UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

PLATAFORMA COMPUTACIONAL PARA INTEGRAR Y ANALIZAR DATOS DEL ATLAS DEL GENOMA DEL CÁNCER (TCGA) EN EL ESTUDIO DEL FENÓMENO DE COMPENSACIÓN DE DOSIS GÉNICA Y SU EFECTO EN LA SOBREVIDA DE PACIENTES CON CÁNCER

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Ciencias Biomédicas para optar al grado de Maestría Académica en Bioinformática y Biología de Sistemas

GUILLERMO OVIEDO BLANCO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2021

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Ciencias Biomédicas de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica de Bioinformática y Biología de Sistemas"

MARIELA EUGENIA ARIAS HIDALGO (FIRMA)
Firmado digitalmente por MARIELA EUGENIA ARIAS HIDALGO (FIRMA)
Fecha: 2022.01.24 16:41:31 -06'00'

Dra. Mariela Arias Hidalgo
Representante de la Decana
Sistema de Estudios de Posgrado

RODRIGO MORA RODRIGUEZ (FIRMA)
PERSONA FISICA, CPF-01-1035-0030.
Fecha declarada: 24/01/2022 03:51:32 PM
Esta representación visual no es fuente de confianza. Valide siempre la firma.

PhD. Rodrigo Mora Rodríguez
Director de Tesis


Firmado digitalmente por JOSE ARTURO MOLINA MORA (FIRMA)
Fecha: 2022.01.25 09:48:24 -06'00'

PhD. José Arturo Molina Mora
Asesor

WARNER ENRIQUE ALPIZAR ALPIZAR (FIRMA)
Firmado digitalmente por WARNER ENRIQUE ALPIZAR ALPIZAR (FIRMA)
Fecha: 2022.01.25 17:14:45 -06'00'

Dr. Warner Alpízar Alpízar
Asesor

CHRISTOPHER RIDGELY VAGLIO CEDEÑO (FIRMA)
Firmado digitalmente por CHRISTOPHER RIDGELY VAGLIO CEDEÑO (FIRMA)
Fecha: 2022.01.26 19:24:19 -06'00'

PhD. Christopher Vaglio Cedeño
Representante de la Directora
Programa de Posgrado

GUILLERMO OVIEDO BLANCO (FIRMA)
Firmado digitalmente por GUILLERMO OVIEDO BLANCO (FIRMA)
Fecha: 2022.01.26 23:04:50 -06'00'

Guillermo Oviedo Blanco
Sustentante

TABLA DE CONTENIDO

RESUMEN

La enfermedad del cáncer se va gestando a lo largo del tiempo mediante la acumulación de cambios en el comportamiento celular que son producto de alteraciones epigenéticas y mutaciones en el material genético. Esta actividad da forma a un fenotipo que logra, de manera sorprendente y hasta contradictoria, proliferar bajo condiciones internas que serían mortales para células sanas. Una de estas condiciones letales es la aneuploidía, entendida como el desbalance en la cantidad de cromosomas y por ende en la cuota normal de genes, lo cual incide consecuentemente en la abundancia o escasez de proteínas. La compleja respuesta celular a esta situación incluye circuitos biológicos o redes de interacción de diversos elementos moleculares que a su vez pueden estar siendo afectados por dicha inestabilidad genómica. Representando una ventaja biológica en el caso del cáncer, el cambio en la dosis de ciertos genes -cuya expresión podría activar procesos de muerte celular- es compensada de alguna forma dentro de esos circuitos biológicos, permitiendo la proliferación de las células que conforman las masas tumorales.

Determinar con exactitud la topología y propiedades de esas redes de interacción permitiría vislumbrar estrategias para modularlas y por ende propiciar terapias para detener la progresión de la enfermedad. Este es uno de los retos de un nuevo tipo de desarrollo científico y tecnológico que ha sido impulsado por la generación de grandes volúmenes de datos genómicos, tal como el proyecto del Atlas del Genoma del Cáncer (TCGA).

En el presente trabajo se creó una plataforma computacional que posibilita no sólo la adquisición, integración, transformación y adaptación de datos del TCGA, sino también la aplicación sobre ellos de métodos matemáticos de análisis, con el fin de apoyar el proceso de caracterización del fenómeno de compensación de dosis génica, el estudio de otros problemas científicos relacionados con la regulación de la expresión génica mediada por micro-ARN (miARN) y su relación con la sobrevida global de los pacientes de cáncer.

El principal aporte de esta tesis al estudio de la compensación de dosis génica consistió en la confirmación de la presencia de dicho fenómeno en muestras de cáncer de mama, utilizando como un criterio inicial el nivel de tolerancia a la variación en la expresión con respecto a la variación en el número de copias de genes. Posteriormente y mediante el uso de datos de expresión adaptados a las alteraciones en el número de copias se definió una medida más robusta para representar el grado de compensación de dosis génica. Esta nueva métrica permitió identificar genes compensados, cuya amplificación diferencial entre grupos de pacientes con alta y baja sobrevida los potencia como candidatos de análisis para encontrar posibles dianas terapéuticas.

Además se ajustó un modelo matemático de las interacciones de una red miARN y factores de transcripción para el oncogen MYC -gen candidato- para cada paciente, lo que permitió caracterizar la heterogeneidad de este fenómeno en datos de muestras clínicas tumorales y facilitó la identificación de una propiedad emergente a nivel de sistema de ciclos de retroalimentación. La compensación de dosis génica del oncogen MYC en cáncer de mama es regulada principalmente por el miARN miR-19a y no por los otros miARN del modelo matemático, lo cual subraya la importancia el estudio personalizado de este fenómeno para identificar las potenciales dianas terapéuticas para cada paciente.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

ADN	Ácido desoxirribonucleico
API	Interfaz de programación de aplicaciones
ARNcirc	ARN circulares
ARNlnc	ARN largos no codificantes
ARNm	ARN mensajero
ARNnc	ARN no codificante
BioNetUCR	Plataforma bioinformática para la identificación de nuevas inferencias del fenómeno de compensación de dosis génica
CCLE	Enciclopedia de líneas de células cancerosas
CES	Fuentes estimadas de consenso
consenso-ICA	Componentes de consenso independientes
CNA	Alteraciones del número de copias de genes
CV-ADN	Coeficiente de variación del número de copias
CV-ARN	Coeficiente de variación de la expresión génica
GDCDP	Portal de datos comunes de datos genómicos
GDSC	Genómica de la sensibilidad a fármacos en el cáncer
GEO	Ómnibus de expresión génica
GENCODE	Anotación del genoma humano de referencia
GMM	Modelo de mezclas gaussiana
GRCh38	Archivo con anotaciones del genoma humano
HGNC	Comité de Nomenclatura Genética de la Organización del Genoma Humano
LabQT	Laboratorio de Quimio-Sensibilidad Tumoral
miARN	micro-ARN
NCI-60	Repositorio de 60 líneas celulares de tumores humanos del Instituto Nacional de Cáncer
ODE	Ecuaciones diferenciales ordinarias
RFE	Eliminación recursiva de características
SBML	Lenguaje de marcado para biología de sistemas
SNP	Polimorfismos de un nucleótido
SVM	Máquina de soporte vectorial
TACNA	Adaptación transcripcional a las alteraciones en el número de copias
TCGA	Atlas del Genoma del Cáncer

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Guillermo Oviedo Blanco, con cédula de identidad 107050390, en mi condición de autor del TFG titulado Plataforma computacional para integrar datos del Atlas del Genoma del Cáncer (TCGA) en el estudio del fenómeno de compensación de dosis génica y su efecto en la sobrevida de los pacientes con cáncer

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción:** _____ **año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

GUILLERMO OVIEDO BLANCO (FIRMA)

Firmado digitalmente por GUILLERMO OVIEDO BLANCO (FIRMA)
Fecha: 2022.01.26 23:15:48 -06'00'

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

1. ANTECEDENTES

1.1. Cáncer y mutaciones

El cáncer es un conjunto de enfermedades relacionadas que pueden iniciarse en casi cualquier parte del cuerpo humano y que tienen en común alteraciones en mecanismos moleculares de procesos que están relacionados con funciones vitales como la proliferación, el desarrollo y la muerte celular programada. En todos los tipos de cáncer el proceso ordenado de envejecimiento, muerte y reemplazo celular se descontrola y esto conduce a un escenario en el que células dañadas sobreviven y proliferan en lugar de morir. En muchos casos dichas células adquieren la capacidad de invadir tejidos cercanos o moverse a sitios lejanos del origen de la enfermedad, afectando a otros órganos y sistemas del cuerpo humano (Hanahan & Weinberg, 2011).

Según estadísticas de la "International Agency for Research on Cancer" (https://gco.iarc.fr) sólo en 2020 más de 19 millones de personas fueron diagnosticadas con la enfermedad y se reportaron 9.9 millones de muertes. Ese mismo año en Costa Rica se presentaron 13139 casos y el número de muertes llegó a 6000. Se prevé que para 2040 la cantidad de personas diagnosticadas de cáncer en el mundo ascenderá a 29.5 millones y las muertes a 16.4 millones.

Las primeras ideas, surgidas a finales del siglo XIX y principios del XX, de que el cáncer está formado por clones defectuosos de células que contienen anomalías en el material hereditario, fueron asentándose cuando -posterior al descubrimiento de la estructura del Ácido Desoxirribonucleico (ADN) y su rol como sustrato molecular de herencia- se demostró que puede ser causado por agentes que generan mutaciones y dañan el ADN (Loeb & Harris, 2008). El desarrollo de los análisis citogenéticos (de los cromosomas) en células cancerígenas mostraron que la incidencia recurrente de ciertas aberraciones cromosómicas se asociaba con ciertos tipos de cáncer; esto condujo a que por primera vez -en 1982- se identificara una mutación como origen de esta enfermedad (Reddy *et al.*,

1982; Tabin *et al.*, 1982). A partir de ese momento, la intensa búsqueda de genes anormales subyacentes al desarrollo del cáncer ha llevado a comprender el mismo como la consecuencia fenotípica de un proceso que se ajusta a varios principios de evolución Darwiniana alimentado por la acumulación de alteraciones genéticas, genómicas y epigenéticas y el impacto de éstas en la conformación del microambiente de un tejido específico.

El avance científico de los últimos cuarenta años ha permitido comprobar la relación entre el cáncer y los cambios genómicos, al descubrir y catalogar un número importante de mutaciones que confieren ganancia de función en genes específicos (oncogenes) o pérdida de función en otros (supresores tumorales) (Hanahan & Weinberg, 2000). Para la prevención, detección y tratamiento del cáncer, es vital completar este catálogo de alteraciones genéticas y alcanzar un mayor entendimiento no sólo de las redes de transducción de señales en las que se ven envueltos los oncogenes y los supresores de tumores, sino también de cómo varían las consecuencias biológicas de las mutaciones dependiendo del genotipo general, tipo de célula, el estado de desarrollo y el microambiente (Chin *et al.*, 2011). Para acelerar los procesos que hacen posible la transformación de descubrimientos genómicos en un producto aplicable desde el punto de vista clínico, es indispensable la resolución de retos científicos y logísticos de los cuales interesa resaltar la disposición de un genoma de referencia de cáncer, el desarrollo de análisis bioinformáticos sobre ese genoma y el descubrimiento de la bases moleculares y bioquímicas de la actividad oncogénica de posibles blancos terapéuticos (Chin *et al.*, 2011).

Contar con un genoma de referencia de cáncer resulta indispensable para la toma de decisiones de índole médica frente al perfil molecular del tumor de un paciente, ya que permitiría disponer de puntos de control terapéutico en los que se esconden dos tipos de mutaciones que están presentes en cáncer: aquellas que proporcionan ventajas de crecimiento a las células cancerosas (mutaciones conductoras) y otras que no inciden en la actividad oncogénica (mutaciones pasajeras) (Stratton *et al*., 2009). Un esfuerzo notable en

esta línea es el llevado a cabo por el Atlas del Genoma del Cáncer (TCGA), el cual no sólo consolida una gran base de datos de información genómica sino que también ha establecido todo un cuerpo de entes, normas y estándares a nivel internacional para garantizar la calidad y homogeneidad de la recolección de muestras biológicas y datos relacionados.

En esencia el cáncer surge a partir de la acumulación de mutaciones conductoras que afectan el funcionamiento de circuitos biológicos, los cuales gobiernan actividades tales como la comunicación, división celular, adaptación, proliferación y homeostasis; sin embargo, estas alteraciones no se presentan en el mismo orden ni en el mismo momento aún en el mismo tipo de tejido. Considerando la diversidad de tipos de cáncer y tumores, buscar un pequeño grupo de circuitos común a todos ellos es la meta del desarrollo científico en este campo para los próximos años (Hanahan and Weinberg, 2000). Descubrir la arquitectura de las redes de señales en las que están involucradas las mutaciones conductoras es el quehacer de un nuevo tipo ciencia abarcado por la biología de sistemas. Buscar formas de diferenciar mutaciones conductoras y pasajeras es una tarea que se realiza sobre esas bases de datos genómicas con análisis bioinformáticos en los que se requiere la convergencia de distintas ramas del conocimiento: biología, matemática, estadística, ingeniería y computación.

1.2. Aneuploidía

Los rasgos distintivos de prácticamente todos los tipos de cáncer incluyen ocho capacidades funcionales (Mantenimiento de las señales proliferativas, Evasión de los supresores de crecimiento, Desregulación de la energética celular, Evasión de la destrucción inmune, Resistencia a la muerte celular, Habilitación del potencial replicativo ilimitadio, Inducción de la angiogénesis y la Activación de la invasión tisular y la metástasis) y dos características habilitadoras (Inflamación favorecedora al tumor y la Inestabilidad genómica) (Hanahan and Weinberg, 2011). De hecho, la inestabilidad genómica impulsa la robustez del cáncer, visto como un sistema cuyos componentes corresponden a

subpoblaciones de células tumorales. Tal robustez se entiende como la capacidad de proliferar y evolucionar (propiedades del sistema) a pesar de las terapias anticáncer y respuestas imunológicas y microambientales del organismo (perturbaciones externas). La robustez es habilitada en gran parte por la redundancia funcional. Un ejemplo de redundancia funcional se aprecia en una subpoblación particular de células (componente funcionalmente redundante del sistema) que sobrevive a la terapia y conduce a la recurrencia del tumor; esto es posible por la heterogeneidad intratumoral que es causada por la inestabilidad genómica (Kitano, 2004). La aneuploidía -ganancia o pérdida, de forma parcial o total, de cromosomas (Hassolt *et al.*, 2007)- es uno de los principales tipos de inestabilidad genómica. A la fecha no se conocen los mecanismos que le permiten a las células cancerosas tolerar los efectos adversos de la aneuploidía para proliferar y volverse inmortales (Holland and Cleveland, 2009; Schvartzman *et al.*, 2010). Este tipo de anomalía genómica se considera como una pérdida de equilibrio (Torres & Amon, 2008) con efectos tan profundos sobre la fisiología y división celular (Torres *et al.*, 2007) que debería ser fatal para el cáncer (Williams & Amon, 2009). A pesar de eso, la aneuploidía representa una paradoja ya que el cariotipo desequilibrado también puede conducir a beneficios para una célula en particular (Torres *et al.*, 2010; Sheltzer & Amon, 2011; Pavelka *et al.*, 2010).

En un lado de esta paradoja, la aneuploidía que se considera un sello distintivo del cáncer -está presente en el 90% de los tumores sólidos y el 75% de los tumores hematológicos- ha sido valorada como la fuerza impulsora de la evolución genómica del cáncer (Sheltzer y Amon, 2011). En efecto, un estudio ha señalado a la aneuploidía como la principal fuente de inestabilidad genómica autocatalítica, donde las células con la mayor aneuploidía tienen también la mayor inestabilidad de sus genomas (Duesberg, *et al*., 1998). La aneuploidía parece tener un papel en la oncogénesis dado que aparece antes de la transformación maligna, presenta alteraciones genéticas clonales que comprometen la fidelidad de la segregación cromosómica; de hecho existen mutaciones hereditarias en puntos de control del ciclo celular y segregación cromosomal que conducen a la aneuploidía y predisponen al cáncer (Sheltzer & Amon, 2011) .

La aneuploidía es letal tanto para células normales como para organismos complejos y representa la causa más frecuente de abortos y retraso mental en humanos. La mayoría de las monosomías (con la excepción del cromosoma X) nunca se observan en humanos y también muy raramente se observan en abortos espontáneos (Hardy & Hardy, 2015), presumiblemente porque esta condición no es compatible con la supervivencia celular o la implantación del óvulo fertilizado (Kojima & Cimini, 2019). Por otra parte, 20 de las 23 posibles trisomías (ganancia de un cromosoma adicional) son letales y de las otras tres, sólo la trisomía 21 (síndrome de Down) puede sobrevivir y alcanzar la adultez (Sheltzer & Amon, 2011). También conduce a muchos defectos a nivel celular como lo son la disminución en la proliferación y la viabilidad celular, el aumento del estrés proteotóxico, los requisitos metabólicos, la producción de lactato y la inducción de defectos de recombinación que llevan a la inestabilidad genómica, entre otros. La explicación más probable de todos esos efectos negativos de la aneuploidía está dada por la alteración de la dosis del gen: las ganancias o pérdidas de cromosomas enteros alteran inmediatamente las dosis de cientos de genes en la célula, lo que lleva a una carga desequilibrada de productos génicos (i.e. proteínas o ARN no codificantes) críticos, alterando los requerimientos energéticos y homeostasis celular (Sheltzer & Amon, 2011). De hecho, durante la carcinogénesis, la aneuploidía autocataliza la inestabilidad genómica de las células cancerosas, lo que conduce a muchos cariotipos inestables y muerte celular debido a un errores catastróficos (Solé & Deisboeck, 2004) .

En el otro lado de la paradoja, en algunas ocasiones (aunque muy poco frecuentes), se da una combinación específica de diversas alteraciones que supera esos umbrales de error anteriormente mencionados y propicia la evolución del cáncer, dando lugar a células malignas que son capaces de sobrevivir a pesar de la aneuploidía y la inestabilidad genómica. Este cuello de botella en la dinámica evolutiva del cáncer representa una puerta para la generación de cariotipos malignos que conducen a la farmacorresistencia y la metástasis (Li *et al*., 2009). Dentro del caos de la inestabilidad cromosómica, algunos

patrones conservados en las configuraciones cariotípicas sugieren la presencia de un mecanismo estable, cuya función debe mantenerse para asegurar la supervivencia: i) existen aneusomías específicas en diferentes etapas de la transformación celular (Fabarius, *et al.*, 2008), ii) hay cariotipos clonales que evolucionan durante las fases celulares (Fabarius, *et al.*, 2002) , iii) los cariotipos que causan cáncer tienen un equilibrio cromosómico entre la aneuploidía desestabilizadora y la selección estabilizadora para la función oncogénica (Li *et al*., 2009) y iv) un estudio a gran escala reveló dos vías distintas hacia la aneuploidía donde las células ganan o pierden cromosomas para restaurar el equilibrio de sus proteínas alteradas y mantener la viabilidad (Ozery-Flato, *et al.*, 2011). Por tanto, a pesar de la inestabilidad genómica del cáncer, estas observaciones sugieren la existencia de un mecanismo estable para hacer frente a los efectos negativos de la aneuploidía en las células cancerosas.

1.3 Compensación de dosis génica

Se desconoce cómo las células cancerosas se enfrentan a tanta aneuploidía mientras que las células normales son muy sensibles a ella. Una pista sobre la naturaleza de estos mecanismos desarrollados por las células cancerosas para minimizar los efectos negativos de la aneuploidía es el hecho de que la hiperdiploidía (número de cromosomas superior al normal) se observa con más frecuencia que la hipodiploidía (número de cromosomas inferior al normal) (Weaver & Cleveland, 2006; Cimini, 2008), y que por esto, en la mayoría de los casos en los que ocurre la aneuploidía es muy probable que las células posean un número de cromosomas por encima de la norma. Entonces, una posible explicación viene dada por la hipótesis de compensación de la dosis génica, un mecanismo por el cual se modula la expresión de ciertos genes para compensar las diferencias en la dosis génica cuando hay cromosomas adicionales debido a la aneuploidía (Kojima& Cimini, 2019).

La compensación de la dosis génica ha sido descrita en estudios muy tempranos para otros organismos que compensan los efectos negativos de la aneuploidía (Devlin, *et al.*, 1982). De hecho, el concepto de compensación de la dosis génica, o equilibrio de la dosis génica, es un fenómeno generalizado que se descubrió en los primeros días de la genética y hay cúmulos de pruebas de que tiene un efecto sobre la expresión de genes, los rasgos cuantitativos, los síndromes aneuploides, la dinámica poblacional resultante de variantes en el número de copias y el destino evolutivo diferencial de genes producto de duplicaciones completas o parciales del genoma (Birchler & Veitia, 2012).

Se han identificado varios mecanismos potenciales de compensación de dosis génica en distintos niveles regulatorios de la expresión de un gen y en la formación de un producto biológico final. A nivel de síntesis de proteínas, se ha planteado la hipótesis de que los efectos de la compensación son el resultado de diferencias estequiométricas entre los miembros de los complejos macromoleculares, el interactoma y las vías de señalización (Birchler & Veitia, 2012; Veitia *et al*., 2008). Se ha demostrado que los niveles de ARN mensajero (ARNm) generalmente correlacionan bien con el número de copias de ADN, pero estos cambios no se reflejan a nivel de proteína para ciertos genes (Stingele et al., 2012). Se ha observado que las múltiples consecuencias de la aneuploidía que se regulan a nivel protéico se deben a un efecto secundario de los defectos de plegamiento y al aumento de la degradación por parte del proteosoma y la autofagia (Donnelly & Storchová, 2014). También, un estudio mostró que cuando hay un exceso las subunidades se degradan o agregan y que la agregación es casi tan efectiva como la degradación para reducir el nivel de proteínas funcionales (Brennan et al., 2019). Un enfoque para identificar genes con compensación de dosis, mediante el aumento del número de copias de genes individuales utilizando la técnica del “tug-of-war” genético, mostró que aproximadamente el 10% del genoma muestra compensación de dosis génica. Esta consiste, predominantemente, en la degradación o agregación de subunidades en exceso que son parte de complejos de múltiples proteínas, lo cual significa que sus niveles están regulados de manera dependiente

de la estequiometría (Ishikawa *et al*., 2017). Cabe señalar que este enfoque se diseñó para identificar genes compensados solo a nivel de proteínas.

La regulación a nivel transcripcional podría ser otro mecanismo eficaz para compensar los cambios de dosis génica en las células aneuploides, ya que mantiene la estequiometría y conserva la energía necesaria para la transcripción, traducción y eventual degradación de las proteínas adicionales (Donnelly & Storchová, 2014). En este nivel, se han utilizado varios modelos experimentales para investigar los efectos de la aneuploidía en la expresión génica con hallazgos contradictorios. Hay estudios que han encontrado que la adquisición de un cromosoma adicional da como resultado un aumento proporcional en la expresión de los genes que se encuentran en ese cromosoma Otros, por su parte, han informado que la aneuploidía también puede perturbar los niveles de expresión génica en otros cromosomas. Sin embargo, varios estudios han indicado de algún tipo de compensación, por el cual los genes en el cromosoma aneuploide no se expresan en los niveles esperados según la dosis génica, reportándose que esto ocurre para números variables de genes según el contexto y cambiando además el grado de compensación. Adicionalmente, se reporta que los efectos de la compensación de la dosis pueden ocurrir simultáneamente sobre otros cromosomas (Kojima & Cimini, 2019).

No hay consenso tampoco con respecto a la extensión de la compensación. Algunos estudios indican que sólo se compensan unos pocos genes. Por ejemplo la inserción de un cromosoma 5 adicional reveló que la mayoría de las proteínas codificadas en los cromosomas adicionales son más abundantes que las proteínas de los cromosomas diploides, lo que indica que no existe un mecanismo general eficaz para la compensación de la dosis génica en este sistema. Sin embargo, algunas proteínas específicas se mantienen a niveles diploides, especialmente las correspondientes a quinasas y subunidades ribosómicas. La mayoría de estos genes se compensan a nivel de proteínas, pero algunos otros también se compensan a nivel de ARNm (Stingele *et al*., 2012). Por el contrario, para ciertos modelos experimentales, este fenómeno de compensación de dosis parece ser

amplio y afectar a una gran fracción del contenido del gen aneuploide. Un informe sobre cepas de levadura salvaje aneuploide mostró una compensación de dosis en el 10-30% de los genes amplificados en comparación con cepas euploides isogénicas o estrechamente relacionadas, y que la aneuploidía no condujo a defectos de crecimiento. También predijeron que la compensación de dosis ocurre en genes que son más tóxicos cuando se sobreexpresan y que su expresión también puede estar bajo una mayor restricción evolutiva (Hose *et al*., 2015).

El Laboratorio de Quimio-Sensibilidad Tumoral (LabQT), del Centro de Investigación de Enfermedades Tropicales de la Universidad de Costa Rica, llevó a cabo una investigación (Acón *et al.*, 2016) en la que se argumentó que el mecanismo de compensación de dosis génica podría proporcionar estabilidad al cáncer a pesar de su inestabilidad genómica. Dicho estudio, además, contribuye a explicar por qué las células cancerosas toleran altos niveles de aneuploidía, mientras que las células normales son muy sensibles con la ganancia de un solo cromosoma. También planteó la idea de que ese mecanismo estuviera mediado, al menos en parte, por las propiedades emergentes de redes complejas de miARN y factores de transcripción que controlan la expresión de genes que tienen alterado el número de copias. Finalmente, en ese mismo estudio se propuso el diseño de una red reguladora de un conjunto de genes candidatos obtenidos con base en datos genómicos del NCI-60 (Boyd, 1997).

Los miARN, de los cuales se detallará más adelante, constituyen uno de los grupos de ARN no codificante (ARNnc) más predominantemente representados en la investigación clínica en años recientes. De hecho, durante la última década los ARNnc que existen dentro de la célula han pasado de ser considerados productos transcripcionales "basura" a moléculas con importante funciones reguladoras que median los procesos celulares, incluida la remodelación de la cromatina, la transcripción, las modificaciones postranscripcionales y la transducción de señales (Anastasiadou *et al.,* 2018). Además de los miARN -y gracias a las tecnologías de secuenciación- se han descubierto ARN circulares (cirARN) y ARN largos

no codificantes (ARNlnc). Las redes en las que participan los ARNnc pueden influir en numerosos objetivos moleculares para impulsar respuestas y destinos biológicos celulares específicos (Yamamura *et al*., 2018). En consecuencia, los ARNnc actúan como reguladores clave de los programas fisiológicos en contextos de desarrollo y enfermedades. Los ARNnc han demostrado ser particularmente relevantes en la biología del cáncer dadas sus propiedades oncogénicas y de supresión de tumores en todos los tipos principales de cáncer (Anastasiadou *et al.,* 2018).

Los miARN son pequeñas moléculas de ARN endógeno que se unen a los ARNm y reprimen la expresión génica (Fabian *et al*., 2010). Un miARN típico se procesa a partir de una secuencia larga de ARN primario hasta dar pie a una versión funcional madura y corta de alrededor de 22 nucleótidos de longitud. Una característica común de los miARN es su capacidad para actuar pleiotrópicamente uniéndose a los ARNm codificados por varios genes (Hanna *et al*., 2019;Ritchie *et al*., 2013). De hecho, las estimaciones actuales apuntan a que el genoma humano contiene 1917 precursores anotados y 2654 secuencias maduras de miARN (Kozomara *et al*., 2019), que se presume regulan directamente más del 60% de los ARNm humanos (Kim *et al.*, 2016). En consecuencia, existe una buena posibilidad de que las interacciones miARN y factor de transcripción -que forman diferentes tipos de motivos reguladores (retrocontrol positivo, retrocontrol negativo entre miARN, factores de transcripción y genes diana)- controlen la expresión de genes amplificados o eliminados en el cáncer.

1.4. Perfil de adaptación transcripcional a las alteraciones en el número de copias (TACNA)

La identificación de genes con compensación de dosis se dificulta debido a que la correlación entre las alteraciones del número de copias de genes (CNA) y la expresión génica podría verse alterada por muchos factores diferentes. De hecho, debido a los mecanismos de adaptación de la transcripción, las alteraciones en el número de copias no siempre se traducen proporcionalmente en niveles de expresión alterados. Para explorar

asociaciones entre CNA y niveles de expresión génica se utiliza, actualmente, el enfoque de genómica genética que combina perfiles de expresión genética y genómica obtenidos de muestras tumorales. No obstante, resulta muy difícil detectar los efectos particulares de los CNA sobre los niveles de expresión génica. Esto se debe a que el perfil de expresión génica se realiza a menudo en muestras de tejido que incluyen tanto células tumorales como no tumorales del microambiente tumoral, lo cual implica que se mida la expresión promedio de todos los tipos de células presentes en dichas muestras. Esto significa que los efectos de los CNA en los niveles de expresión génica, además de estar influenciados por factores experimentales y no genéticos, a menudo se ven eclipsados por los efectos de las células no tumorales. Consecuentemente, el grado de adaptación transcripcional a las CNA sigue sin estar claro para la mayoría de los genes a pesar de los esfuerzos actuales por esclarecer este aspecto (Bhattacharya *et al*., 2020)

Para determinar con precisión el grado de adaptación transcripcional a los CNA utilizando genómica genética, se necesita de un gran número de parejas de perfiles de expresión y de número de copias provenientes de muestras tumorales. Desafortunadamente, se encuentran disponibles pocos conjuntos de este tipo de datos a gran escala, además de que muchos de los perfiles de expresión de muestras tumorales de acceso público no tienen un perfil de número de copias emparejado, lo que implica que estas muestras sean actualmente excluidas de los análisis de genómica genética. Se requieren, por lo tanto, criterios novedosos para desentrañar de mejor manera -a escala genómica- el grado de adaptación transcripcional a los CNA y, por ende, la identificación de genes con compensación de dosis. *Bhattacharya et al* han desarrollado un método independiente de plataforma llamado "Adaptación Transcripcional al Perfil de CNA" (perfil TACNA), el cual extrae los efectos de los CNA a partir de un único perfil de expresión génica, generado a partir de una biopsia de tumor, sin la necesidad de un perfil de número de copias emparejado. Con este método el nivel de expresión neto de un gen se determina mediante los efectos combinados de varios factores reguladores de la transcripción, incluidos factores experimentales, genéticos (como las CNA) y no genéticos. Para identificar los efectos de estos factores en los niveles

de expresión génica, primero se aplica un método computacional denominado "Componentes de Consenso Independientes" (consenso-ICA), en el cual los perfiles de expresión génica se calculan como el resultado de la suma de "Fuentes Estimadas de Consenso" (CES), con los CES estadísticamente independiente entre ellos tanto como sea posible. Los autores plantearon dos hipótesis. La primera es que cada CES describe el efecto de un factor latente de regulación de transcripción sobre los niveles de expresión génica, lo cual significa que para cada CES cada gen tiene asociado un peso que describe cuán fuertemente y en qué dirección su nivel de expresión está influenciado por cada factor latente de regulación transcripcional. La segunda hipótesis es, que los CES que capturaban el efecto de las CNA sobre los niveles de expresión génica, son aquellos que mostraron consistentemente el siguiente patrón: sólo los genes que se ubican en una región genómica contigua específica tenían un peso absoluto alto (Bhattacharya *et al*., 2020). Los perfiles TACNA se obtuvieron de aproximadamente 28.000 muestras de tejido tumoral provenientes de TGCA, Cancer Cell Line Encyclopedia (CCLE), Gene Expression Omnibus (GEO) y Genomics of Drugs Sensitivity in Cancer (GDSC).

El desarrollo de estos perfiles ha proporcionado una nueva herramienta para las investigaciones que, como las del LabQT, tienen por objetivo caracterizar los mecanismos subyacentes de la adaptación transcripcional al CNA. En el caso particular del fenómeno de compensación de dosis génica dicha exploración puede conducir a la identificación de genes compensados que, por sus niveles de expresión estrechamente controlados, representen potenciales dianas terapéuticas .

1.5. El Altas del Genoma del Cáncer (TCGA)

El proyecto TCGA fue iniciado en 2005 por el ”National Institute of Health, U.S. Department of Health and Human Services” con el propósito de crear un atlas exhaustivo de perfiles genómicos de cáncer para catalogar y descubrir las alteraciones genéticas implicadas en el desarrollo y progresión del cáncer. En este esfuerzo, con una inversión de

al menos 50 millones de dólares en sus primeros tres años, ha convergido el trabajo y recursos del "National Cancer Institute of National Institute of Health", "National Human Genome Research Institute" y diversas entidades alrededor de Estados Unidos y Europa (Tomczak *et al.*, 2015), contando en la actualidad, muestras biológicas de más de 84.000 pacientes en 66 tipos de tumores. A partir de muestras biológicas recolectadas, el TCGA lleva a cabo secuenciaciones de alto rendimiento y análisis bioinformáticos que derivan en datos de diversos tipos tal como expresión génica, expresión de miARN, variación del número de copias, polimorfismos de un nucleótido (SNP), pérdida de heterocigosidad, mutaciones, metilación de ADN y expresión de proteínas.

Para generar esa gama de datos el TCGA está organizado por un grupo de entidades que se encargan de las tareas de recolección, procesamiento de muestras y análisis bioinformáticos. El proceso inicia en las que recolectan sangre y tejido de pacientes con cáncer (Tissue Source Sites, TSSs), de ahí estos bioespecímenes se trasladan a un centro de almacenamiento (Biospecimen Core Resource, BCR) en donde se cataloga, procesa y verifica la calidad de cada uno. Luego los datos clínicos y metadatos son enviados a un centro de coordinación (Data Coordinating Center, DCC), por su parte los analitos moleculares son sometidos a caracterizaciones genómicas y secuenciaciones de alto rendimiento (Genome Characterization Centers, GCCs y Genome Sequencing Centers, GSCs). Los datos sobre secuenciación de esos centros se envían al DCC, datos sobre secuencias y alineamientos van a un repositorio seguro (NCI´s Cancer Genomics Hub, CGHub). Los datos genómicos van a entidades que aplican nuevos análisis y generan información adicional (Genome Data Analisys Centers, GDACs).

Todo el flujo de datos generado por esta red de entidades es administrado por el DCC y está disponible a la comunidad científica a través del portal del TCGA (Genomic Data Commons Data Portal, GDCDP), ya sea mediante el acondicionamiento de un sitio web o la interacción con una interfaz de programación de aplicaciones (API). La colección de datos genómicos de cáncer que el TCGA pone a disposición de los investigadores posibilita

la expansión del conocimiento de la carcinogénesis e impulsa un mayor entendimiento de la biología del cáncer y el desarrollo de nuevos métodos de diagnóstico y terapias. Algunos ejemplos de nuevos descubrimientos utilizando datos del TCGA se describen a continuación.

Investigadores de la Universidad de Melbourne (Trigos *et al.*, 2017) analizaron datos de siete tipos de cáncer del TCGA para investigar posibles vínculos entre los tumores y propiedades de organismos unicelulares ancestrales. En este estudio se descubrió una estrecha asociación entre la edad del gen y el nivel de expresión. Los genes conservados de organismos unicelulares estaban fuertemente sobre expresados, mientras que los genes que regulan procesos multicelulares (de origen metazoario) estaban principalmente desregulados a la baja (niveles de expresión menores). Se sugiere que interacciones anormales entre vías multicelulares y vías unicelulares podrían ser blancos de estrategias clínicas aplicables a una amplia gama de cánceres. Este estudio proporcionó el primer análisis molecular a gran escala entre los tumores y la historia evolutiva, con el fin de aportar evidencia a la exploración de la teoría -llamada atavismo- de que el cáncer revierte a procesos celulares antiguos. Esta teoría considera que las características distintivas del cáncer como, el crecimiento ilimitado, la evasión de la muerte celular programada y la desdiferenciación hacia un tipo de células menos especializadas; son propiedades comunes en organismos unicelulares que prosperan porque crecen y se dividen sin detenerse; no obstante, al evolucionar los organismos multicelulares desarrollaron células especializadas con mayor control sobre el crecimiento y la muerte celular.

Las inmunoterapias con inhibidores de puntos de control inmunitario -que evitan que las células tumorales evadan el ataque inmune- son actualmente muy exitosas para el tratamiento de muchos tipo de cáncer. A pesar de esto, sólo cierto tipo de pacientes reaccionan de forma positiva a la terapia. Con el objetivo de comprender de mejor forma los atributos genéticos que median en la respuesta a la inmunoterapia, se llevó a cabo un estudio sobre datos de doce tipos de cáncer del TCGA, con el fin de comprender el impacto

que las alteraciones en el número de copias de su genoma tendría en la enfermedad (Davoli *et al.*, 2017). Se encontró que los cánceres que albergan muchas alteraciones tienden a mostrar menos compromiso inmunitario y peor respuesta a las inmunoterapias mencionadas. La consideración de estos hallazgos junto a otros conocimientos de la respuesta inmunitaria puede ayudar a los científicos y más adelante a los médicos, a distinguir cuáles pacientes se beneficiarán o no de la terapia, mejorando así los resultados y minimizando los efectos secundarios.

Como parte de los procedimientos actuales para la atención del cáncer de pulmón, se realizan análisis histológicos de imágenes bajo el microscopio para distinguir los diferentes tipos, etapas y grados del tumor. Sin embargo, dichos análisis frecuentemente presentan mucha subjetividad por las diferencias de criterio entre los patólogos que evalúan las muestras. Por ejemplo, para el caso de cáncer de pulmón de células no pequeñas el porcentaje de concordancia en el diagnóstico entre observadores es de aproximadamente 60% y la clasificación clínica no permite establecer un pronóstico del período de sobrevivencia luego del diagnóstico del cáncer en un paciente. Para desarrollar un mejor método de diagnóstico y predicción se aplicó un enfoque de aprendizaje automático sobre un conjunto de 2.000 imágenes de cáncer de pulmón del TCGA (Yu *et al.*, 2016). Haciendo uso de tecnología de procesamiento de imágenes, los investigadores construyeron una plataforma de análisis que logró detectar 10.000 características de las cuales 240 fueron usadas por los algoritmos de aprendizaje para diagnosticar cánceres de pulmón de células no pequeñas. Dentro de esas características se encuentran las texturas de los núcleos celulares y las distribuciones de intensidad de los pixeles que son muy difíciles de reconocer para el ojo humano. Los algoritmos encontraron patrones en una cuadrícula de 1.000.000 de pixeles. Se emplearon como métodos de aprendizaje clasificadores Bayesianos, variantes de máquinas de soporte vectorial y variantes de bosques aleatorios. El análisis estadístico de los resultados demostró que se puede predecir con éxito la supervivencia a corto o largo plazo del carcinoma de células escamosas y el diagnóstico de adenocarcinoma.

El glioblastoma es el tipo de tumor cerebral más común y mortal en los adultos debido a la alta frecuencia de recaídas. Se analizaron los datos genómicos de 114 muestras de este tipo de tumor maligno que fueron tomadas en el momento del diagnóstico y al momento de la recaída, además de datos de tumores recurrentes del TCGA (Wang *et al*., 2016). El estudio encontró que es muy probable que los tumores recurrentes aparecieran antes del diagnóstico inicial, ya que se encontraron mutaciones en los glioblastomas iniciales que no estaban presentes en los recurrentes. Considerando la baja probabilidad de que una mutación revierta a su secuencia original, es posible que el clon que se descubrió después del tratamiento haya divergido del clon original antes del diagnóstico, y que el tratamiento, actuando como una presión selectiva, haya sido el detonante para que estos clones se expandieran y repoblaran el tumor. Al analizar la frecuencia de las nuevas mutaciones se determinó que el tiempo de divergencia entre el diagnóstico y los clones recurrentes fue más de 10 años, lo cual advierte la presencia de sub-tumores o clones al momento de diagnosticar un glioblastoma. Esta perspectiva contribuye a explicar porque el glioblastoma es difícil de curar y mejora la comprensión de la historia evolutiva de este tipo de cáncer, lo cual puede servir de base para investigar más a fondo y, en última instancia, desarrollar tratamientos más efectivos.

1.6. Reconocimiento de patrones y aprendizaje automático

Esfuerzos como los llevados a cabo por el TCGA, propulsados por el desarrollo de tecnologías de secuenciación, han producido un notable crecimiento en la cantidad de datos biológicos disponibles para análisis (Figura 1A). Éstos son, precisamente, el insumo que se requiere para avanzar en uno de los mayores retos de la biología computacional: el desarrollo de herramientas y métodos capaces de transformar esa masa de datos en conocimiento biológico. Dos campos del conocimiento que pueden aportar instrumentos para este tipo de tareas son el reconocimiento de patrones y el aprendizaje automático.

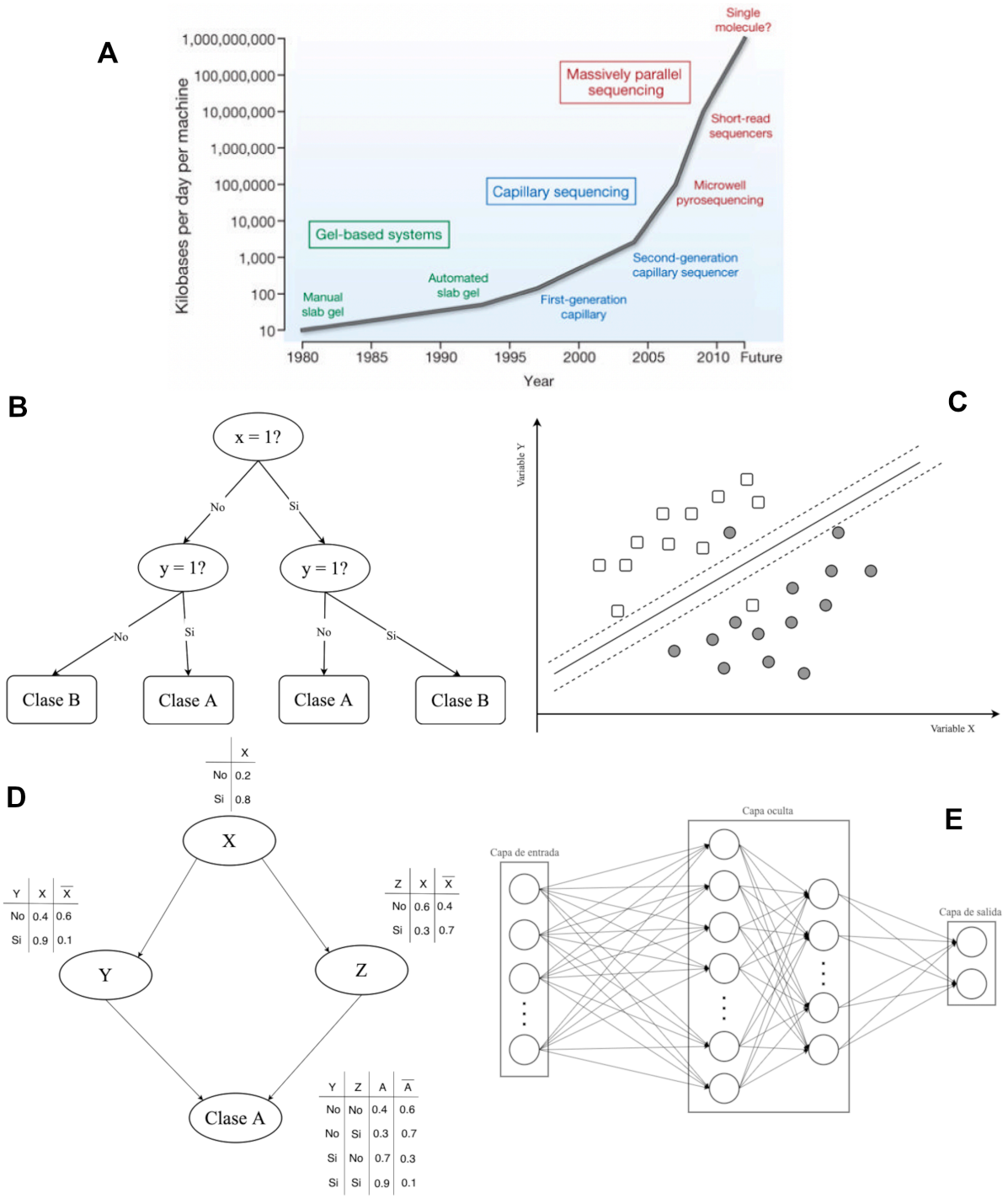


Figura 1. Reconocimiento de patrones y aprendizaje automático. A. Tasa de incremento en la secuenciación de ADN en función del avance en las tecnologías de secuenciación (Straton et al., 2010). **B.**Estructura de un árbol de decisión. De acuerdo a los valores de las características de un registro de datos -que se comparan en lo nodos internos- éste se clasifica en alguna de las clases de los nodos hoja. **C.**En un conjunto de datos de sólo dos características (X y Y) un hiperplano -en este caso una recta en un espacio de dos dimensiones- permite clasificar los datos en dos clases. Se aprecia un par de datos mal clasificados. **D.**En una red Bayesiana los arcos representan relaciones de interdependencia entre las características de un conjunto de datos y a cada nodo se asocia una distribución de probabilidad a partir de sus padres. **E.**En una red neural artificial cada nodo de la capa oculta y de salida es un perceptron. Los nodos de la capa de entrada corresponden cada uno a una característica del conjunto de datos. Conforme la red se entrena las configuraciones internas de los nodos de la capa oculta se auto-ajustan para "aprender" a clasificar los datos, el resultado de esa clasificación se envía a los nodos de salida.

El reconocimiento de patrones se origina en la ingeniería y está relacionado con la aplicación de algoritmos computacionales para el descubrimiento de regularidades en grandes conjuntos de datos y con el uso de esas regularidades para la toma de decisiones. El reconocimiento de patrones realizado por máquinas resulta muy útil en el reconocimiento de voz, identificación de huellas dactilares, reconocimiento óptico de caracteres e identificación de secuencias de ADN, por citar algunas aplicaciones (Duda *et al.*, 2012; Bishop, 2006). El contexto general en el que se aplica el reconocimiento de patrones implica la existencia de una masa de datos -registros con valores para una serie de características que describen una realidad específica-, que servirá para entrenar algoritmos que al detectar patrones puedan categorizar o agrupar nuevos datos -con la misma serie de características de los datos de entrenamiento- que nunca antes habían sido examinados.

Un sistema o proceso de reconocimiento de patrones incluye métodos y herramientas para llevar a cabo tareas de recopilación de datos, preprocesamiento, selección de características, entrenamiento y validación. Una vez que se adquieren datos es común que éstos se preprocesen para llevar a cabo tareas de control de calidad que eliminen o permuten ruido, valores atípicos, duplicados y valores nulos; también puede resultar útil, en esta fase, que los datos sean transformados en un nuevo espacio de variables donde el problema de reconocimiento de patrones sea más sencillo de resolver. Otras labores involucran la ejecución de técnicas como la reducción de la dimensionalidad y la selección y/o extracción de características con el fin de eliminar redundancia, acelerar los procesos de cómputo y amoldar de mejor forma el conjunto de datos a los métodos de búsqueda de patrones. Finalmente, en el entrenamiento y la validación se aplican métodos que buscan esos patrones en los datos y los representan en la forma de funciones o modelos matemáticos que pueden entonces ser utilizados para analizar nuevos datos para categorizarlos o agruparlos (Hastie *et al*., 2005; Duda *et al.*, 2012; Bishop, 2006).

Gestado en las ciencias de la computación e informática el aprendizaje automático trata con la creación y evaluación de algoritmos que buscan maximizar las tasas de reconocimiento

de patrones, clasificación y predicción mediante la construcción de modelos matemáticos que aprenden de los datos. Tanto el reconocimiento de patrones como el aprendizaje automático han evolucionado en conjunto desde sus raíces en ingeniería e inteligencia artificial, integrando ideas, métodos y herramientas hasta convertirse en facetas de un mismo campo (Bishop, 2006).

El aprendizaje automático se puede subdividir a su vez en dos categorías. Se denomina supervisado cuando un conjunto de datos etiquetados (cuando las características de un fenómeno están asociadas a una serie de clases) es usado para detectar el mapa o función entre las variantes de los datos y las etiquetas; luego ese mapa es utilizado para clasificar nuevos datos en las posibles clases o categorías. En el caso de aprendizaje automático no-supervisado, los datos no están etiquetados (no se conocen las clases a las que pertenecen) y los métodos de aprendizaje detectan patrones o descubren grupos sobre esos datos; esos patrones son empleados para asignar nuevos datos al grupo con características similares. Cuando un fenómeno está descrito por un conjunto de datos cuyos registros no están totalmente etiquetados se pueden utilizar métodos que combinan el aprendizaje supervisado y no-supervisado; a este enfoque se le llama aprendizaje semi-supervisado. Otras técnicas que se conocen como aprendizaje reforzado buscan encontrar, para una situación determinada, las acciones que generen una máxima ganancia o recompensa. En este último caso de aprendizaje, el algoritmo no recibe datos categorizados sino que los descubre mediante un proceso de prueba y error que le permite interactuar con su entorno mediante una secuencia de acciones y estados. Usando este tipo de técnicas un algoritmo puede, por ejemplo, aprender a jugar Backgammon con un alto nivel de desempeño (Hastie *et al*., 2005; Bishop, 2006).

Posterior a las tareas de recolección de datos, preprocesamiento y selección de características, la fase de entrenamiento puede realizarse probando y evaluando técnicas de aprendizaje que incluyen, por mencionar algunos, árboles de decisión, máquinas de soporte vectorial, redes Bayesianas, redes neurales artificiales, clasificadores Bayesianos, regresión

logística, análisis de discriminantes, bosques aleatorios, el vecino más cercano, agrupamiento particional o jerárquico, modelos de mezclas y modelos de Markov. Características y conceptos claves de algunos de esos algoritmos se describen a continuación.

Los árboles de decisión (de Rider *et al.*, 2013; Pang *et al*, 2006) estructuran el conocimiento como grafos dirigidos en forma de árbol. En estas estructuras los nodos internos representan valores o comparaciones sobre las características del conjunto de datos, mientras que los nodos hojas corresponden a las etiquetas o clases asociadas. Son muy utilizados para propósitos de clasificación. Cuando se tienen nuevos registros de datos que requieren ser clasificados se usan los valores de sus características para recorrer el árbol y predecir la clase asociada (Figura 1B).

La inducción de conjuntos de reglas, o aprendizaje de reglas, es un modelo de clasificación basado en un conjunto de sentencias de la forma "si ocurre un condicional entonces ocurre una consecuencia". Cada regla tiene una conjunción de valores de características en la parte condicional y una etiqueta de clase en el consecuente. Como alternativa a tales reglas lógicas, se pueden inducir reglas probabilísticas. Además de la etiqueta de la clase predicha, el resultado de estas reglas también consiste en una lista de probabilidades o números de instancias de entrenamiento cubiertas para cada posible etiqueta de clase (Clark y Boswell, 1991). Las ideas clave para aprender tales conjuntos de reglas son bastante similares a las ideas utilizadas en la inducción del árbol de decisiones. Sin embargo, en lugar de dividir de forma recursiva el conjunto de datos optimizando la medida de pureza en todos los nodos sucesores, los algoritmos de aprendizaje de reglas solo se expanden un solo nodo sucesor a la vez, aprendiendo así una regla completa que cubre parte de los datos de entrenamiento. Una vez que se ha aprendido una regla completa, todos los ejemplos cubiertos por esta regla se eliminan del conjunto de entrenamiento y el procedimiento se repite con los ejemplos restantes (Fürnkranz, 1997). Las reglas de decisión son probablemente los modelos de predicción más interpretables.

Las máquinas de soporte vectorial (SVM) (de Rider *et al.*, 2013; Pang *et al*, 2006) crean un mapa entre los valores del conjunto de datos y un espacio multidimensional -cuyo tamaño depende de la cantidad de características de los datos- y luego, tomando en cuenta las etiquetas, encuentra un hiperplano que separa los datos en dos clases, para finalmente maximizar las distancias que separan los puntos más cercanos al hiperplano. Para clasificar nuevos registros se utilizan los valores de sus características para ubicarlos como puntos en el espacio multidimensional y determinar el grupo de pertenencia de acuerdo al hiperplano (Figura 1C).

Las redes Bayesianas (de Rider *et al.*, 2013; Pang *et al*, 2006), además de utilizarse en problemas de clasificación se emplean como medio de representación de razonamiento y conocimiento, estructurando éste como un grafo dirigido acíclico que representa dependencias probabilísticas entre las variables del conjunto de datos y las posibles clases o etiquetas. Para construir una red Bayesiana primero se genera el grafo que mejor representa las relaciones de interdependencia entre las características del conjunto de datos y luego, para cada nodo, se calcula una distribución local de probabilidad en función de sus nodos padres; existe un nodo que contiene la distribución de probabilidad de la clase o etiqueta (Figura 1D).

Las redes neurales artificiales (de Rider *et al.*, 2013; Pang *et al*, 2006) son un paradigma de programación basado en la interconexión de módulos llamados perceptrones. Un perceptrón es un modelo de neurona artificial (Rosenblatt, 1958), que permite ajustar una serie de parámetros internos asociados con los valores de las características de un conjunto de datos para generar un resultado que podría estar ligado a los valores de una clase o etiqueta. Una red neural artificial está formada por grupos de perceptrones interconectados entre sí, formando capas de procesamiento en donde cada neurona artificial se auto-ajusta dependiendo de los resultados de los ajustes de las otras neuronas artificiales. Este tipo de procesamiento hace que la forma en que la red neural “aprende” sea una especie de caja negra ya que no se puede saber cómo la red estructuró su conocimiento (Figura 1E).

El algoritmo K-medias identifica clústeres en los datos tratando de separar muestras en grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo. El algoritmo requiere que se especifique el número de clústeres “k” y divide el conjunto de datos en “k” grupos disjuntos, cada uno descrito por la media de los elementos del grupo denominada centroide. Básicamente, el algoritmo tiene tres pasos. Primero se eligen los centroides iniciales; comúnmente se seleccionan “k” elementos del conjunto de datos. Seguidamente se asigna cada elemento a su centroide más cercano, luego se crean nuevos centroides tomando el valor medio de todas las muestras asignadas a cada centroide anterior. Finalmente, se calcula la diferencia entre el antiguo y el nuevo centroide y el algoritmo repite estos dos últimos pasos hasta que este valor sea menor que un umbral o, lo que es equivalente, se repite hasta que los centroides no se mueven significativamente (Bishop, 2006).

DBSCAN es otro algoritmo para la identificación de clústeres a los cuáles identifica como regiones continuas de alta densidad. Debido a esta visión bastante genérica, los grupos encontrados por DBSCAN pueden tener cualquier forma, a diferencia de k-medias que asume que los grupos tienen forma convexa. Para cada elemento del grupo de datos, el algoritmo cuenta cuántos elementos están distanciados de él dentro de una pequeña distancia ε (épsilon). Esta región se denomina vecindad ε del elemento y éste se considera un elemento central si tiene al menos cierta cantidad (un parámetro del algoritmo) de elementos en su vecindario ε (incluyéndose a sí mismo). En otras palabras, los elementos centrales son aquellos que están ubicados en regiones densas. Todos los elementos cercanos a un elemento central pertenecen al mismo clúster. Esto puede incluir otros elementos centrales, por lo tanto, una secuencia larga de elementos centrales vecinos forma un solo clúster. Cualquier elemento que no sea un elemento central y no tenga uno en su vecindario se considera una anomalía (Ester, 1996).

Un algoritmo adicional para la identificación de clústeres es un modelo probabilístico denominado modelo de mezcla gaussiana (GMM). Este modelo asume que los puntos de datos se generaron a partir de una mezcla de varias distribuciones gaussianas cuyos parámetros se desconocen. Todos los elementos generados a partir de una única distribución gaussiana forman un grupo que normalmente parece un elipsoide. Cada grupo puede tener una forma, tamaño, densidad y orientación elipsoidal diferente (Bishop, 2006).

La relación entre el aprendizaje automático y la biología se puede calificar como extensa e intensa. Por ejemplo, el perceptron fue propuesto en 1958 para representar el comportamiento neuronal (Rosenblatt, 1958) y a su vez, la misma biología del sistema nervioso visual sirvió de inspiración para inventar arquitecturas de redes neuronales artificiales (Carpenter and Grossberg, 1988; Fukushima, 1980). La evolución de las técnicas de aprendizaje automático, así como desarrollos matemáticos para medir la confiabilidad de dichas técnicas, ha facilitado que esos algoritmos contribuyan a mejorar la eficiencia del descubrimiento y entendimiento del gran volumen y complejidad de datos biológicos. Por ejemplo, en el campo de la estimación del riesgo de cáncer de mama, una red neural artificial con un capa de entrada de 36 características -correspondientes a datos clínicos acerca de edad, terapias, historial personal o familiar de cáncer de mama, descriptores mamográficos y categorizaciones radiológicas- de 48.744 mamogramas realizados a 18.269 pacientes y una capa oculta conformada por 1.000 nodos se usó para mapear hacia un nodo de salida la indicación de que un mamograma presentara una lesión maligna (Ayer *et al.*, 2010). Para ello la red se entrenó usando iteraciones de validaciones cruzadas dividiendo la totalidad de mamogramas en 9 grupos de entrenamiento y 1 de prueba. Usando sensibilidad y especificidad como métricas del rendimiento de la red se determinó que ésta alcanzó un 96.5% en comparación con un 93.9% de los radiólogos.

En el mismo campo de cáncer de mama, pero enfocado en predecir el riesgo de metástasis luego de una cirugía, se desarrolló una plataforma de máquinas de soporte vectorial para analizar datos de micro-arreglos de expresión génica de pacientes posterior a una

intervención quirúrgica (Xu *et al*., 2012). Esto con la intención de determinar la conveniencia de aplicar sesiones de quimioterapia. El algoritmo se utilizó para seleccionar entre miles de genes un grupo de 50 cuya expresión fuera relevante para el diagnóstico hacia dos categorías: metástasis o libre de metástasis, esto con el fin de determinar casos en los que se recomendaría la quimioterapia. Se emplearon datos de expresión de 25.000 genes de 295 pacientes, de los cuales 208 pertenecían al grupo de metástasis y 87 al libre de metástasis. Usando el coeficiente de correlación de Pearson el grupo de genes se redujo a 300 y, sobre este conjunto se aplicaron iteraciones de un método recursivo, basado en máquinas de soporte vectorial, para la eliminación de características. Al terminar cada iteración de entrenamiento -que usó validación cruzada de 1 muestra para prueba y el resto para entrenar- se calcularon los coeficientes de los vectores de soporte para encontrar los 5 genes con menor peso; estos genes se eliminaban del conjunto y se iniciaba otra iteración. Este flujo de trabajo se aplicó hasta que no quedaron genes en el conjunto de datos y en cada iteración se calculó el rendimiento en función de la exactitud, sensibilidad y especificidad, determinándose que el máximo fue alcanzado por un grupo de 50 genes. Un análisis comparativo entre la expresión de esos 50 genes y la clase del diagnóstico, permitió identificar 37 genes sobre-expresados asociados al grupo de metástasis y 13 infra-expresados asociados al grupo libre de metástasis.

En un estudio que empleó redes Bayesianas (Gevaert *et al*., 2006) se analizaron datos clínicos (historial del paciente, resultados de análisis de laboratorio y parámetros de ultrasonido) junto a datos de expresión génica obtenidos mediante micro-arreglos que superaron el rendimiento de sistemas de predicción de cáncer que utilizaban sólo datos de expresión. Se usaron datos del “Integrated Tumor Transcriptome Array and Clinical data Analysis Database” para crear un conjunto de datos de entrenamiento proveniente de 78 pacientes, de ellos 34 con diagnóstico de recurrencia de cáncer dentro de los 5 años luego del diagnóstico y 44 libres de la enfermedad en el mismo período. Los datos de prueba se formaron de 19 pacientes, 12 con recurrencia y 7 sin la enfermedad. Se crearon entonces tres estructuras. La primera (integrada) incorporó datos clínicos y de expresión, creando

una red que contenía interrelaciones entre las características de los dos tipos. En la segunda (decisión) se generaron dos redes separadas, una para los datos de expresión y otra para los datos clínicos, cuyas predicciones se fusionaron combinando las dos distribuciones de probabilidad mediante un sistema de coeficientes o pesos. Finalmente, la tercera red (parcial) se formó a partir del primer paso de construcción de una red Bayesiana, generando inicialmente y por separado sólo los dos grafos de dependencia para cada conjunto de datos, y luego uniéndolos por el nodo de la clase, o sea el único nodo en común; seguidamente se generaron las distribuciones locales de probabilidad. Se comparó el rendimiento de cinco métodos: las tres redes descritas, predicciones considerando sólo datos clínicos y predicciones con sólo datos de expresión. Los modelos sobre los datos por separado tuvieron mejor rendimiento que la red integrada, pero inferior a las redes de decisión y parcial, esta última fue la superior. Esto mostró la conveniencia de considerar datos clínicos junto a datos de expresión en los sistemas de diagnóstico de cáncer, pero restringiendo la forma en que las variables se correlacionaban. Este tipo de conclusiones y resultados de análisis más profundos sobre el comportamiento de los conjuntos de datos pueden ser obtenidos por la capacidad descriptiva y gráfica de las redes Bayesianas.

Otros casos en los que ha resultado útil el aprendizaje automático son el entrenamiento algoritmos para determinar multiresistencia a drogas (Vargas *et al.,* 2018), el reconocimiento de patrones de espectroscopía de absorción para la clasificación de enfermedades (Garro, 2016), la búsqueda de reguladores de la mitosis (Wurzenberger *et al*., 2012) y, el control de respuestas al estrés celular (Wippich *et al*., 2013). Además, este enfoque ha sido muy útil para identificar factores relacionados con la biogénesis de ribosomas (Wild *et al*., 2010) y factores celulares del huésped relacionados a infecciones por virus (Mercer *et al.*, 2012). También ha resultado valioso para estudiar la heterogeneidad de la respuesta celular a diversas drogas (Loo *et al.*, 2009), construir perfiles de interacción genética (Hom *et al.*, 2011), así como para el estudio de los estados de la progresión mitótica (Zhong *et al.*, 2012) y el diseño de plataformas

biocomputacionales para la representación y análisis de la respuesta a la quimioterapia (Coto *et al*., 2016).

2. PROBLEMA

2.1. Justificación

El diseño de redes de interacción de miARNs y factores de transcripción que regulan la expresión génica es coherente con el esfuerzo científico para la caracterización de circuitos complejos que representan procesos biológicos. En el caso del fenómeno de compensación de dosis génica, estos procesos pueden estar involucrados en proporcionar estabilidad funcional a las células cancerosas a pesar de su inestabilidad genómica. En esta línea de investigación, el LabQT ha desarrollado una serie de instrumentos tecnológicos que permiten modelar ese tipo de redes sobre genes posiblemente compensados. Sin embargo, el análisis de datos genómicos para la selección de esos genes candidatos se ve limitado al utilizar como fuente de datos el NCI-60, ya que contiene pocos casos de referencia para este tipo de estudios: 60 líneas celulares tumorales humanas adquiridas de 9 tipos de cáncer (9 de pulmón, 7 de colon, 6 de mama, 9 de melanoma, 8 de riñón, 2 de próstata, 7 de ovario, 6 de sistema nervioso central y 6 de leucemia).

El TCGA dispone de un mayor volumen y complejidad de datos (66 tipos de tumores, ~84.000 casos y ~618.000 archivos de datos). La principal ventaja, respecto a iniciativas como el NCI-60, es que estos datos provienen de muestras tumorales primarias humanas, las cuales están asociadas con datos de sobrevida -fechas de diagnóstico, seguimiento y defunción- por lo que podrán ser correlacionados con los patrones de compensación. Esto no es posible con el NCI-60 porque los datos provienen de células cultivadas *in-vitro* que no tienen referencia hacia información de sobrevida.

El aumento de la complejidad de los datos para el estudio de la compensación de dosis génica utilizando TCGA implica un consecuente aumento en la complejidad de la definición de un gen compensado. Para los datos de NCI60 se definió un gen compensado como aquel con una alta variación en su número de copias y una baja tolerancia a la variación en su expresión. Este criterio se vuelve insuficiente ante un mayor número de

muestras, ya que los niveles de expresión de un gen no dependen sólo del número de copias sino también de las regulaciones de muchos otros factores en la red. Ala fecha, no se han definido en la literatura umbrales que permitan clasificar *a priori* si un gen está compensado. En el presente trabajo se propone una estrategia para aproximar los umbrales de compensación mediante criterios de clasificación o patrones del comportamiento de la compensación que emerjan de los propios datos genómicos, al buscar correlaciones entre las magnitudes del material genético y expresión de miles de genes en miles de pacientes.

El volumen de datos manejados por el TCGA hace que la identificación, extracción y organización de los archivos de miles de pacientes, clasificándolos por conveniencia a nivel de la clase de tumor, tipo de contenido, tejido y bioespecimen, sean tareas que no puedan llevarse a cabo manualmente y de manera práctica. De igual forma sucede con el proceso para la creación, partir de dichos archivos, de conjuntos de datos sobre los cuáles aplicar técnicas analíticas para la búsqueda de patrones de compensación.

La identificación de genes candidatos, utilizando patrones de compensación obtenidos a partir de la información genómica del TCGA, potenciará la utilidad de la tecnología de modelaje de redes de regulación desarrollada por el LabQT. La presente estrategia permitirá la construcción de modelos matemáticos que representen interacciones entre miARNs y factores de transcripción -obtenidos a partir de una topología de red-, que puedan ser ajustados contra miles de datos de tumores primarios de pacientes registrados en TCGA.

Bajo esta perspectiva se hace necesaria la implementación de una plataforma computacional que facilite la caracterización del fenómeno de compensación de dosis génica y su relación con la sobrevida de los pacientes. Para ello, dicha plataforma permitirá la integración de datos genómicos del TCGA entre sí y con otras fuentes de información biológica, además, posibilitará la creación de conjuntos de datos *ad-hoc* y la búsqueda de patrones de compensación que conduzcan a la identificación de genes candidatos. Estos genes se utilizarán para la construcción de modelos matemáticos de regulación génica

ajustados por miles de datos de pacientes clasificados según el efecto del fenómeno de compensación en la sobrevida de estos pacientes.

2.2. Hipótesis

La implementación de una plataforma computacional para el estudio de datos genómicos de cáncer humanos del TCGA permitirá una caracterización ampliada del fenómeno de compensación de dosis génica y su relación con la sobrevida de pacientes con cáncer.

2.3. Objetivo general

Implementar una plataforma computacional para el estudio de datos genómicos de cáncer humano del TCGA para la caracterización del fenómeno de compensación de dosis génica y su relación con la sobrevida de pacientes de cáncer.

2.4. Objetivos específicos

2.4.1. Crear módulos de software que interactúen con el GDCDP para identificar, extraer y organizar datos sobre cantidad de material genético, expresión génica, expresión de miARN y sobrevivencia de pacientes de diferentes tipos de cáncer.

2.4.2. Crear módulos de software que implementen procesos de reconocimiento de patrones y técnicas analíticas (entre ellas aprendizaje automático) sobre los datos extraídos del TCGA para identificar patrones de compensación de dosis génica relacionados con la sobrevida de los pacientes.

2.4.3. Ajustar modelos matemáticos de redes de interacción de miARNs y factores de transcripción para modelar la regulación de la expresión de genes candidatos, obtenidos usando los patrones de compensación de dosis génica.

2.4.4 Analizar el modelo de regulación de la compensación de dosis génica para identificar posibles puntos de control con potencial terapéutico.

3. METODOLOGIA

3.1. Adquisición de archivos de datos complementarios

Se dispuso de archivos provistos por Ensembl con datos de coordenadas cromosómicas y nombres de especies registradas en el genoma humano, así como identificadores y equivalencias entre ellos, como por ejemplo aquellas entre los transcritos y miRBase.

3.2. Elaboración de módulos de interacción con GDCDP

Se utilizó Python como lenguaje de programación para escribir los módulos que recopilaron, extrajeron y organizaron los metadatos sobre:

- Pacientes, tal como identificadores, el tipo de cáncer y fechas de diagnóstico, seguimiento y defunción.

- Archivos de datos, como por ejemplo identificadores, el tipo de contenido (variación en el número de copias, expresión génica y expresión de miARN), referencias al bioespecimen (muestra, porción, analito y alícuota) y al tipo de tejido (tumoral o no tumoral).

Estos módulos consideraron la arquitectura del repositorio de archivos, la especificación de los procesos bioinformáticos que generan esos archivos, el diseño del diccionario de datos, el modelo de datos (Figura 2) y el protocolo de comunicación del GDCDP.

Otros módulos se encargaron de la identificación, descarga y organización de los archivos de datos de interés; en primera instancia, se extrajeron los datos de los tipos de cáncer con más volumen de información (cerebro, mama, riñón, pulmón, colon, útero y ovario), lo cual equivale aproximadamente a 35.000 archivos de 6.000 pacientes.

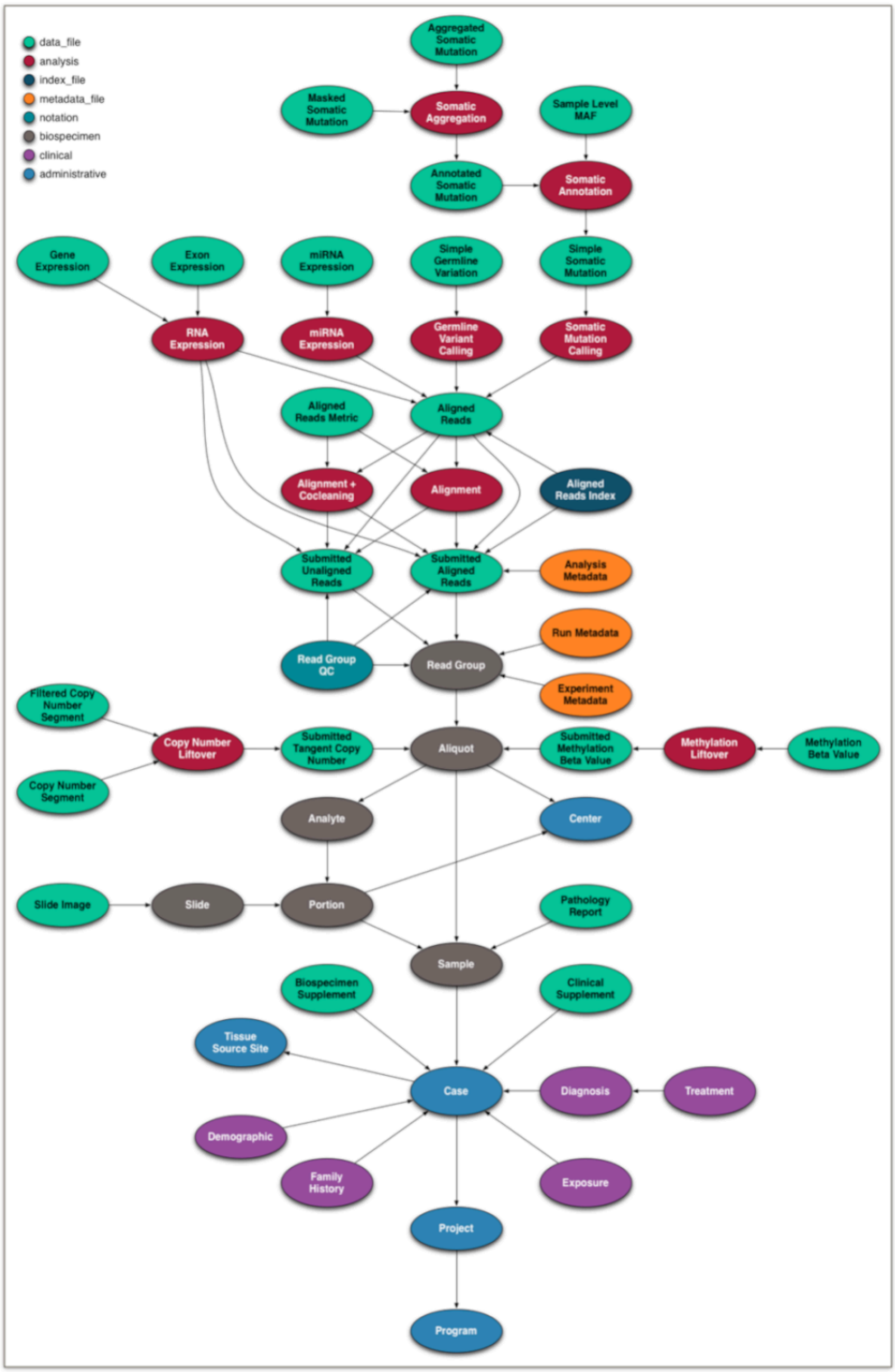


Figura 2. El núcleo del modelo de datos del GDCDP es el caso (paciente), alrededor de él información clínica, de sus bioespecímenes y datos crudos son asociados co datos generados por procesos bioinformáticos (nodos rojos) llevados a cabo dentro del TCGA (https://gdc.cancer.gov/developers/gdc-data-model/gdc-data-model-components).

Los dos tipos de módulos descritos requirieron adaptarse a las especificaciones del API publicadas por el TCGA para interactuar con el GDCDP y extraer de forma masiva registros de metadatos y archivos de datos.

Los valores acerca del número de copias son generados en el TCGA a nivel de segmentos de cromosomas, por lo que un grupo de módulos se ocupó de calcular los valores de número de copias por gen, utilizando las coordenadas de mapa del genoma humano publicado por Ensembl (Zerbino *et al.*, 2018). También controlaron las equivalencias entre distintos tipos de identificadores en los casos de miARN.

3.3. Construcción de conjuntos de datos e implementación de módulos para analítica

Con el fin de apoyar la búsqueda de patrones de compensación de dosis génica relacionados con la sobrevida de los pacientes, se implementaron módulos de programación para crear conjuntos de datos *ad-hoc* que facilitaron la prueba de los criterios propuestos para identificar genes candidatos. Para estas pruebas se construyeron escenarios de análisis, en algunos de los cuales se aplicaron modelos de aprendizaje automático tales como SVM, K-medias, DBSCAN y GMM. Para tal fin se utilizó la biblioteca especializada en aprendizaje automático de Python (scikit-learn).

Para realizar la clasificación mediante SVM se utilizó la clase *SVC.* Los hiper-parámetros de *SVC* se optimizaron no sólo mediante búsqueda exhaustiva sobre valores predeterminados, sino también escaneando un espacio de valores a partir de una distribución específica, para ello se utilizaron las clases *GridSearchCV* y *RandomizedSearchCV*. Se escanearon hiper-parámetros para el tipo de kernel (*lineal, polinomial, radial*), regularización (*C y gamma*), grados (*degree y coef* para el caso de polinomios). Se utilizó *acurracy* como métrica para la selección de la mejor configuración de hiper-parámetros con un enfoque de validación cruzada de 5 grupos (“ 5-fold cross-validation”), lo cual equivalió a 5 rondas de entrenamiento cada una usando 80% de datos

para el aprendizaje y 20% de datos para validación por cada configuración evaluada.

El modelo de SVM con el mejor rendimiento fue utilizado como insumo de un método de Eliminación Recursiva de Características (RFE), esto con el fin de identificar un subconjunto de características altamente discriminante e informativo en el proceso de clasificación. La clase utilizada para implementar el REF fue *RFECV.*

Para la búsqueda de clústeres se utilizaron las clases *Gaussian Mixture*, *KMeans* y *DBSCAN.* Para todos los casos el hiper-parámetro de la cantidad de clústeres por identificar fue asignado en 2.

3.4. Construcción de modelos matemáticos de compensación de dosis génica

La caracterización de patrones de compensación derivó en la identificación de genes candidatos. La regulación de la expresión de esos genes fue modelada mediante el ajuste de la representación matemática de una red de interacciones de miARN´s y factores de transcripción. La red de interacciones se diseñó con la plataforma biocomputacional desarrollada en (Acón *et al*, 2016) y su subsecuente evolución a BioNetUCR (Acón *et al*, 2021) en la que se amplió la cantidad de interacciones de ~65000 a ~450000. Se programaron módulos para la modificación de los modelos matemáticos escritos en SBML -especificación basada en XML para representar modelos en biología de sistemas- (Hucka *et al.*, 2003) que fueron utilizados por COPASI (Hoops *et al.*, 2006) para identificar los miARN´s que controlan la red y que serían los posibles puntos de control con potencial terapéutico.

4. RESULTADOS

4.1. Una plataforma computacional centrada en datos facilita el proceso de caracterización del fenómeno de compensación de dosis génica y su relación con la sobrevida de pacientes.

Se desarrolló una plataforma computacional para recopilar e integrar datos del TCGA y otras fuentes con el fin de transformarlos y analizarlos. Este desarrollo estuvo guiado por la manera en que evolucionó tanto el proceso de identificación de los genes candidatos a encontrarse bajo compensación de dosis, como por el ajuste de los modelos matemáticos de las redes de interacción de factores de transcripción y miARNs. A lo largo de estos procesos se plantearon criterios iniciales para identificar esos genes candidatos y se aplicaron, para confirmar o desechar dichos criterios, diversas técnicas para la búsqueda de patrones entre ellas algoritmos de aprendizaje automático. Por lo tanto, la implementación de la plataforma estuvo centrada en la transformación continua de datos y en su capacidad de evolucionar y crecer conforme se desarrolló el proceso de investigación. Se consideró como requisito base de diseño, dotar a la plataforma de la flexibilidad necesaria para crear diversos escenarios de análisis de datos que permitieran probar y modificar los criterios de compensación y adaptar los ajustes a los modelos matemáticos. De esta forma, el principio de diseño de la plataforma -desde un punto de vista conceptual- corresponde a un grupo de anillos concéntricos que fueron generando una base de datos que creció progresivamente conforme avanzó el proceso de investigación. Así, los datos de un anillo interior estuvieron disponibles para todos los exteriores a él y todo el flujo de datos inició de un anillo central (A0) que creó una base de datos de propósito general (Figura 3).

El TCGA tiene 10 categorías generales de datos, dentro de ellas son de interés para esta línea de trabajo del LabQT las de “Variación en el Número de Copias” y “Perfil de Transcriptoma”, dentro de las cuales hay 103535 y 91737 archivos, respectivamente. Ambas categorías se subdividen en tipos de datos; la primera se distribuye en 5 tipos, uno de ellos es el “Copy Number Segment”, mientras que en la segunda -que se desglosa en 6

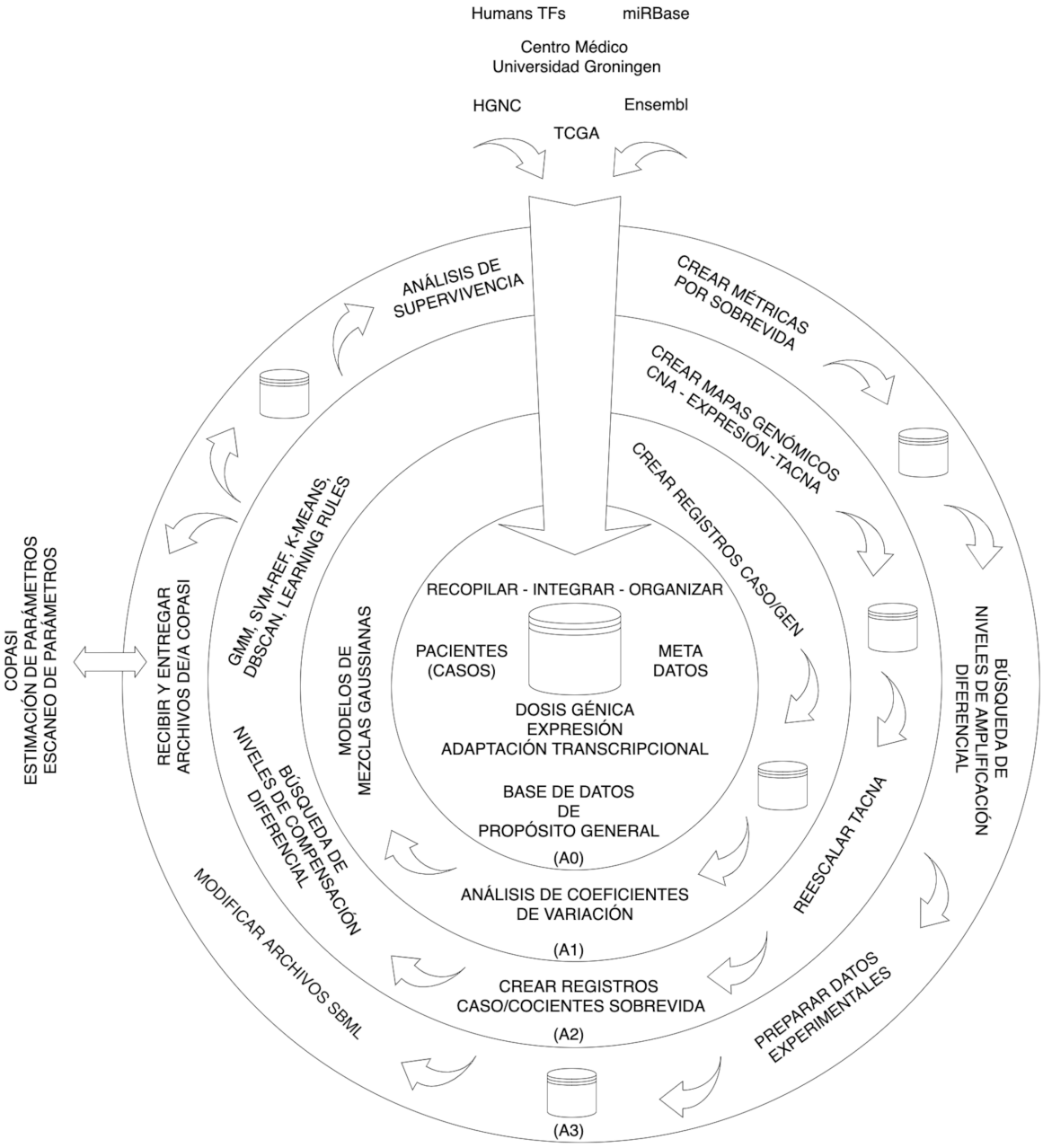


Figura 3. Diseño conceptual de la plataforma computacional que corresponde a un patrón expandible de anillos concéntricos que generan una base de datos que crece progresivamente conforma avanza el proceso de investigación, los datos de un anillo están disponibles para todos los anillos exteriores. El flujo inicia desde una base de datos de propósito general ubicada en el anillo 0 (A0). **Anillo A1:** Creación de conjuntos de datos y escenarios para aplicar el enfoque de coeficientes de variación. **Anillo A2:** Creación de mapas genómicos para reescalar los datos de los perfiles TACNA. Creación de un cociente de compensación para buscar comportamiento diferencial de acuerdo a la sobrevida, aplicando para ello modelos de aprendizaje automático. **Anillo A3:** Creación de nuevas métricas que permitieron identificar genes con niveles de amplificación diferencial de acuerdo a la sobrevida. Interacción con COPASI a nivel de datos. Análisis de supervivencia de pacientes pertenecientes a grupos con distintos niveles de amplificación génica y capacidad de compensación.

tipos- se encuentran el “Gene Expression Quantification” y “miARN Expression Quantification”. Estos tres tipos suman en total 95029 archivos que son la materia prima para la validación y ajuste de los criterios del fenómeno de compensación de dosis génica. Estos archivos tienen estructuras distintas, los de expresión de genes cuantifican la magnitud de los genes usando identificadores “Ensembl”, los de expresión de miARN utilizan identificadores de “mirRBase “ y los de número de copias vinculan sus valores a las coordenadas de segmentos cromosómicos y a los identificadores de las alícuotas de las cuales se obtuvieron.

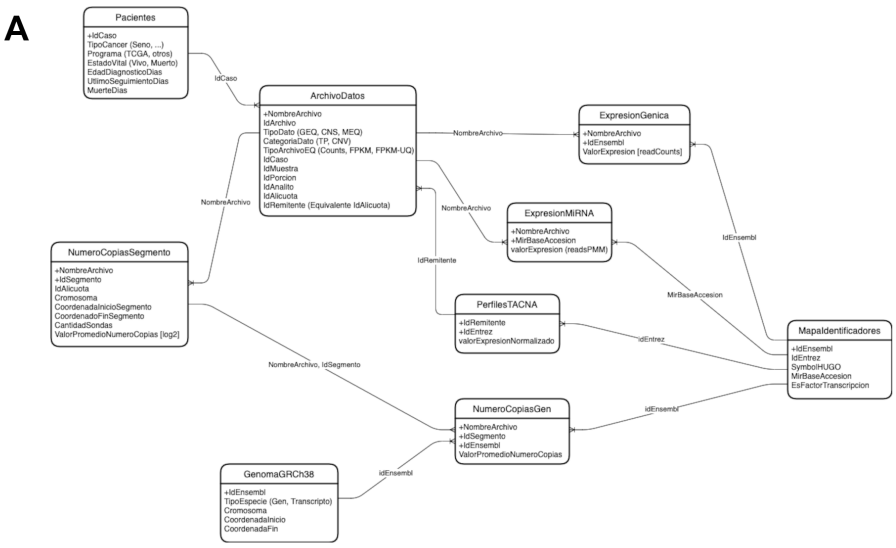
El anillo central recopiló del TCGA cerca de 35000 archivos de datos sobre número de copias por segmento cromosómico, expresión génica y expresión de miARN, de aproximadamente 6000 pacientes de cáncer de cerebro, mama, riñón, pulmón, colon, útero y ovario; por ser estos los tipos de cáncer con más datos. De igual forma extrajo los metadatos de esos archivos y de los pacientes asociados. Los metadatos de los archivos permitieron relacionar estos a los pacientes, identificar el tipo de datos que contenían, el tipo de cáncer, el tipo de tejido (tumoral o no tumoral) y la codificación de indicadores de la muestra, porción, analito y alícuota (bioespecimen) de la que se obtuvieron. La información de los pacientes incluyó el estado vital y las fechas del diagnóstico, último seguimiento y defunción. Este anillo permitió la inclusión de archivos complementarios provenientes de diversos proyectos y organizaciones. Se obtuvieron los perfiles de adaptación transcripcional (TACNA) del Centro Médico de la Universidad de Groningen, las anotaciones del genoma humano (GRCh38) de Ensembl (Zerbino *et al.*, 2018), un catálogo de factores de transcripción procedente de “The Human Transcription Factors” (Lambert *et al.*, 2018). Desde miRBase (Griffiths-Jones *et al.*, 2008) se descargó un catálogo de miARNs y listas de conversiones de sus identificadores con los de transcriptos de Ensembl y además otras listas, descargadas de “Hugo Gene Nomenclature Committe (HGNC)” (Tweedie et al., 2021) con conversiones entre los símbolos HUGO y los identificadores de genes Ensembl y miRBase. Utilizando los archivos con datos sobre el número de copias por segmento cromosómico y las coordenadas cromosómicas del genoma

humano se generaron nuevos archivos con valores a nivel de gen. Empleando las listas de conversiones y la información del bioespecimen se acoplaron los datos propios de TCGA con los perfiles de adaptación transcripcional. Luego, todos los datos de número de copias, expresión de genes, expresión de miARN y adaptación transcripcional se organizaron en una base de datos de propósito general que integró además los datos de diagnóstico de los pacientes (Figura 4A).

Un conjunto de módulos de software escritos en el lenguaje de programación Python tanto para la implementación de las actividades de recopilación, integración y transformación inicial de datos de TCGA como también para sus posteriores adaptaciones, que serán descritas en los siguientes capítulos, forman parte de la plataforma computacional desarrollada en este trabajo.

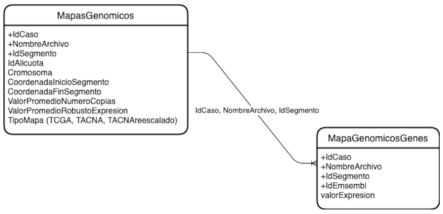
4.2. Análisis sobre datos de cáncer de mama de TCGA confirman la presencia del fenómeno de compensación de dosis génica

Para identificar genes bajo compensación de dosis se propuso caracterizar el comportamiento de los genes candidatos utilizando los datos sobre muestras de cáncer de mama del TCGA. Además de ser el tipo de cáncer que tiene la mayor cantidad de datos, es para el cual el LabQT cuenta con modelos experimentales y líneas celulares estables con varios sensores en caso de una futura validación experimental. Se desarrolló dentro de la plataforma un anillo en donde los datos sobre el número de copias por gen y sus respectivos valores de expresión se adecuaron en un conjunto de registros a nivel de caso y gen (Figura 3-anillo A1 y Figura 4B). En este mismo anillo se crearon escenarios de análisis para una primera exploración sobre los datos que utilizó el enfoque de Acón M *et al*., 2016, sobre los datos de NCI60, en el cual un gen candidato a compensación se definió como aquel con alta variación en el número de copias pero baja tolerancia al cambio en su expresión. Inicialmente se calculó la desviación estándar, tanto para el número de copias como para su expresión, sin embargo, debido a los amplios valores de dispersión se trabajó con los coeficientes de variación del número de copias (CV-ADN) y de la expresión génica



B

	Número de copias TCGA			Expresión TCGA		
Caso 1	Gen 1	...	Gen m	Gen 1	...	Gen m
...						
Caso n						



D

	Grupo Alta Sobrevida			Grupo Baja Sobrevida		
Gen 1	CocientePromedio	%CasosAmplificados	%CasosCompensados	CocientePromedio	%CasosAmplificados	%CasosCompensados
...						
Gen m						

Figura 4. A. Diseño conceptual de la base de datos de propósito general creada en el anillo 0 (A0). **B.**Conjunto de datos caso/gen con los valores del número de copias y expresión TCGA **C.**Diseño de las estructuras que albergan los mapas genómicos basados en los vecindarios de los archivos de número de copias por segmento del TCGA. **D.**Conjunto de datos con métricas por gen, obtenidas de la población de casos

normalizada con respecto al nivel de expresión de los genes diploides (CV-ARN). Además, se integró un tercer parámetro para representar la correlación entre el CV-ADN y el CV-ARN. Se calculó para cada gen el z-score de la expresión génica y luego el coeficiente de correlación de Pearson. Un conjunto de datos con los tres parámetros por gen fue ingresado a un Modelo de Mezclas Gaussianas (GMM) para identificar y seleccionar grupos de genes con una alta variación en el número de copias, baja variación en la expresión y baja correlación entre ellos. Este enfoque identificó 5 grupos de genes (Figura 5B), uno de estos grupos contenía un subconjunto de genes con alta variación en el número de copias y baja variación en la expresión (grupo verde Figura 5A). Dado que este grupo todavía contenía una gran cantidad de otros genes, se realizó un GMM anidado que reveló la presencia de un clúster de genes con alta variación en el número de copias, baja variación en la expresión génica y baja correlación entre ellos (Figura 5C). Posteriormente, se calculó para cada gen candidato una regresión lineal con base en los datos de expresión y número de copias. Una pendiente inferior a 0.5 significa que un gen en particular tiene una expresión más baja en comparación a lo esperado por su amplificación en el números de copias (Figura 5D). Se identificaron un total de 36 genes con pendientes inferiores a 0.5. Estos resultados demuestran la presencia de genes bajo el criterio de compensación de dosis de Acón *et al*., 2016, lo que sugiere un mecanismo activo de compensación de dosis de genes.

La anotación funcional de los 36 genes candidatos reveló que la mayoría de los genes están ubicados en los cromosomas 8, 11 y 17 (Figura 6A). Este resultado es similar a los hallazgos de Acón M *et al.*, 2016, los cuales mostraron que estos cromosomas tenían un número significativo de candidatos que incluían miARN y genes codificadores de proteínas. La anotación estructural de productos génicos candidatos bajo compensación de dosis identificó diversos biotipos genéticos. Usando la anotación GENCODE actual se hallaron ARN codificantes y no codificantes (ARNnc) de proteínas con distribución en diferentes cromosomas (Figura 6B). De los ARNnc, los ARN largos no codificantes (ARNlnc) fueron los más presentes, con transcripciones intrónicas antisentido y sentido. La gran presencia de diversos biotipos de ARNlnc representa un hallazgo interesante, dada su

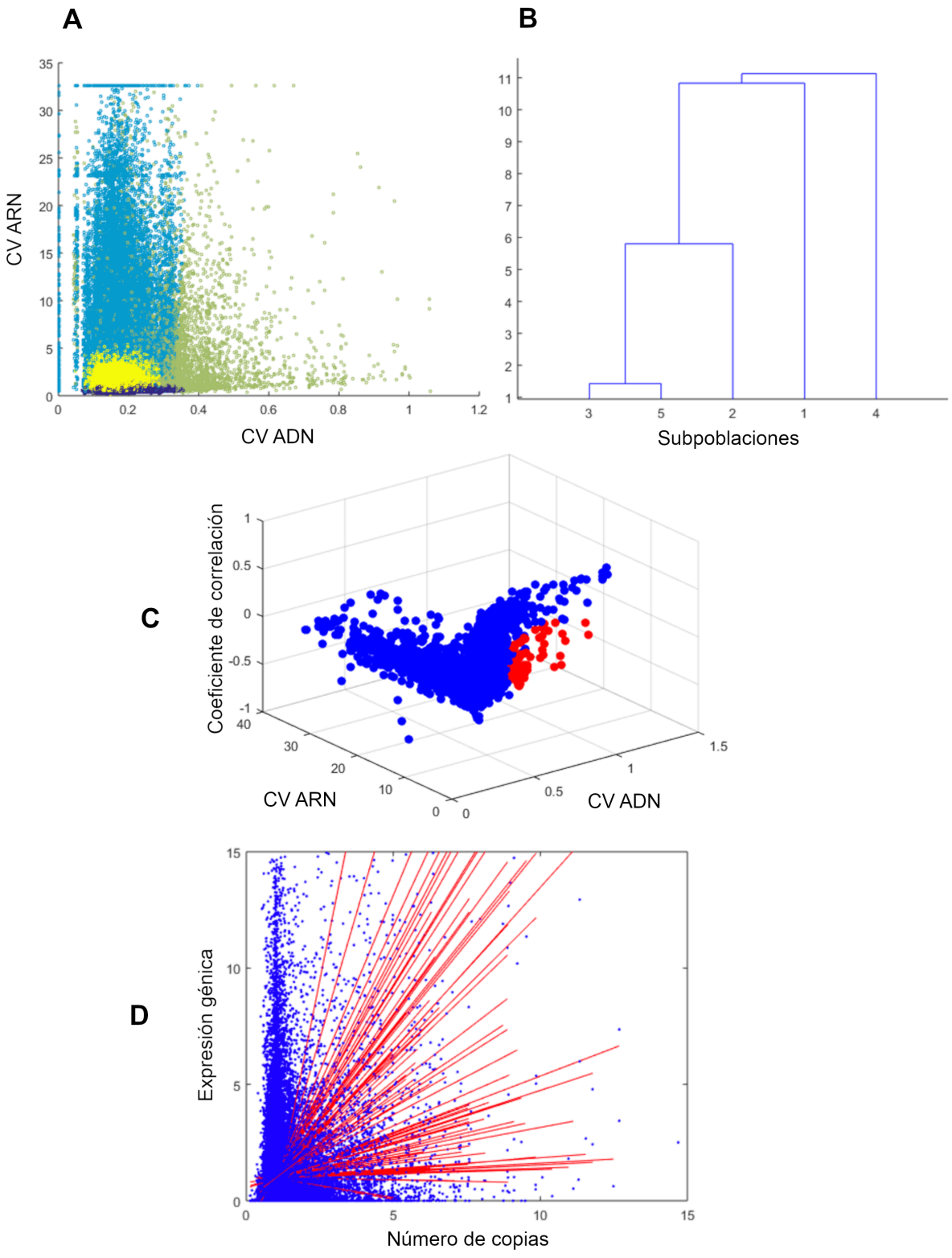


Figura 5. Resultados del enfoque de Acón et al. al comparar la variación entre el número de copias y la expresión de datos de cáncer de seno del TCGA. **A-B.**Grupos de genes identificados por GMM. **C.**En rojo los genes con alta variación en el número de copias, baja variación en la expresión y baja correlación entre ellos. **D.**Regresiones lineales sobre la expresión normalizada, se observa una gran variación a través de los genes candidatos.

