

Priorización de genes y búsqueda de fármacos por medio de herramientas informáticas y técnicas de aprendizaje de máquinas en osteosarcoma

Raúl Alejandro Cabrera-Andrade

Tesis doctoral UDC / 2021

Directores: Cristian Robert Munteanu y Humberto González Díaz.

Tutor: Alejandro Celestino Pazos Sierra.

Programa de doctorado en Tecnologías de la Información y las Comunicaciones



UNIVERSIDADE DA CORUÑA



UNIVERSIDADE DA CORUÑA

Dr. Cristian Robert Munteanu, Profesor Titular de Universidad en el área de Ciencias de la Computación e Inteligencia Artificial, perteneciente al Departamento de Ciencias de la Computación y Tecnologías de la Información, Facultad de Informática, Universidade da Coruña

y

Dr. Humberto González Díaz, Prof. Investigador del Departamento de Química Orgánica e Inorgánica, Universidad del País Vasco (UPV/EHU), BIOFISIKA, Insituto Vasco de Biofísica (UPV/EHU, CSIC), e IKERBASQUE, Fundación Vasca para la Ciencia, Bilbao, España.

HACEN CONSTAR QUE:

La memoria “**Priorización de genes y búsqueda de fármacos por medio de herramientas informáticas y técnicas de aprendizaje de máquinas en osteosarcoma**” ha sido realizada por Alejandro Cabrera-Andrade bajo nuestra dirección en el Grupo de Investigación RNASA-IMEDIR, y constituye la Tesis que presenta para optar al Grado de Doctor en Informática de la Universidad de Coruña.

A Coruña, 28 de enero de 2021.

Fdo. Cristian Robert Munteanu

Fdo. Humberto González Díaz

*A mi madre y padre,
a Esteban y su mamá,
a mis hermanos y querida familia*

Agradecimientos

Este proyecto de Tesis Doctoral ha sido realizado en la Facultad de Informática de la Universidad de la Coruña, bajo la dirección del Prof. Dr. Cristian Robert Munteanu y Dr. Humberto González Díaz, a quienes quiero agradecer por su dirección científica, acompañamiento a lo largo de este programa y constante apoyo.

Agradezco a la Universidad de las Américas por haberme brindado la apertura en el desarrollo de este trabajo de investigación. De manera especial a Eduardo Tejera por haber sido parte esencial en el desarrollo de todo este trabajo, por transmitirme todos sus conocimientos y amistad; a Yunierkis Pérez por su guía y ayuda; a Tannya Lozada por su apertura, apoyo y confianza. A todos mis compañeros que se involucraron de cierta forma en esta investigación.

De igual forma, a mis amigos y colegas Andrés López y Gabriela Jaramillo por sus valiosos aportes y su compañía durante este período de tesis doctoral

Por último, a toda mi familia y amigos por haber hecho de este tiempo algo especial.

¡Gracias totales!

Resumen

El osteosarcoma es el subtipo más común de cáncer de hueso primario y afecta principalmente a adolescentes. En los últimos años, varios estudios se han centrado en dilucidar los mecanismos moleculares de este sarcoma; sin embargo, su etiología molecular aún no se ha determinado con precisión. Por otro lado, su diagnóstico clínico es generalista y sus terapias no han cambiado en las últimas décadas. Aunque hoy en día las tasas de supervivencia a 5 años pueden alcanzar hasta el 60-70%, las complicaciones agudas y los efectos tardíos del tratamiento del osteosarcoma son dos de los factores limitantes de los tratamientos. Así, el objetivo de esta tesis doctoral es desarrollar una estrategia de priorización que permita la identificación de genes asociados con la patogenicidad del osteosarcoma y explicar de forma más completa la etiología de esta enfermedad. Por otro lado, se busca desarrollar algoritmos de predicción de fármacos basados en aprendizaje de máquinas que permitan proponer nuevos agentes terapéuticos para el tratamiento de esta enfermedad. Todos los resultados obtenidos se publicaron en revistas científicas internacionales con importante factor de impacto JCR.

Abstract

Osteosarcoma is the most common subtype of primary bone cancer, affecting mainly adolescents. In recent years, several studies have focused on elucidating the molecular mechanisms of this sarcoma; however, its molecular etiology has not yet been accurately determined. On the other hand, the clinical diagnosis is generalist and therapies have not changed in recent decades. Although nowadays 5-year survival rates can reach up to 60-70%, acute complications and late effects of osteosarcoma therapy are two of the limiting factors in treatments. Thus, the objective of this doctoral thesis is to develop a prioritization strategy that allows the identification of genes associated with the pathogenicity of osteosarcoma, and to explain more fully the etiology of this disease. On the other hand, it seeks to develop drug prediction algorithms based on machine learning techniques that allow proposing new therapeutic agents for the treatment of this disease. All the results obtained in this research were published in international scientific journals with an important JCR impact factor.

Resumo

O osteosarcoma é o subtipo máis común de cancro óseo primario, que afecta principalmente a adolescentes. Nos últimos anos, varios estudos centráronse en dilucidar os mecanismos moleculares deste sarcoma; con todo, a súa etioloxía molecular aínda non foi determinada con precisión. Por outra banda, o seu diagnóstico clínico é xeralista e as súas terapias non cambiaron nas últimas décadas. Aínda que hoxe as taxas de supervivencia a 5 anos poden chegar ata o 60-70%, as complicacións agudas e os efectos tardíos do tratamento con osteosarcoma son dous dos factores limitantes dos tratamentos. Deste xeito, o obxectivo desta tese de doutoramento é desenvolver unha estratexia de priorización que permita a identificación de xenes asociados á patoxenicidade do osteosarcoma e explicar máis plenamente a etioloxía desta enfermidade. Por outra banda, buscamos desenvolver algoritmos de predición de medicamentos baseados na aprendizaxe automática que permitan propoñer novos axentes terapéuticos para o tratamento desta enfermidade. Todos os resultados obtidos publicáronse en revistas científicas internacionais cun importante factor de impacto JCR.

ÍNDICE

1. INTRODUