of Hepatology*, *7*(26), 2648. https://doi.org/10.4254/wjh.v7.i26.2648

Waller, L. P., Deshpande, V., & Pyrsopoulos, N. (2015). Hepatocellular carcinoma: A comprehensive review. *World Journal of Hepatology*, *7*(26), 2648–2663. https://doi.org/10.4254/wjh.v7.i26.2648

What Is Cancer? (s/f). [CgvArticle]. National Cancer Institute. Recuperado el 26 de octubre de 2017, de https://www.cancer.gov/about-cancer/understanding/what-is-cancer

WHO. (2018). *Cancer , Health topic*. https://www.who.int/news-room/fact-sheets/detail/cancer

WHO | Cancer. (s/f). WHO. Recuperado el 26 de octubre de 2017, de http://www.who.int/mediacentre/factsheets/fs297/en/