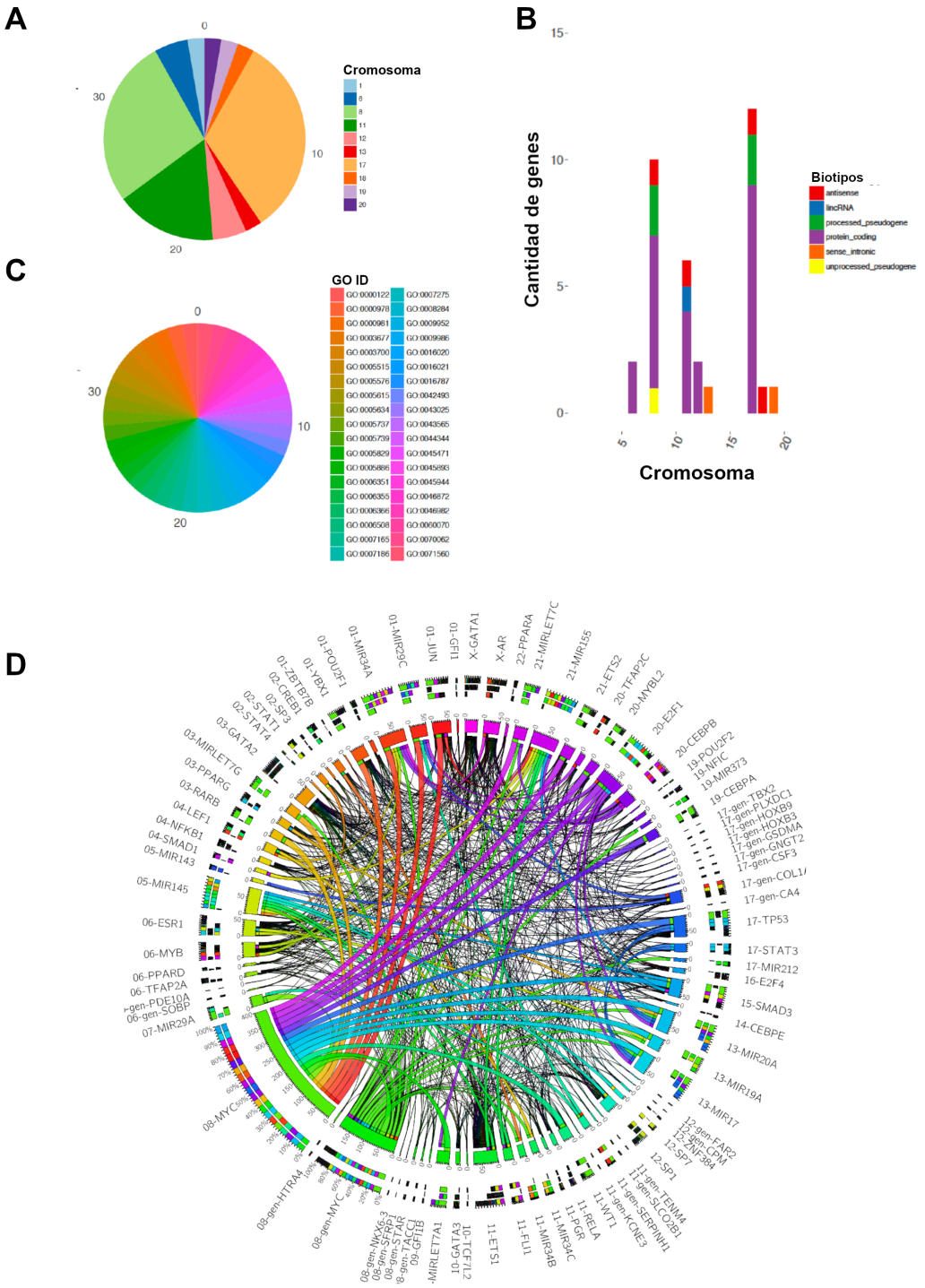


Figura 6. Distribución de la anotación funcional de los genes candidatos identificados por GMM. **A.** Distribución de los biotipos de genes para cada cromosoma, basado en GENCODE. **B.** Distribución de los genes candidatos en cada cromosoma. **C.** Términos GO identificados en los genes que codifican proteínas. **D.** Conectividad de MYC

asociación con múltiples procesos como: regulación de la expresión génica, interacción con complejos remodeladores de cromatina e interacción con complejos proteicos (Oviedo, 2018). La identificación de ARN antisentido sugiere la participación de los ARNlnc en eventos relacionados con la biología del cáncer y la compensación de la dosis de genes. También se identificaron pseudogenes, que pueden adquirirse somáticamente durante el desarrollo del cáncer, debido a la activación de retrotransposones (Coke, 2014). Aunque la evidencia ha asociado los ARNnc y el cáncer, la relación entre la dosis de genes y los ARNnc sigue sin estar clara y requiere mayor estudio. El término de ontología genética más representado fue GO: 0005515 (Figura 6C), que corresponde a la función molecular que caracteriza interacciones selectivas y no covalentes con proteínas o complejos proteicos. La presencia de productos génicos asociados a proteína-proteína y proteína ribonucleoproteína podría representar a elementos que interactúan no solo con otras proteínas sino con transcripciones no codificantes, que se sabe regulan la expresión génica. El segundo término más representado fue GO: 0003677 el cual se refiere a cualquier interacción selectiva y no covalente con el ADN. Estos son productos genéticos que forman interacciones con el ADN que son diferentes de los factores de transcripción (GO: 0003700). Finalmente, fue de interés la presencia del término GO: 000098, correspondiente a la función molecular de los productos génicos que interactúan de forma selectiva y no covalente con una secuencia de ADN específica, para modular la transcripción por la ARN polimerasa II. La presencia de enriquecimiento de la función del factor de transcripción RNA-pol II (13% de los objetivos) en la lista de candidatos muestra la posible asociación entre los elementos reguladores y la compensación de la dosis de genes.

Como seguimiento a la hipótesis de que la compensación de la dosis de genes puede estar mediada por las propiedades emergentes de una compleja red de interacciones de miARN y factores de transcripción, se indagaron las interacciones reportadas de los genes identificados bajo compensación de dosis en el cáncer de mama para construir dicha red. La red resultante se muestra en la figura 6D y destaca la alta conectividad de la red y sugiere e

indica que el oncogen MYC tiene un papel central en la regulación de esta red de compensación de dosis.

En conjunto, estos resultados indican que la compensación de la dosis génica está presente en tumores de pacientes con cáncer de mama. Además, el grupo de genes identificados bajo compensación de dosis tiene una expresión más baja de lo que podría esperarse en casos con amplificación de genes. Lo anterior implica que su expresión regulada juega un papel importante en las funciones de supervivencia del cáncer tal y como sugiere su análisis de enriquecimiento funcional. Finalmente, estos análisis apoyan la hipótesis de que la compensación de la dosis de genes está presente en el cáncer y los genes involucrados tienen una regulación común por una red a gran escala de interacciones entre miARN y factores de transcripción. Sin embargo, es difícil establecer un límite para clasificar si un gen está compensado o no, y la utilización de regresión lineal para calcular las posibles pendientes está sujeta a mucha variación, al supuesto de un comportamiento lineal de esa variable y a niveles diferentes de ajuste entre los genes, haciéndolo muy sensible a "outliers" y por lo tanto a error.

4.3. Análisis del grupo de pacientes de cáncer de mama del TCGA no evidenció niveles de compensación diferencial en relación con la sobrevida de los pacientes

Para identificar genes con niveles de compensación diferencial en pacientes con distintos tiempos de sobrevida se requirió diseñar una medida más robusta para representar el grado de compensación génica. Luego esta medida fue utilizada por un modelo de análisis de datos para evaluar que tan bien el grado de compensación de un gen contribuiría a la separación de un conjunto de pacientes en distintos grupos de sobrevida.

El gráfico de regresiones lineales sobre la expresión normalizada contra el número de copias (Figura 5D) reveló una gran variación a través de los genes candidatos, lo que indica que en un conjunto grande de datos de expresión génica que contiene ruido, el criterio de comparación de coeficientes de variación, usado para identificar genes que podrían

encontrarse bajo compensación de dosis, no funciona adecuadamente. Ese ruido es producto de la influencia de factores experimentales y no genéticos en los enfoques de genómica genética utilizados para medir los efectos que ejercen las alteraciones en el número de copias (CNA) sobre la expresión de genes, y -probablemente- por la heterogeneidad en la expresión génica de una población. Para obtener esa medida robusta del grado de compensación génica se utilizaron los perfiles TACNA. Módulos del anillo 2 de la plataforma (Figura 3-anillo A2) fueron utilizados para adecuar la escala de dichos perfiles a la del CNA. Basándose en los archivos de número de copias por segmento de cada paciente (caso) los módulos definieron vecindarios de genes de acuerdo con las coordenadas cromosómicas de esos segmentos. En estos vecindarios los genes comparten el valor del número de copias del segmento, pero tienen su propio valor de expresión. Se construyeron vecindarios tanto para la expresión original de TCGA como para los perfiles TACNA. A partir de estos vecindarios, la base de datos de propósito general se expandió y se crearon mapas genómicos por paciente (Figura 4C) que fueron utilizados para calcular el promedio robusto de los valores tanto de expresión TCGA como de TACNA, de acuerdo con el segmento CNA en que se encontraba el gen correspondiente en la muestra dada, y se compararon con los valores de CNA (Figura 7A Panel 1 y 2, basada en una muestra de cáncer de mama). Luego, se reescaló el perfil de TACNA para ajustarlo a la escala de los datos del número de copias. Para ello el valor del perfil de cada gen se multiplicó por el cociente del respectivo número de copias del segmento dividido por el promedio robusto de TACNA de ese segmento (Figura 7A Panel 3). De esta forma el perfil reescalado de TACNA se pudo comparar directamente con los valores de CNA. Finalmente, el cociente TACNA/CNA se utilizó como una medida más robusta de compensación (Figura 7A Panel 4) donde un cociente por debajo de 1 indica compensación de dosis génica, alrededor de 1 indica ausencia de compensación y por arriba de 1 amplificación de dosis.

Mediante módulos del anillo A2, se dividió el conjunto de pacientes en cuartiles de acuerdo con el tiempo de sobrevida y se nombró a los del primer cuartil como el grupo de baja sobrevida y a los del cuarto cuartil como el de alta sobrevida (212 casos en cada grupo).

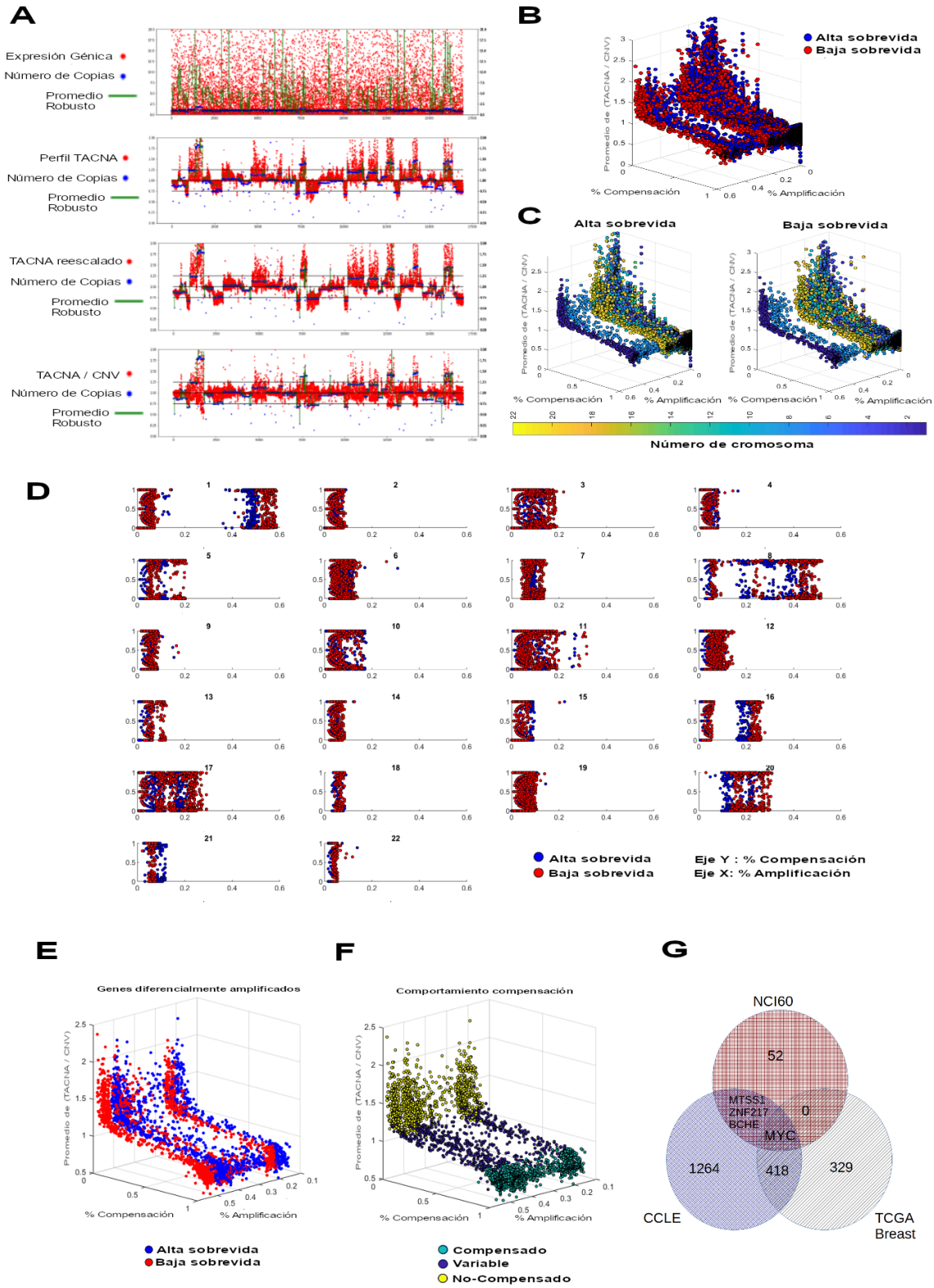

A
Expresión Génica
Número de Copias
Promedio Robusto
Perfil TACNA
Número de Copias
Promedio Robusto
TACNA reescalado
Número de Copias
Promedio Robusto
TACNA / CNV
Número de Copias
Promedio Robusto
B
Alta sobrevida
Baja sobrevida
Promedio de (TACNA / CNV)
% Compensación
% Amplificación
C
Alta sobrevida
Baja sobrevida
Promedio de (TACNA / CNV)
% Compensación
% Amplificación
Número de cromosoma
D
Alta sobrevida
Baja sobrevida
Eje Y : % Compensación
Eje X: % Amplificación
E
Genes diferencialmente amplificados
Promedio de (TACNA / CNV)
% Compensación
% Amplificación
Alta sobrevida
Baja sobrevida
F
Comportamiento compensación
Promedio de (TACNA / CNV)
% Compensación
% Amplificación
Compensado
Variable
No-Compensado
G
NCI60
52
MTSS1
ZNF217
BCHE
0
MYC
1264
418
329
CCLE
TCGA Breast

Figura 7

Luego, se formó un conjunto de datos con los niveles de compensación de cada gen (16945 genes) para cada caso seleccionado (424 casos). A continuación se calculó CV-ADN, CV-ARN y la correlación entre ellos, luego se aplicó GMM para reducir a 196 genes por los mismos 424 casos clasificados en alta y baja sobrevida. Sobre este conjunto de datos se utilizó el método de Eliminación Recursiva de Características basado en Máquina de Soporte Vectorial (SVM-RFE) para identificar un subconjunto de características -niveles de compensación de cada gen- altamente discriminante e informativo en la clasificación de pacientes con respecto a los dos grupos de sobrevida.

SVM-RFE es un método muy utilizado para detectar este tipo de subconjuntos -llamados firmas distintivas- para construir clasificadores muy confiables en el estudio del cáncer (Xu *et al.*, 2012; Guyon *et al.*, 2002; Duan *et al.*, 2005; Zhang *et al.*, 2006). SVM es muy eficiente para la clasificación en espacios de alta dimensión, encontrando el hiperplano que separara de mejor forma las clases usando vectores de soporte, los cuales corresponden a los casos que generan la mayor distancia con respecto al hiperplano, tal y como se muestra en la Figura 1C en un problema de clasificación binaria. Cuando los datos no se pueden separar linealmente en un espacio de baja dimensión estos se asignan a un espacio de alta dimensión mediante una función transformadora. Además, se utilizan penalizaciones para los datos clasificados erróneamente si aún es imposible encontrar un hiperplano incluso en un espacio de muy alta dimensión. El método SVM-RFE es un proceso de selección secuencial hacia atrás. Se comenzó con el conjunto de todas las características y la menos importante para la clasificación se eliminó de forma iterativa de acuerdo con el criterio de clasificación del SVM; el criterio estuvo formado por pesos que multiplicaron cada característica. Esos pesos se usaron como coeficientes de importancia, así las características que estuvieron ponderadas por los valores más grandes se consideraron como las más influyentes en la decisión de clasificación. Por lo tanto, las entradas con los pesos más grandes correspondieron a las características más informativas. En cada iteración del clasificador SVM se midió su rendimiento (precisión) hasta que todas las características

fueron eliminadas. Se identificó un grupo de 196 genes cuyos cocientes de compensación tenían mayor peso en la clasificación de sobrevida.

Sin embargo, la precisión de la clasificación fue en extremo pobre, alrededor del 50%, por lo que no se logró encontrar un comportamiento diferencial entre los niveles de compensación del grupo de genes con respecto al tipo de sobrevida. Con el fin de corroborar este hallazgo se aplicaron varios métodos de aprendizaje no supervisado (K-Means, DBSSCAN y GMM), con el fin de localizar clústeres naturales, pero ninguno logró separar en dos el conjunto de pacientes. Posteriormente, se extendió el conjunto de datos con nuevas características por cada gen, estas complementaron los valores de TACNA y número de copias y se aplicó aprendizaje de reglas para encontrar el conjunto de reglas que mejor clasificara la sobrevida y, por ende, los genes involucrados en ellas. A pesar de esto, los resultados fueron parecidos a los de SVM-REF dado que las reglas encontradas sólo lograron clasificar de forma correcta la mitad de los casos. Dado lo anterior, no se logró evidenciar la existencia de genes con una compensación diferencial entre pacientes en los extremos de alta y baja de sobrevida.

4.4. Reconfiguración de los datos sobre niveles de compensación de dosis génica basados en TACNA posibilita la identificación de genes candidatos en conjuntos con gran volumen de casos.

Considerando que el intento para identificar genes con niveles de compensación de dosis génica diferencial respecto de la sobrevida de pacientes resultó infructuoso, se configuraron -utilizando módulos del anillo A3- dos nuevos conjuntos de datos (uno por cada grupo de sobrevida). Para esto se utilizaron tres métricas: el promedio de los cocientes TACNA/ CNA, el porcentaje de casos amplificados -CNA mayor que 1- y el porcentaje de casos compensados -con cociente menor a 1- (Figura 4D). Se encontraron diferencias de interés en la distribución de la dispersión de las tres métricas entre los grupos de alta y baja sobrevida, en particular para el porcentaje de casos amplificados (Figura 7B). Los datos codificados por color de acuerdo al cromosoma (Figura 7C) indican que estas diferencias se

deben a un aumento en el porcentaje de casos amplificados en genes de ciertos cromosomas en el grupo de baja sobrevida, mientras que las distribuciones del promedio de cocientes y porcentaje de casos compensados permanecen muy similares. A continuación, esta distribución diferenciada de genes entre los dos grupos se exploró para cada cromosoma identificando así un comportamiento bimodal de genes frecuentemente amplificados en los cromosomas 1, 8, 11 y 16, en contraste con otros que mostraron baja frecuencia de amplificación (Figura 7D). Los genes altamente amplificados fueron seleccionados ya que la mayoría de ellos mostraron un comportamiento diferencial entre los dos grupos de sobrevida. Cabe destacar que los casos del grupo de baja sobrevida presentaron mayor frecuencia en la amplificación (Figura 7E, puntos rojos). Finalmente, esos genes se clasificaron en tres categorías de acuerdo con el porcentaje de casos compensados (Figura 7F); se denominó como “Compensado” el gen con más de 90% de casos compensados (todos con cociente TACNA/CNA menor que 1), “No-compensado” el gen con menos de 10% de casos compensados (todos con cociente TACNA/CNA mayor que 1) y “Variable” el gen cuyo porcentaje de casos compensados se ubicó entre 10% y 90% (con un cociente TACNA/CNA alrededor de 1). De este análisis se identificó una lista de 51 factores de transcripción compensados en ambos grupos de sobrevida.

Estos resultados confirman que no hay genes con compensación diferencial en pacientes con distintos tiempos de sobrevida; por el contrario, los genes bajo compensación de dosis en cáncer de mama mantienen ese comportamiento en la mayoría de los casos independientemente de la progresión de la enfermedad. También, los resultados sugieren que estos genes podrían habilitar el avance del cáncer en medio del cambio en el número de copias de sus respectivos cromosomas, lo cual provoca amplificaciones en genes compensados y no compensados. Si lo anterior fuera correcto y la compensación de dosis génica se conserva en los mismos genes a través de los diferentes casos, esos genes amplificados diferencialmente respecto a la sobrevida representarían dianas terapéuticas de mucho interés contra el cáncer aneuploide.

Adicionalmente, se realizó un análisis de enriquecimiento funcional sobre el grupo de 1338 genes frecuentemente amplificados (748 compensados y 590 no compensados); luego sólo sobre los 123 factores de transcripción (51 compensados y 72 no compensados) a nivel de procesos biológicos, funciones moleculares y vías biológicas definidas en "PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System" (Mi *et al.*, 2019) (Figuras 8 y 9). Se encontró que el 87% de todos los genes que tienen que ver con interacción entre especies, 85% de adhesión biológica, 81% de sistema inmune, 81% de desarrollo y 80% de organismo multicelular son compensados. Adicionalmente resulta llamativo la cantidad de genes involucrados en respuesta al estímulo (188) y regulación biológica (379), de los cuales el 65% y 54%, respectivamente, también son compensados. Más del 73% de los factores de transcripción amplificados se asocian a procesos celulares, regulación biológica y metabolismo, en todos ellos los factores compensados representan aproximadamente el 45%. En desarrollo, organismo multicelular y señalización los factores de transcripción compensados abarcan el 88%, 78% y 67%, respectivamente. En lo relacionado con funciones moleculares, casi un 28% de los genes frecuentemente amplificados están vinculados con actividad catalítica y con unión, un 10% con regulación molecular. En esas tres categorías los genes compensados representan en promedio el 50%. Se observó, además, que el 93% de los genes ligados a la actividad transductora molecular son compensados. Por otra parte, el 46% de los factores de transcripción emparentados a la regulación molecular y a la unión son también compensados. Finalmente, se denota que entre el 87% y 95% de los genes que forman parte de la vía de señalización de proteína G heterotrimérica son genes compensados, así como también el 69% de los que están incluidos en la vía de señalización de quimiocinas y citocinas que median la inflamación. Las implicaciones de la compensación de dosis génica sobre estas vías y procesos celulares requiere más investigación incluyendo la identificación de redes complejas de regulación génica que involucren uno o más de estos procesos y vías celulares.

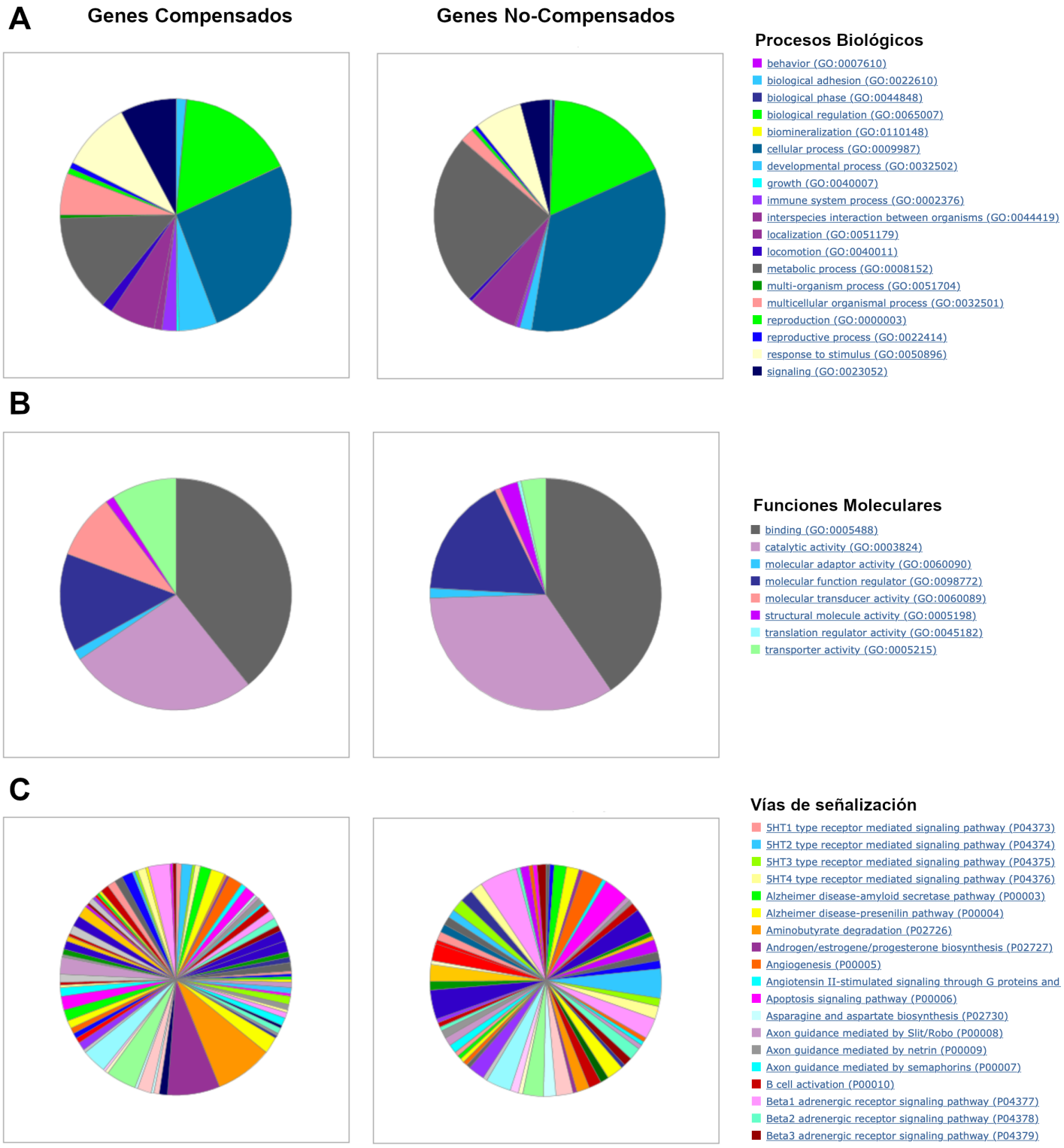


Figura 8. Enriquecimiento funcional sobre genes amplificados (compensados y no-compensados). A. Procesos biológicos. **B.** Funciones moleculares. **C.** Vías de señalización.
Análisis basado en "PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System".

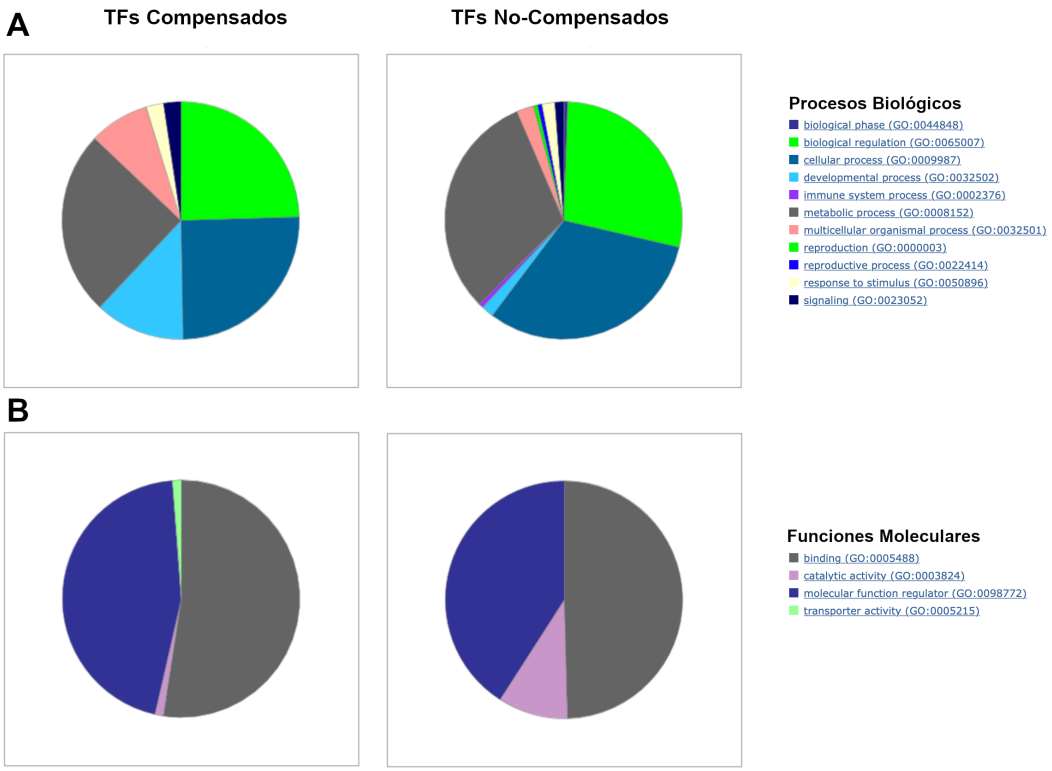


Figura 9. Enriquecimiento funcional sobre factores de transcripción amplificados (compensados y no-compensados). A. Procesos biológicos. **B.** Funciones moleculares.
Análisis basado en "PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System".

4.5. El ajuste de un modelo matemático de las interacciones de una red de micro-ARNs y factores de transcripción permite modelar la compensación de dosis de genes candidatos

Para encontrar propiedades a nivel de sistema de una red de miARNs y factores de transcripción que permitan modelar la compensación de dosis de genes candidatos se crearon y ajustaron modelos matemáticos utilizando datos del TCGA. Se construyeron, entonces, redes de interacciones reguladoras para cada gen candidato (BioNetUCR, LabQT) y se seleccionaron sólo aquellas que presentaran ciclos reguladores de uno, dos y tres nodos, que permitieran propagar y regresar las señales desencadenadas por factores de transcripción. Estos ciclos podrían compensar la expresión génica en respuesta a cambios

en la dosis. De las 51 redes sólo ATF3, RUNX1T1, MYC y POU2F1 presentaron esos tipos de ciclos (Figura 10A y 10B). El oncogen MYC surge como un gen altamente conservado bajo compensación de dosis tanto en TCGA como en estudios paralelos realizados sobre NCI60 y CCLE (Figura 7G), por lo que se profundizará el análisis en los posibles mecanismos involucrados en este fenómeno.

Se crearon desde BioNetUCR los archivos SMBL usados por COPASI para ajustar el modelo de compensación de MYC. Del conjunto de datos de la Figura 4B, se extrajeron los datos experimentales de los 1069 casos de cáncer de mama para las 77 especies presentes en la red de MYC. Con el fin de favorecer la búsqueda de los valores de los parámetros que conducen a la compensación, se incrementaron los pesos de la función objetivo de la expresión de MYC para cada línea experimental. Para tal fin y para evitar la alteración manual del modelo vía la interfaz de aplicación de COPASI, se implementaron módulos (Figura 3, anillo-3) que modificaron automáticamente los archivos SBML. El ajuste del modelo de MYC resultó con una función objetivo muy alta. Se notó una enorme heterogeneidad en los valores de la función objetivo de los casos de pacientes individuales, lo que sugiere de hecho una configuración heterogénea en esta red. Se decidió probar con el modelo mínimo obtenido en Acón *et al.*, 2006 (Figura 10C) presentándose el mismo comportamiento. Dado lo anterior, se utilizó el modelo como herramienta para caracterizar los casos de los pacientes ajustándolo a los datos de cada experimento individual (caso) y se obtuvo un conjunto de parámetros cinéticos que describen las interacciones entre miARNs y factores de transcripción para cada paciente. A continuación, se realizaron exploraciones de tales parámetros individuales para caracterizar aún más los casos en función de su respuesta a los aumentos del número de copias MYC. Para tal fin, se llevaron a cabo simulaciones *in silico* en las que se aumentó el número de copias del gen de 1 a 5 y se observó el aumento en la expresión génica: un aumento sub-lineal denota la capacidad de la configuración de red para lograr la compensación de la dosis. Se observó una gran heterogeneidad en la compensación de dosis de MYC (Figura 11A).

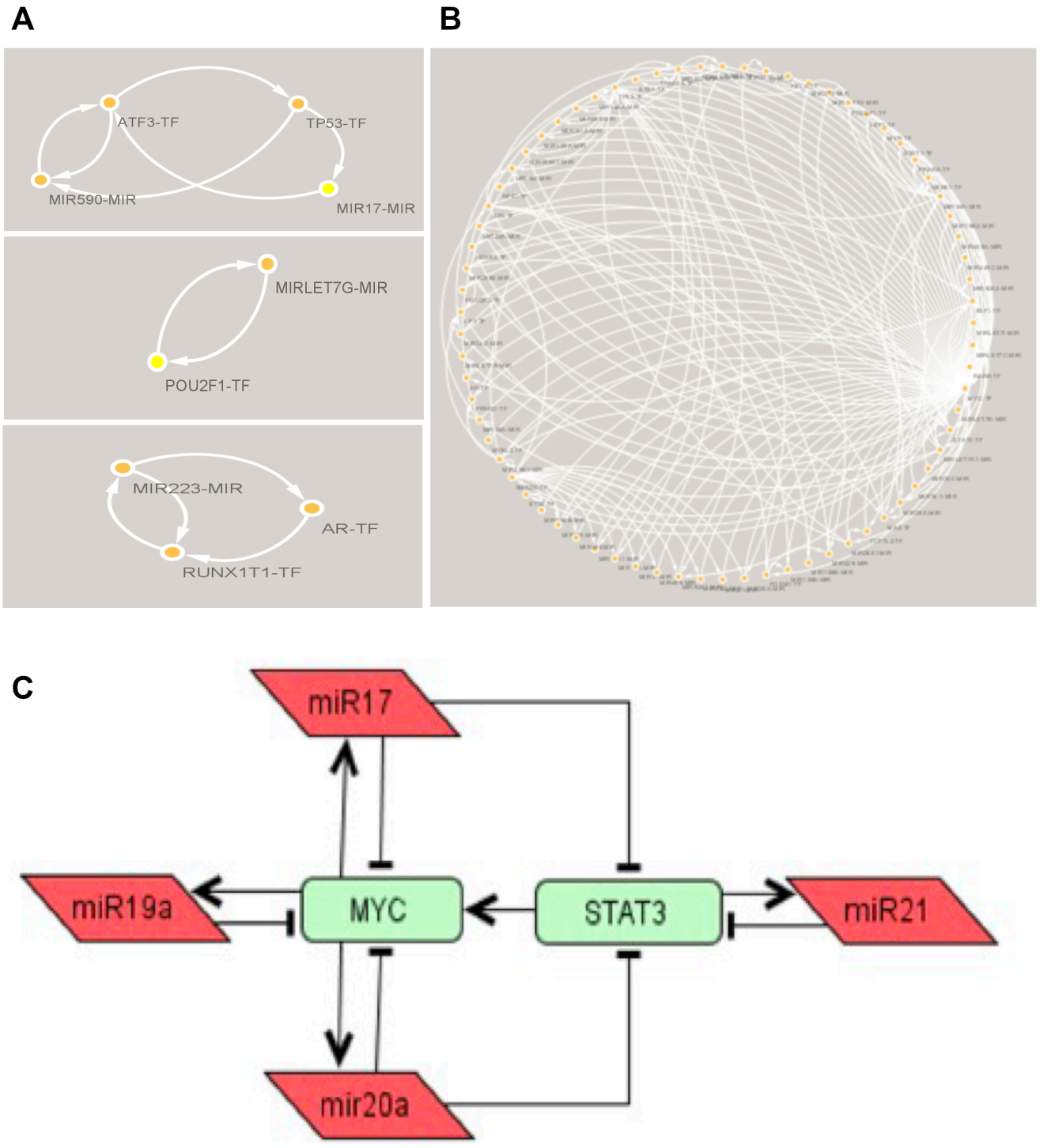


Figura 10. Topologias. A-B. Redes de transcripción. ATF3 (A.Panel superior), POU2F1 (A.Panel central), RUNX1T1 (A.Panel inferior) y MYC (B). **C.** Modelo mínimo obtenido en Acón *et al*., 2006.

Luego, se clasificaron los pacientes según su amplificación génica y la capacidad de compensación de sus modelos ajustados. La amplificación en un gen sugiere que el tumor está progresando en su malignidad y plantea la cuestión de si esta amplificación se compensa o no. Por lo tanto, se trazó el número real de copias de genes frente a la expresión génica simulada en un número de copias de 5 para evaluar la capacidad de compensación de la configuración de la red. Cuanto menor sea la expresión génica simulada, mayor será la capacidad de la configuración de red correspondiente para realizar la compensación. Se observaron muchas amplificaciones para MYC con una separación entre los casos compensados y no compensados que se hizo más evidente con un mayor número de copias. Se logró separar los casos en tres grupos: baja amplificación (CNV menor o igual a 1,5) y para aquellos con mayor amplificación un grupo con baja compensación (Expresión MYC superior a 3,7) y un grupo con alta compensación (Expresión MYC inferior a 3,7) (Figura 11B). Además se trazaron curvas de supervivencia implementando un gráfico de Kaplan-Meier para comparar las probabilidades de supervivencia entre los grupos (Figura 11C). La importancia de sus diferencias se evaluó mediante una regresión de riesgos proporcionales de Cox. No se observó diferencia significativa entre los casos no amplificados y los casos con amplificaciones no compensadas ($p = 0,1$), por el contrario, se observó una disminución significativa en la probabilidad de supervivencia de los pacientes con amplificaciones compensadas en comparación con los no compensados ($p = 0,009$) y en comparación con los casos sin amplificaciones ($p = 0,007$). Este resultado sugiere que la compensación de la dosis de MYC permite la progresión del cáncer hacia la malignidad y de hecho representaría un importante objetivo terapéutico para prevenir esta progresión.

Finalmente, se determinó un espacio multidimensional de parámetros, donde se identifica una región de compensación y otra de no compensación para las posibles combinaciones de parámetros cinéticos de un motivo de regulación. Esto con el propósito de caracterizar aún más la heterogeneidad de la compensación de dosis de MYC en los modelos de pacientes con cáncer de mama. Las dimensiones de este espacio corresponden a los 3 parámetros

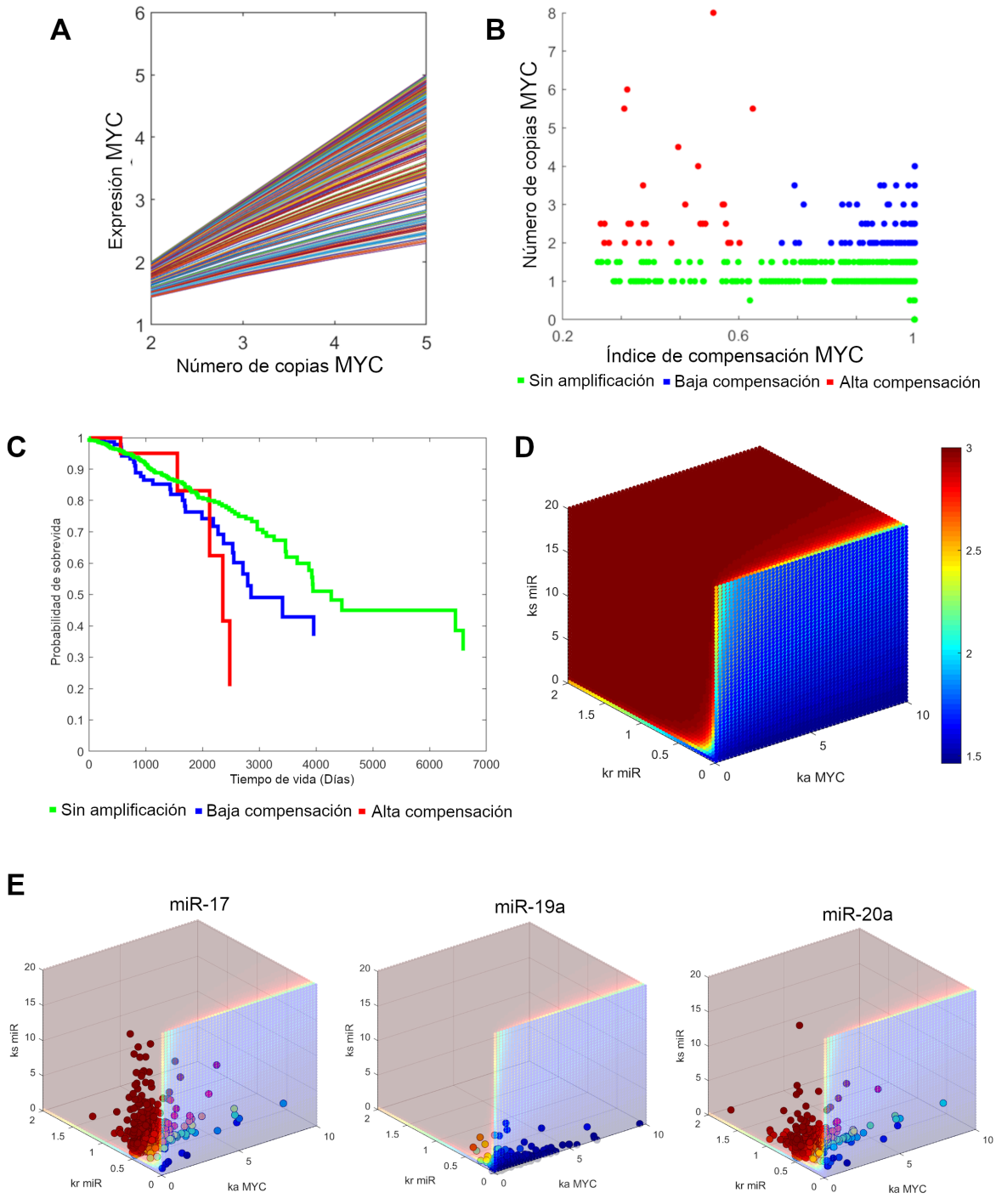


Figura 11.A. Comportamiento de la compensación de MYC a través de cada caso (paciente). **B.**Grupos de casos de acuerdo a la capacidad de compensación de la red. **C.**Análisis de supervivencia para los grupos de la figura B. **D.**Mapa de calor con regiones de compensación de acuerdo a parámetros de interacción entre especies (Zona azul corresponde al espacio de parámetros de dosificación-compensación, la zona roja al espacio que conduce a un aumento lineal de la expresión en función de la dosis génica). **E.**Ubicación para cada motivo de retroalimentación del los microARN del modelo de la figura 7E del valor de compensación de cada paciente. La mayoría de casos de pacientes de cáncer de seno son compensados por el miR-19a, lo cual sugiere que representa un objetivo terapéutico para bloquear la compensación de dosis de MYC en este tipo de cáncer.

cinéticos que mejor separan los niveles de compensación: las tasas de síntesis del miARN (ks), de activación del factor de transcripción (ka) y de represión de miARN (kr). Usando COPASI se realizaron simulaciones en estado de equilibrio para monitorear sistemáticamente el efecto de los 3 parámetros sobre la expresión de MYC. Se utilizaron dos condiciones de número de copias de MYC: con 1 y 3 copias, y se calcularon los valores de MYC en estado de equilibrio para la correspondiente combinación de parámetros. Luego, se normalizaron los datos de expresión en la condición de 3 copias dividiéndolos por la correspondiente expresión con 1 copia. Esto permitió obtener un espacio de compensación, representado por un mapa de calor, en el que la región azul corresponde al espacio de parámetros donde ocurre compensación, mientras que la región roja representa el espacio de parámetros que conducen a un aumento lineal de la expresión génica en función del número de copias (sin compensación) (Figura 11D). Usando los valores de los parámetros ajustados para cada paciente, se utilizó el algoritmo de búsqueda del vecino más cercano para obtener el valor de compensación para cada paciente para ese motivo de retroalimentación mediado por un miARN en particular. Se observó que miR-17 y miR-20a median la compensación de la dosis de MYC solo en una pequeña fracción de los casos, mientras que la mayoría de los casos de pacientes con cáncer de mama son compensados por el motivo de retroalimentación regulado por miR-19 (Figura 11E). Esto sugiere que la inhibición de miR-19 representaría un objetivo terapéutico potencial para bloquear la compensación de la dosis de MYC en el cáncer de mama.

En conjunto, estos resultados basados en datos de pacientes indican que la compensación de la dosis de MYC es mediada por una propiedad emergente de retroalimentación y bucles de retroalimentación. Las interacciones que forman los motivos de esta red tienen un conjunto específico de parámetros cinéticos que permiten la compensación. Se validaron experimentalmente los circuitos de compensación para MYC y se determinó que este mecanismo puede bloquearse agotando el miARN clave, afectando en mayor medida a aquellas células con mayor número de copias de esos genes compensados. El estudio de la compensación de la dosis de MYC en pacientes con cáncer de mama mostró

heterogeneidad con respecto a qué motivos reguladores median esta regulación. También sugiere que la compensación de la dosis de MYC podría conducir a una menor supervivencia del paciente, destacando el potencial de este tipo de intervenciones para atacar el cáncer aneuploide y prevenir la progresión del cáncer.

5. DISCUSIÓN

Actualmente, el tema de la regulación mediada por miARN está ganando mucha atención en el campo de la biología molecular y celular. De hecho, la identificación de un miARN que controla la robustez del cáncer tiene un enorme potencial terapéutico, ya que los miARN se están convirtiendo en dianas más atractivas para la terapia, como lo demuestra por primera vez el miARN-122 contra la infección por hepatitis C y el cáncer hepático (Lindow & Kauppinen, 2012). Hoy en día, al menos 7 mímicos o inhibidores de miARN se prueban en ensayos clínicos gracias al desarrollo de modificaciones químicas para aumentar su estabilidad, para mejorar la focalización en los sitios de la enfermedad o el transporte mediante varias opciones de sistemas de administración (Rupaimoole & Slack, 2017). De hecho, hay dos aprobados, uno para el tratamiento de la atrofia muscular espinal (Wurster et al., 2019) y otro para el tratamiento de amiloidosis heriditaria por transtiretina (Adams *et al.*, 2018). Sin embargo, es muy difícil identificar interacciones miARN-objetivo individuales con funciones biológicas relevantes. Los enfoques clásicos comienzan con la identificación de miARN desregulados relacionados con enfermedades y se requiere un extenso trabajo de biología celular y molecular para validar los genes diana o candidatos relacionados con el fenotipo de interés, lo cual es ineficiente porque cada miARN puede alterar la expresión de cientos de genes (Vera *et al*., 2013) y es el efecto cooperativo de las redes de miARN lo que las convierte en reguladores robustos (Herranz & Cohen, 2010; Matsuo *et al.*, 2010). Por lo tanto, la identificación de miARN críticos que puedan afectar la respuesta de esas redes requiere el análisis de un modelo matemático de esas interacciones mediante análisis de sensibilidad combinado con simulaciones predictivas para sugerir procesos bioquímicos clave para convertirse en posibles objetivos terapéuticos (Lai *et al.*, 2013; Vera *et al.*, 2013).

El Laboratorio de Quimiosensibilidad Tumoral de la UCR (LabQT) ha diseñado una metodología para la identificación de miARN con un enfoque inverso al tradicional, inicia con una lista de genes diana de interés, los cuales podrían ser determinados por análisis de

expresión diferencial o personalizados por los investigadores. Posteriormente, se construye una topología de red con todas las interacciones informadas de esos genes diana con miARN y factores de transcripción. Sin embargo, la topología de la red no es suficiente para identificar los objetivos más robustos para controlar el fenotipo de interés (Lai *et al.*, 2013; Vera *et al.*, 2013). Por lo tanto, el enfoque va más allá de la configuración de una red y establece un sistema dinámico completo basado en sistemas de ecuaciones diferenciales ordinarias (ODE), que se calibran con datos experimentales existentes. Finalmente, el análisis de sensibilidad y los experimentos *in silico* permiten identificar los nodos objetivo más robustos para regular el fenotipo de interés, determinar el mecanismo que subyace detrás de ese fenotipo y las intervenciones en esos nodos objetivo.

Previo a ese nuevo enfoque los esfuerzos para la identificación de genes diana se han basado en datos de expresión diferencial entre tejidos tumorales y no tumorales. Sin embargo, esas diferencias surgen como consecuencia de la inestabilidad genética y como tal podrían ser muy heterogéneas entre los tipos de cáncer. El enfoque propuesto por el LabQT y descrito en Acón *et al*., 2016, explora la estabilidad de la expresión génica en el cáncer, representada por un grupo reducido de genes críticos que podrían mediar la supervivencia celular a pesar de la inestabilidad genómica del cáncer. De hecho, la aneuploidía es un sello distintivo de la mayoría de las células cancerosas avanzadas y, por ende, deben haber desarrollado mecanismos para minimizar los efectos negativos de dicho fenómeno. Se sugiere, por lo tanto, que esos genes críticos forman parte de un núcleo central de estabilidad que se selecciona o mantiene durante el proceso evolutivo que conlleva a la inestabilidad típica del cáncer. También, se propone que este núcleo central de estabilidad prevalece mediante un mecanismo de compensación de la dosis de esos genes esenciales, que se sabe tienen una tolerancia muy pequeña a la variación. Si bien la expresión de genes con dosis compensada se controla en proporción al número de copias del gen, muchos otros genes escapan a la compensación de dosis en respuesta a la aneuploidía y pueden contribuir significativamente a la variación fenotípica en el cáncer (Hose *et al.*, 2015). Esta línea de investigación del LabQT generó un primer criterio para caracterizar el fenómeno de

compensación de dosis génica, el cual identifica genes candidatos como aquellos que tienen baja tolerancia en la variación de su expresión, independientemente de una gran variación en su número de copias.

Como se mencionó en la justificación, el análisis de datos genómicos para la selección de genes candidatos se vio limitado al utilizar una fuente de datos pequeña como el NCI-60. Por otra parte, el TCGA dispone de un mayor volumen y complejidad de datos que además provienen de muestras tumorales primarias humanas, las cuales están asociadas con datos de sobrevida. Este tipo de datos -por ejemplo- ha facilitado investigaciones para evaluar varios aspectos en la biología del cáncer y cómo esto puede repercutir en el manejo clínico de los pacientes. Por ejemplo, se ha estudiado si los perfiles transcripcionales de tejido normal adyacentes al tumor mejoran la predicción de supervivencia (Huang *et al.,* 2016). De la misma forma, estos han servido para ampliar la comprensión de las características moleculares del carcinoma hepatocelular y así mejorar el pronóstico y desarrollo de nuevas estrategias de terapia (Zhu *et al.,* 2018). Adicionalmente este aservo de datos ha sido de utilidad para probar hipótesis sobre el efecto que tiene la expresión de ciertos miARN en el aumento de la sobrevida en cáncer de mama (Kin *et al.,* 2018), identificar ARNlnc que funcionen como biomarcadores en nuevos mecanismos de diagnóstico en cáncer gástrico (Zhang, 2019) y realizar análisis de supervivencia de datos multiómicos para identificar potenciales biomarcadores para pronóstico del Adenocarcinoma Pancreático Ductal (Kumar, 2019). Siguiendo estas mismas líneas, de explotar las interrelaciones entre datos clínicos y ómicos, la plataforma computacional diseñada e implementada en este trabajo tuvo como fin particular el poner a disposición de los investigadores del LabQT datos del TCGA en escenarios de análisis hechos a la medida y por demanda, para la validación y ajuste de criterios para caracterizar el fenómeno de compensación.

Para esta investigación la plataforma computacional automatizó la descarga de los archivos y el proceso de transformación e integración de 2172 millones de registros en una base de datos de propósito general (Figura 4A). Dado que la implementación de la plataforma se

realizó en un ambiente de desarrollo, esa cantidad de registros procesados se obtuvo de aproximadamente 35000 archivos de datos de unos 6000 pacientes de cánceres de cerebro, mama, riñón, pulmón, colon, útero y ovario. En un ambiente en producción la plataforma podría procesar de manera automática los 95029 archivos y su metadata para la totalidad de 84609 casos y 66 tipos y subtipos de cáncer que tiene actualmente el TCGA. El proceso de transformación e integración comprende dos grandes grupos de tareas. Primero, mapear hacia cada gen los valores del número de copias asociados a los segmentos cromosómicos, para lo cual la plataforma permite incorporar datos de otras entidades que manejan información biológica, tal como, las anotaciones del genoma humano (GRCh38) que contienen las coordenadas cromosómicas de cada gen. Segundo, fusionar los datos a través de los diferentes tipo de identificadores utilizados. Cabe resaltar que la plataforma cuenta con la flexibilidad de incorporar más entidades a la base de datos de propósito general si fuera necesario. A pesar de que el mismo TCGA y otros organizaciones cuenta con facilidades para la descarga de estos datos, no se encontró una herramienta similar a la aquí desarrollada para la integración de los datos, metadatos y convertidores en estructuras que puedan ser configuradas a la medida para estudio del fenómeno de compensación de dosis génica.

A partir de esta base de datos se acondicionaron estructuras con registros a nivel de caso y gen (Figura 3-anillo A1 y Figura 4B), que facilitaron el estudio de 1066 casos de cáncer de mama en términos de expresión génica y cambios del número de copias, aplicando un Modelo de Mezclas Gaussianas sobre los coeficientes de variación de número de copias y la expresión génica y su correlación. Lo anterior condujo a la identificación de 98 genes candidatos bajo compensación de dosis con alta variación en sus números de copias pero baja tolerancia a la variación en la expresión génica. Se confirmó la compensación de la dosis mediante un ajuste lineal para 36 genes que comprenden una diversidad de funciones, como se mostró mediante el análisis de enriquecimiento funcional. Todos estos genes pueden estar conectados por una red de interacción a gran escala de miARN y factores de transcripción, lo que sugiere que los genes con compensación de dosis comparten una red

reguladora común para garantizar la supervivencia del cáncer a la inestabilidad genómica (Oviedo *et a*l., 2013).

Esos resultados son consistentes con lo informado previamente por Acón *et a*l., 2016, donde la distribución de genes candidatos mostró una tendencia hacia ciertos cromosomas sobre otros. Se identificó un mayor número de genes diana con dosis compensadas en los cromosomas 8, 11 y 17. Curiosamente, para estos cromosomas, el biotipo génico más abundante, según lo definido por GENCODE, fue el codificador de proteínas (Figura 6A y 6B). Estos hallazgos son de especial interés, ya que se ha informado que estos cromosomas presentan grupos de genes en los que parecen estar presentes dianas con dosis compensadas (Acón *et a*l., 2016). Sin embargo, los nuevos análisis también presentaron hallazgos novedosos con la identificación de varios biotipos reconocidos como ARNlnc. Estos incluían genes antisentido y ARNlnc, así como intrónicos de sentido que se encuentran en el cromosoma 19. Esto tiene gran importancia porque se ha demostrado que los elementos no codificantes participan en los mecanismos relacionados con la regulación de la expresión génica y se ha demostrado que sus funciones se ven afectadas por diferentes aberraciones genómicas, incluidas las variaciones estructurales (Diederichs *et al.*, 2016; Anastasiadou *et al.,* 2018; Yamamura *et al.*, 2018).

El análisis muestra que existe una pequeña superposición en lo que respecta a la lista de genes candidatos. Esto se puede atribuir a las diferencias en el conjunto de datos analizados. El panel de la línea celular NCI60 tiene diferentes condiciones experimentales con respecto a las del conjunto de datos de cáncer de mama TCGA. El número de muestras, las diferencias en los tratamientos experimentales, así como los orígenes de las muestras, añaden capas de complejidad que podrían explicar las diferencias en los candidatos identificados. Sin embargo, la presencia constante de MYC representa una doble validación de la solidez del enfoque y refuerza el papel previamente propuesto de este gen como modelo para la compensación de dosis, cuya regulación es fundamental para la supervivencia de las células cancerosas.

El gen MYC (protooncogén MYC, factor de transcripción bHLH), también conocido como C-MYC, es un factor de transcripción bien conocido y estudiado. Activa genes implicados en la proliferación, crecimiento celular, diferenciación celular y apoptosis. Se estima que MYC regula el 15% de los genes. También es un oncogén, generalmente sobreexpresado en muchos tipos de cáncer (Dang, 1999), estando activo en el 70% de los cánceres humanos, pero también está relacionado con la apoptosis (Zhang *et al.*, 2013). La desregulación de MYC puede conducir al cáncer, pero también al suicidio celular (Nilsson & Cleveland, 2003; Hsieh *et al.*, 2015). De hecho, se informa que tiene una función dual de oncogén a supresor de tumores en la leucemia (Uribesalgo *et a*l., 2012).

El análisis de enriquecimiento del término GO identificó la presencia de varias anotaciones. En total, se encontraron 38 términos que estaban presentes en más de 3 entradas. Estos incluyeron la función molecular, el proceso biológico y el componente celular. La presencia de múltiples términos GO que se refieren a productos génicos asociados con la unión de ADN y ARN, así como al complejo de proteínas (Figura 6C) apoya la noción de que los genes dentro del conjunto de candidatos desempeñan funciones importantes en múltiples vías reguladoras o que forman parte de bucles reguladores que deben controlarse en condiciones normales. La presencia de factores de transcripción de ARN polimerasa II, que regulan la expresión no sólo del ARN mensajero sino también de otros elementos como los ARNlnc, representa un hallazgo interesante. Las funciones potenciales de los ARNlnc en la expresión génica, la epigenética y las enfermedades son bien conocidas, pero su impacto y extensión en un modelo de compensación de dosis génica aún no se han aclarado, excepto para Xist (X-inactive specific transcript) el cual regula el silenciamiento transcripcional del cromosoma X en las hembras de los mamíferos (Sahakyan, 2018). Curiosamente, la lista de candidatos forma una red interna, donde aparecen ARNlnc, genes codificadores de proteínas y otros elementos reguladores, con estos últimos anotados con funciones moleculares características para regular la expresión de los elementos no codificantes presentes en la lista. Cabe además resaltar que el criterio de selección empleado está

diseñado para la identificación de compensación de dosis génica a nivel transcripcional o epigenético, pero existen además otros mecanismos de compensación mediada nivel proteico (Brennan *et al*., 2019). El tipo de transcripción que permaneció fuera de la lista fue miARN, que se incorporó previamente en el modelo base para la compensación de dosis de genes descrito por Acón *et a*l., 2016, usando la línea celular NCI60.

Aunque los resultados indiquen que la compensación de la dosis génica se puede dar en el cáncer de mama y que los genes involucrados tienen una regulación común por una red a gran escala de interacciones miARN y factores de transcripción, el criterio basado en regresiones lineales sobre los coeficientes de variación de la expresión y de la dosis génica se debilita por varias razones. En primer lugar, el ruido presente en un conjunto grande de datos proveniente de los factores experimentales y no genéticos utilizados para obtenerlos y digitalizarlos (Figura 7A Panel 1). Otros factores son la premisa de un comportamiento lineal de las variables, la pluralidad natural en la expresión génica de una población y los niveles diferentes de ajuste entre los genes. Esa limitación se evidencia en la amplia heterogeneidad de pendientes mostradas de la Figura 5D.

La búsqueda de una medida más robusta para representar el grado de compensación génica hizo necesario cuantificar la correlación entre las alteraciones del número de copias (CNA) y la expresión génica. Dicha cuantificación -como se mencionó en los antecedentes- se altera por el enfoque de genómica genética que combina de forma conjunta perfiles de expresión genética y genómica obtenidos de muestras tumorales. Debe recordarse que el perfil de expresión génica se realiza a menudo en biopsias que incluyen tanto células tumorales como no tumorales del microambiente tumoral, midiendo por lo tanto la expresión promedio de todos los tipos de células presentes en las biopsias de tumores. Esto significa que los efectos de los CNA en los niveles de expresión génica, además de estar influenciados por factores experimentales y no genéticos, a menudo se ven eclipsados por los efectos de las células no tumorales (Bhattacharya *et al.*, 2020).

En respuesta a lo anterior, los perfiles TACNA proveen una cuantificación, libre de ruido, del nivel de expresión génica que responde directamente a las alteraciones en la dosis de genes (Figura 7A Panel 2). Los datos de los perfiles TACNA están estructurados en forma de matriz cuyas dimensiones corresponden a identificadores Entrez para los genes y de alícuotas TCGA para los casos, esto permitió asociarlos con los datos correspondientes al número de copias a nivel de caso y gen mediante los convertidores de identificadores de la plataforma computacional. Sin embargo, para lograr el cálculo de una métrica de compensación se requirió adecuar la escala de los perfiles a la del número de copias. Para ello fue necesario revisar los detalles del mecanismo que se diseñó en Bhattacharya *et al.*, 2020, para crear los perfiles TACNA. En un primer paso, la expresión TCGA fue procesada por un algoritmo para análisis de señales -"Independent Component Analysis (ICA)" (Hyvärinen & Oja, 2000)- que la separó en componentes aditivos cada uno asociado con un peso que cuantifica el efecto que tiene el componente en la reconstrucción de la señal. Posteriormente, se identificaron los componentes que capturaron el grado de adaptación transcripcional a las alteraciones en el número de copias. Estos componentes fueron los que presentaron un patrón en el que a los genes contiguos se les asignaron pesos extremos. Este detalle de contigüedad fue la base para el proceso automatizado de reescalado. La plataforma computacional creó mapas genómicos por cada caso (Figura 4C), esto facilitó el cálculo de promedios robustos de expresión de genes que se encontraban en los segmentos cromosómicos de los archivos originales de número de copias, estos promedios robustos por segmento junto a los correspondientes valores de número de copias fueron utilizados en el ajuste de escala del perfil TACNA (Figura 7A Panel 3).

El perfil reescalado de TACNA se pudo así comparar directamente con los valores de CNA y finalmente el cociente TACNA/CNA se utilizó como una medida más robusta de compensación (Figura 7A Panel 4). En dicha medida, un cociente por debajo de 1 indica compensación de dosis génica, alrededor de 1 indica ausencia de compensación y por arriba de 1 amplificación de dosis. Los mapas genómicos correspondientes a casos de cáncer de mama están compuestos por 29 millones de registros que ampliaron la base de datos de

propósito general. La plataforma computacional permitirá generar de manera automática este tipo de mapa para los casos del TCGA de cualquier tipo de cáncer.

Esta nueva medida de compensación de dosis génica, y el hecho de contar con datos de diagnóstico, seguimiento y defunción de los pacientes, dio paso al diseño de estrategias para identificar genes cuyos grados de compensación fueran marcadamente diferentes dependiendo del grupo de sobrevida al que pertenecían los pacientes. Sin embargo, a pesar de utilizar métodos de aprendizaje automático, los análisis no evidenciaron la existencia de un comportamiento diferenciado en los niveles de compensación que pudiera asociarse al tiempo de sobrevida de los pacientes. Por esta razón se optó por cambiar el enfoque analítico de los datos y estudiar a los genes como sujeto y al comportamiento poblacional como característica. El análisis de la dispersión de tres nuevas métricas para caracterizar genes (cocientes TACNA/CNA, porcentaje de casos amplificados y porcentaje de casos compensados) llevó a la identificación de un comportamiento bimodal de genes frecuentemente amplificados que mostraron una conducta diferencial entre los dos grupos de sobrevida, destacando que los casos del grupo de baja sobrevida presentaron mayor frecuencia en la amplificación (Figura 7D y 7E). De ese grupo de genes se consideraron como dianas aquellos que tuvieran más de 90% de casos compensados, ya que estos podrían estar propiciando el avance del cáncer en medio del cambio en el número de copias de sus respectivos cromosomas, lo cual provoca amplificaciones en genes compensados y no compensados. La presencia de muchos genes compensados en múltiples procesos y funciones celulares indica que las células de cáncer deben mantener ciertos procesos esenciales intactos para conservar estabilidad, a pesar de estar en un proceso de evolución acelerada debido a su inestabilidad genética. Entonces la compensación de dosis génica evita probablemente que la célula con un genoma inestable perturbe estos procesos esenciales y no cruce límites de error que llevarían a muerte celular (Solé & Deisboeck, 2004) .

La plataforma computacional permitió el planteamiento y ajuste de hipótesis para la búsqueda de genes diana, a partir de los cuales crear redes de interacciones -entre miARN y factores de transcripción- que contribuyeron a caracterizar el comportamiento del fenómeno de compensación de dosis génica. Esto fue posible debido a la capacidad de la plataforma para gestionar de manera integrada, desde dos niveles, los datos genómicos y transcriptómicos. A nivel de los datos puede generar conjuntos a la medida de las necesidades del proceso de investigación. A nivel de análisis -utilizando las bibliotecas especializadas en analítica de Python- facilita la aplicación, dentro de la misma plataforma, de cualquier tipo de modelo, desde descriptivos hasta predictivos.

Considerando la presencia de MYC como gen diana en este y otros estudios paralelos con datos de NCI60 y CCLE, se decidió explorar su red de interacciones reguladoras, con el propósito de encontrar propiedades a nivel de sistema que permitan modelar su compensación de dosis. Por ello, se creó y ajustó un modelo matemático utilizando datos del TCGA. Dado que los resultados del ajuste del modelo completo evidenciaron configuraciones diferenciales en las pacientes de cáncer de mama, se ajustaron modelos matemáticos personalizados utilizando la configuración mínima obtenida en Acón *et al.*, 2006 (Figura 10C) y se observó heterogeneidad en la capacidad de los modelos individuales para lograr la compensación de la dosis de genes (Figura 11A). Esto sugiere que la configuración de los circuitos de compensación puede variar entre las pacientes y, probablemente, entre tipos de tumores. Para estudiar el efecto del nivel de compensación en la sobrevida, se exploró más allá de la clasificación de pacientes por cuartiles. Se generó, entonces, una clasificación por amplificación génica y capacidad de compensación individual (Figura 11B) obteniendo 3 grupos (i) baja amplificación (ii) alta amplificación con alta compensación y (iii) alta amplificación con baja compensación. Para cada grupo se obtuvo su curva de supervivencia Kaplan-Meier y la importancia de sus diferencias se evaluó mediante una regresión de riesgos proporcionales de Cox. Este modelo tiene la ventaja, sobre sólo utilizar los cuartiles, de generar una tasa de riesgo instantáneo (conocida en inglés como “Hazard Ratio”) que considera el evento de interés (en este caso la muerte)

y los tiempos en que ocurre, así como la información de las pacientes de las que no se conoce su desenlace clínico (datos censurados) (Simmons & More, 2002). Se observaron diferencias significativas entre en la probabilidad de supervivencia de las pacientes con amplificaciones compensadas en comparación con las no compensadas, y en comparación con los casos sin amplificaciones. Utilizando los espacios de compensación y no-compensación para analizar los parámetros cinéticos de las interacciones (Figuras 11D y 11E), se observó que la compensación de MYC en los modelos las pacientes con cáncer de mama estuvo mediada principalmente por miR-19a y solo unos pocos casos fueron mediados por los bucles de retroalimentación con miR-17 o miR20a.

Todo lo anterior sugiere que la compensación de la dosis de MYC permite la progresión del cáncer hacia la malignidad, razón por la cual representaría un importante objetivo terapéutico para prevenir esa progresión. Además, la futura identificación de los determinantes moleculares de la compensación, podría conducir al reconocimiento de biomarcadores valiosos para dirigir estrategias de medicina de precisión. La respuesta de la configuración de red correspondiente también arrojó luz sobre la importancia de la compensación de la dosis de genes en el cáncer. Se observó que cuanto mayor es la extensión de la amplificación de MYC, mayor es la separación de casos en términos de compensación, lo que conduce a una clara separación de 2 grupos con una diferencia significativa en la supervivencia del paciente. Estos hallazgos permiten proponer que la compensación de la dosis de MYC contribuye a una especie de estabilidad protumoral que acondiciona a las células cancerosas para evolucionar hacia fenotipos aún más malignos.

6. CONCLUSIONES

Se diseñó e implementó una plataforma computacional que facilita el manejo de grandes y complejos conjuntos de datos genómicos provenientes del TCGA así como la creación de estrategias de análisis sobre esos datos para caracterizar el fenómeno de compensación de dosis génica, en particular para la identificación de genes candidatos cuyas interacciones con miARN y factores de transcripción regulan funciones biológicas relevantes para la progresión del cáncer a pesar de su inestabilidad genómica. Esta plataforma también posibilitó la incorporación de los perfiles de adaptación transcripcional (TACNA) para contar con una cuantificación del nivel de expresión génica asociada directamente a la alteración de dosis de genes y de esta forma consolidar una nueva forma de medir el grado de compensación. Esta infraestructura facilitará la realización de análisis tipo pan-cáncer para identificar genes candidatos y circuitos de compensación comunes en distintos tipos de cáncer y sus relaciones con la sobrevida de los pacientes.

Además, se logró formular un nuevo criterio para identificar genes candidatos al reconocer un comportamiento bimodal de genes frecuentemente amplificados que mostraron una conducta diferencial entre grupos de pacientes con alta y baja sobrevida y cuyo nivel de compensación de dosis génica podría estar habilitando el avance del cáncer en medio del cambio en el número de copias de sus respectivos cromosomas. Este criterio da paso a una nueva estrategia -que puede adicionarse a la plataforma- en la que se parametricen los umbrales para clasificar los grupos de alta y baja compensación y buscar las configuración que presente más diferencia en la sobrevida. Esta estrategia se podrá utilizar para identificar circuitos de compensación de dosis génica que permitan revelar dianas terapéuticas más robustas y globales para impedir la progresión del cáncer aneuploide.

Finalmente, se obtuvo evidencia, para el caso de MYC, de heterogeneidad en los circuitos de compensación entre pacientes de un mismo tipo de cáncer, en este caso de mama. De hecho, se trazaron los parámetros cinéticos que describen las principales interacciones de

los motivos de la red de compensación para MYC dentro de un espacio tridimensional de compensación de dosis génica y se observó que no todos están ubicados en la región de compensación, lo que establece que la topología de la red por sí sola no es suficiente, ya que con diferentes valores de parámetros el modelo perdería la compensación, lo que fortalece la necesidad de modelos dinámicos completos que permitan identificar propiedades emergentes de las redes reguladoras.

REFERENCIAS

Acón M, Siles F, Mora R. 2016. A biocomputational platform for the automated construction of large-scale mathematical models of miARN-transcription factor networks for studies on gene dosage compensation. IEEE 36th Central American and Panama Convention (CONCAPAN XXXVI) doi:10.1109/CONACAPAN.2016.7942348.

Acon M, Geiß C, Torres-Calvo J, Oviedo G, Arias-Arias J.L, Vásquez-Vargas G, Oses-Vargas Y, Guevara-Coto J, Segura-Castillo A, Siles-Canales F, Quirós-Barrantes S, Régnier-Vigouroux A , Mendes P, Mora-Rodríguez R. 2021. MYC dosage compensation is mediated by miARN-transcription factor interactions in aneuploid cancer. bioRxiv: preprint.

Adams D, Gonzalez-Duarte A, O'Riordan W.D, Yang C.C, Ueda M, Kristen A.V, Tournev I, Schmidt H.H, Coelho T, Berk J.L, *et al*. 2018. Patisiran, an RNAi therapeutic, for hereditary transthyretin amyloidosis. New England Journal of Medicine. 379:11-21.

Anastasiadou E, Jacob L, Slack F. 2018. Non-coding RNA networks in cancer. Nature Reviews Cancer 18:5–18.

Ayer T, Alagoz O, Chhatwal J, Shavlik J, Kahn C, Burnside E. 2010. Breast cancer risk estimation with artificial neural networks revisited. Cancer 116:3310-3321.

Bhattacharya A, Bense R, Urzúa-Traslaviña C, de Vires E, van Vugt M, Fehrmann R. 2020. Transcriptional effects of copy number alterations in a large set of human cancers. Nature Communications 11, 715.

Birchler J, Veitia R. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proceedings of the National Academy of Sciences of USA. 109 37: 14746-14753.

Bishop C. 2006. Pattern Recognition and Machine Learning. Springer, New York.

Boyd M. 1997. The NCI In Vitro Anticancer Drug Discovery Screen. Concept, Implementation, and Operation, 1985-1995. Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials, and Approval 23-42.

Brennan C, Vaites L, Wells J, Santaguida S, Paulo J, Strokova Z, Amon A. 2019. Protein aggregation mediates stoichiometry of protein complexes in aneuploid cells. Genes & Development. 33(15-16):1031-1047.

Bussey K, Chin K, Lababidi S, Reimers M, Reinhold W, Kuo W, Gwadry F, Kouros-Mehr, H, Fridlyand J, Jain A, Collins C, Nishizuka S, Tonon G, Roschke A, Gehlhaus K, Kirsch I, Scudiero D, Gray J, Weinstein J. 2006. Integrating data on dna copy number with gene

expression levels and drug sensitivities in the nci-60 cell line panel. Molecular Cancer Therapeutics 5 4:853–867.

Clark P, Boswell R. 1991. Rule induction with CN2: Some recent improvements. Machine Learning - EWSL- 91: 151-163.

Carpenter G, Grossberg S. 1988. The art of adaptive pattern recognition by a self-organizing neural network. Computer 21:77–88.

Chin L, Andersen J, Futreal P. 2011. Cancer genomics: from discovery science to personalized medicine. Nat Med 17:297-303.

Chin L, Hahn W, Getz G, Meyerson M. 2011. Making sense of cancer genomic data. Genes and Development 25 6:534-555.

Cimini D. 2008. Merotelic kinetochore orientation, aneuploidy and cancer. Biochimica et Biophysica Acta - Reviews on Cancer.

Cooke S, Shilen A, J Marshall, Pipinikas C, Martincorena I, Tubio J, Li Y, Menzies A, Mudie L, Ramakrishna M, *et a*l. 2014. Processed pseudogenes acquired somatically during cancer development. Nature communications 5:3644.

Coto J, Siles F, Mora R. 2016. Biocomputing platform module for cancer genomics and chemotherapy. IEEE 36th Central American and Panama Convention (CONCAPAN XXXVI).

Davoli T, Uno H, Wooten E, Elledge S. 2017. Tumor Aneuploidy Correlates with Markers of Immune Evasion and with Reduced Response to Immunotherapy. Science 355:6322

Dang C.V. 1999. c-Myc Target Genes Involved in Cell Growth, Apoptosis, and Metabolism. Molecular and Cellular Biology 19(1).

de Ridder D, de Ridder J, Reinders M. 2013. Pattern recognition in bioinformatics. Briefings in Bioinformatics 14:633-647.

Devlin E, Holm D, Grigliatti T. 1982. Autosomal dosage compensation in *Drosophila melanogaster* strains trisomic for the left arm of chromosome 2. Proceedings of the National Academy of Sciences USA 79 4:1200-1204.

Diederichs S, Bartsch L, Berkmann JC, Fröse K, Heitmann J, Hoppe C, Iggena D, Jazmati D, Karschnia P, Linsenmeier M, Maulhardt T, Möhrmann L, Morstein J, Paffenholz SV, Röpenack P, Rückert T, Sandig L, Schell M, Steinmann A, Voss G, Wasmuth J, Weinberger ME, Wullenkord R. 2016. The dark matter of the cancer genome: aberrations in regulatory

elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. EMBO Molelcular Medicine 8(5):442-57.

Donnelly N, Storchová Z. 2014. Dynamic karyotype, dynamic proteome: buffering the effects of aneuploidy. Biochimica et Biophysica Acta 1843 2:473-481.

Duan K, Rajapakse J, Wang H, Azuaje F. 2005. Multiple SVM-RFE for gene selection in cancer classification with expression data. NanoBioscience, IEEE Transactions on 4(3): 228-234.

Duda R, Hart P, Stork D. 2012. Pattern Classification. Wiley. New York.

Duesberg P, Raush C, Rasnick D, Hehlmann R. 1998. Genetic instability of cancer is proportional to their degree of aneuploidy. Proceedings of the National Academy of Sciences of the United States of America 95 23:13692-13697.

Ester M, Kriegel H.P., Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD-96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining: 226-231.

Fabarius A,Willer A, Yerganian G, Helhmann R, Duesberg P. 2002. Specific aneusomies in chinese hamster cells at different stages of neoplastic transformation, initiated by nitrosomethylurea. Proceedings of the National Academy of Sciences USA 99 10:6778–6783.

Fabarius A, Li R, Yerganian G, Helhmann R, Duesberg P. 2008. Specific clones of spontaneously evolving karyotypes generate individuality of cancers. Cancers Genetics and Cytogenetics 180 2:89-99.

Fabian M.R., Sonenberg N, Filipowicz W. 2010. Regulation of mRNA translation and stability by microRNAs. Annual Review of Biochemistry 79:351-379.

Fukushima K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36:193–202.

Fürnkranz J. 2017. Rule Learning. Springer, Boston, MA.

Garro A. 2016. Reconocimiento de patrones de espectroscopía de absorción para la clasificación de enfermedades. https://www.inii2.ucr.ac.cr/RIINII/pdf/IE/IE-7785.pdf.

Gevaert O, De Smet F, Timmerman, D. Moreau Y. 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 22:e184-190.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Research 36:D154–D158.

Guyon I, Weston J, Barnhill S. 2002. Gene selection for cancer classification using support vector machines. Machine learning 46 (1): 389-422.

Hanahan D, Weinberg A. 2000. The hallmarks of cancer. Cell 100:57-70.

Hanahan D, Weinberg A. 2011. Hallmarks of cancer: The next generation. Cell 144 5:646-674.

Hanna J, Hossain G.S., Kocercha J. 2019. The potencial for microRNA therapeutics and clinical research. Frontiers in Genetics 10 MAY.

Hardy K, Hardy P.J. 2015. 1(st) trimester miscarriage: four decades of study. Translational Pediatrics.

Hassold T, Hall H, Hunt P. 2007. The origin of human aneuploidy: where we have been, where we are going. Human Molecular Genetics 16:R203–R208

Hastie T, Tibshirani R, Friedman J, Franklin J. 2005. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York.

Herranz H, Cohen S.M. 2010. MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. Genes & Development 24(13):1339–1344.

Holland A, Cleveland D. 2009. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nature Reviews Molecular cell biology 10 7:478–487.

Hoops S, Sahle S, Gauges R, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U. 2006. COPASI - a COmplex PAthway SImulator. Bioinformatics 22 24:3067-3074.

Horn T, Sandmann T, Fischer B, Axelsson E, Huber W, Boutros M. 2011. Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. Nature Methods 8:341-346.

Hose J, Young C.M., Sardi M, Wang Z, Newton M.A. Gasch A.P. 2015. Dosage compensation can buffer copy number variation in wild yeast. Elife 4:e05462.

Huang X, Stern D, Zhao H. 2016. Transcriptional Profiles from Paired Normal Samples Offer Complementary Information on Cancer Patient Survival – Evidence from TCGA Pan-Cancer Data. Scientific Reports 6:20567.

Hucka M, Finney A, Sauro H, Bolouri H, Doyle J, Kitano H, Arkin A, Bornstein B, Bray D, Cornish A, Cuellar A, Dronov S, Gilles E, Ginkel M, Gor V, Goryanin I, Hedley W, Hodgman T, Hofmeyr J, Hunter P, Juty N, Kasberger J, Kremling A, Kummer U, Le N, Loew L, Lucio D, Mendes P, Minch E, Mjolsness E, Nakayama Y, Nelson M, Nielsen P, Sakurada T, Scha J, Shapiro B, Shimizu T, Spence H, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J. 2003. The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. Bioinformatics 19 4:524-531.

Hsieh A, Waltonm Z, Altmanm BJ, Stine ZE, Dang CV. 2015. MYC and metabolism on the path to cancer. Seminars in Cell & Developmental Biology 43:11-21.

Hyvärinen A, Oja E. 2000. Independent component analysis: algorithms and applications. Neural Networks 13(4-5):411-30.

Ishikawa K, Makanae K, Iwasaki S, Ingolia NT, Moriya H. (2017) Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes. PLoS Genetics 13(1):e1006554.

Kim D, Sung Y.M., Park J, Kim S, Kim J, Park J, … Baek D. 2016. General rules for functional microRNA targeting. Nature Genetics 48 12:1517-1526.

Kim, S.Y., Kawaguchi T, Yan L, Young J, Qi Q, Takabe K. 2017. Clinical Relevance of microRNA Expressions in Breast Cancer Validated Using the Cancer Genome Atlas (TCGA). Annals of Surgical Oncology 24:2943–2949.

Kitano H. 2004. Cancer as a robust system: implications for anticancer therapy. Nature Reviews. Cancer. 4 3:227-235.

Kojima S, Cimini D. 2019 Aneuploidy and gene expressions: Is there dosage compensation? Epigenomics Vol.11:1827-1837.

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. MiRBase: From microRNA sequences to function. Nucleic Acids Research 47 D1:D155-D162.

Kumar M, Siddesh S, Chittibabu G. 2019. Survival Analysis of Multi-Omics Data Identifies Potential Prognostic Markers of Pancreatic Ductal Adenocarcinoma. Frontiers in Genetics 10:624.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. Cell 172(4):650-665.

Lai X, Bhattacharya A, Schmitz U, Kunz M, Vera J, Wolkenhauer O. 2013. A systems' biology approach to study microrna-mediated gene regulatory networks. BioMed Research International (Ii).

Lengauer C, Kinzler K, Vogelstein B. 1998. Genetic instabilities in human cancers. Nature 396:643-9.

Li L, McCormack A, Nicholson J, Fabarius A, Helhlmann R, Sachs R, Duesberg P. 2009. Cancer-causing karyotypes: chromosomal equilibria between destabilizing aneuploidy and stabilizing selection for oncogenic function. Cancer Genetics 188 1:1–25.

Lindow M, Kauppinen S. 2012. Discovering the first microRNA-targeted drug. The Journal of Cell Biology 199(3):407-412.

Loeb L, Harris C. 2008. Advances in chemical carcinogenesis: a historical review and prospective. Cancer Res. 68:6863–6872.

Loo L, Lin H, Singh D, Lyons K, Altschuler S, Wu L. 2009. Heterogeneity in the physiological states and pharmacological responses of differentiating 3T3-L1 preadipocytes. Journal of Cell Biology. 187:375-384.

Matsuo K, Eno M.L., Im D.D., Rosenshein N.B., Sood A.K.. 2010. Gynecologic Oncology Clinical relevance of extent of extreme drug resistance in epithelial ovarian carcinoma. Gynecologic Oncology 116(1):61–65.

Mercer J, Snijder B, Sacher R, Burkard C, Bleck C, Stahlberg H, Pelkmans L, Helenius A. 2012. RNAi screening reveals proteasome- and Cullin3-dependent stages in vaccinia virus infection. Cell Reports 2:1036-1047.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research 47:D419–D426.

Nilsson J.A., Cleveland J.L. 2003. Myc pathways provoking cell suicide and cancer. Oncogene 22:9007–9021.

Ozery-Flato L, Linhart C, Izraeli S, Shamir R. 2011. Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. Genome Biology 12 6:r61.

Oviedo G, Acón M, Guevara J, Mora R. 2013. Analysis of Large Scale Gene Expression Data sets from TCGA Identifies Potential Candidate Genes under Dosage Compensation in Breast Cancer. Proceedings of the 2018 International Conference on Bioinformatics & Computational Biology 130-137.

Pavelka N, Rancati G, Zhu J, Bradford WD, Saraf A, Florens L, Sanderson B, Hattem G, Li R. 2010. Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. Nature 468:321–325.

Pang-Ning T, Steinbach M, Kumar V. 2006. Introduction to data mining. Pearson Addison Wesley, Boston.

Reddy E, Reynolds R, Santos E, Barbacid M. 1982. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature 300:149–152.

Ritchie W, Rasko J.E.J., Flamant S. 2013. MicroRNA target prediction and validation. Advances in Experimental Medicine and Biology 774:39-53.

Rosenblatt F. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65:386–408.

Rupaimoole R, Slack F.J. 2017. MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases. Nature Reviews Drug Discovery 16(3):203–221.

Sahakyan A, Yang Y, Plath K. 2018. The role of Xist in X-Chromosome Dosage Compensation. Trends in Cell Biology 28(12):999-1013.

Schvartzman J, Sotillo R, Benezra R. 2010. Mitotic chromosomal instability and cancer: mouse modelling of the human disease. Nature Reviews Cancer 10:102–115.

Shankavaram T, Reinhold W, Nishizuka Satoshi, Major S, Morita D, Chary K, Reimers M, Scherf U, Kahn A, Dolginow D, Cossman J, Kaldjian E, Scudiero D, Petricoin E, Liotta L, Lee J, Weinstein J. 2007. Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study. Molecular Cancer Therapeutics 6 3:820–832.

Sheltzer J, Amon A. 2011. The aneuploidy paradox: Costs and benefits of an incorrect karyotype. Trends in Genetics 27 11:446–453.

Symons M, Moore D. 2002. Hazard rate ratio and prospective epidemiologi-cal studies. J Clin Epidemiol 55(9):893-899.

Solé R.V., Deisboeck T.S. 2004. An error catastrophe in cancer? Journal of Theoretical Biology 228 1:47-54.

Stingele S, Stoehr G, Peplowska K, Cox J, Mann M, Storchova Z. 2012. Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. Molecular System Biology 8 608.

Stratton M, Campbell P, Futreal P. 2009. The cancer genome. Nature 458 7239:719-724

Tabin C, Bradley S, Bargmann C, Weinberg R, Papageorge A, Scolnick E, Dhar R, Lowy D, Chang E. 1982. Mechanism of activation of a human oncogene. Nature 300:143–149.

Tomczak K, Czerwinska P, Wiznerowicz M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology 19 1A:A68-A77.

Torres E, Sokolsky T, Tucker C, Chan L, Boselli M, Dunham M, Amon A. 2007. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science 317 5840: 916–924.

Torres E, Amon A. 2008. Aneuploidy: Cells losing their balance. Genetics 179 2:737–746.

Torres E, Dephoure N, Panneerselvam A, Tucker C, Whittaker C, Gygi S, Dunham M, Amon A. 2010. Identification of aneuploidy-tolerating mutations. Cell 143:71–83.

Trigos A, Pearson R, Papenfuss A, Goode D. 2017. Altered Interactions Between Unicellular and Multicellular Genes Drive Hallmarks of Transformation in a Diverse Range of Solid Tumors. Proceedings of the National Academy of Sciences 114 (24):6406-6411.

Tweedie S, Braschi B, Gray K, Jones T, Seal RL, Yates B, Bruford EA. 2021. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Research 49:D939–D946.

Uribesalgo I, Benitah S, Di Croce L. 2012. From oncogene to tumor suppressor: The dual role of Myc in leukemia. Cell Cycle 9(11):1757-1764.

Vargas J, Mora R, Siles F. 2018. Genome copy number feature selection based on chromosomal regions alterations and chemosensitivity subtypes. 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI).

Vera J, Lai X, Schmitz U, Wolkenhauer O. 2013. MicroRNA-regulated networks: the perfect storm for classical molecular biology, the ideal scenario for systems biology. Advances in Experimental Medicine and Biology 774:55–76.

Vielle A, Lang J, Dong Y, Ercan S, Kotwaliwale C, Rechtsteiner A, Appert A, Chen Q, Dose A, Egelhofer T, Kimura H, Stempor P, Dernburg A, Lieb J, Strome S, Ahringer J. 2012. H4K20me1 Contributes to Downregulation of X-Linked Genes for C. elegans Dosage Compensation. PLoS Genetics 8 9:e1002933.

Wang J, Cazzato E, Ladewig E, Frattini V, Rosenbloom D, Zairis S, Abate F, Liu Z, Elliott O, Shin Y, Lee J, Lee I, Park W, Eoli M, Blumberg A, Lasorella A, Nam D, Finocchiaro G, Iavarone A, Rabadan R. 2016. Clonal evolution of glioblastoma under therapy. Nature Genetics 48:768-776.

Weaver B.A., Cleveland D.W. 2006. Does aneuploidy cause cancer? Current Opinion in Cell Biology.

Wild T, Horvath P, Wyler E, Widmann B, Badertscher L, Zemp I, Kozak K, Csucs G, Lund E, Kutay U. 2010. A protein inventory of human ribosome biogenesis reveals an essential function of exportin 5 in 60S subunit export. PLoS Biology 8:e1000522.

Williams B, Amon A. 2009. Aneuploidy: Cancer's fatal flaw? Cancer Research 69 13:5289–5291.

Wippich F, Bodenmiller B, Trajkovska M, Wanka S, Aebersold R, Pelkmans L. 2013. Dual specificity kinase DYRK3 couples stress granule condensation/dissolution to mTORC1 signaling. Cell 152:791-805.

Witten I, Frank E. 2005. Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington.

Wurster C.D, Winter B, Wollinsky K, Ludolph A.C, Uzelac Z, Witzel S, Schocke M, Schneider R, Kocak T. 2019. Intrathecal administration of nusinersen in adolescent and adult SMA type 2 and 3 patients. Journal of Neurology. 266:183-194.

Wurzenberger C, Held M, Lampson M, Poser I, Hyman A, Gerlich, D. 2012. Sds22 and Repo-Man stabilize chromosome segregation by counteracting Aurora B on anaphase kinetochores. Journal of Cell Biology 198:173-183.

Xu X, Zhang Y, Zou Liang, Wang M, Li A. 2012. A gene signature for breast cancer prognosis using support vector machine. International Conference on BioMedical Engineering and Informatics 2012:928-931.

Yamamura S, Imai-Sumida M, Tanaka Y, Dahiya R. 2018. Interaction and cross-talk between non-coding RNAs. Cellular and Molecular Life Science 75:467–484.

Yu K. Zhang C, Berry G, Altman R, Re C, Rubin D, Snyder M. 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature Communications 7 12474 doi:1038/ncomms12474.

Zerbino D, Achuthan P, Akanni W, M Amode R, Barrell D, Bhai J, Billis K, Cummins C, Gall A, García-Girón C, Gil L, et al. 2018. Ensembl. Nucleic Acids Research 46:D754–D761.

Zhang Q, Spears E, Boone DN, Li Z, Gregory MA, Hann SR. 2013. Domain-specific c-Myc ubiquitylation controls c-Myc transcriptional and apoptotic activity. Proceedings of the National Academy of Sciences 110(3):978-983.

Zhang X, Lu X, Shi Q, Xu X, Leung H, Harris L, James D, Miron A, Liu J, Wong W. 2006. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC bioinformatics 7(1): 197.

Zhang X, Zhang W, Jiang Y, Liu K, Ran L, Song F. 2019. Identification of functional ARNlncs in gastric cancer by integrative analysis of GEO and TCGA data. Journal of Cell Biochemistry. October 120(10):17898-17911.

Zhong Q, Busetto G, Fededa J, Buhmann J, Gerlich D. 2012. Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. Natture Methods 9:711-713

Zhu, Q, Sun, Y, Zhou, Q, He, Q, Qian, H. 2018. Identification of key genes and pathways by bioinformatics analysis with TCGA RNA sequencing data in hepatocellular carcinoma. Molecular and Clinical Oncology 9:597-606.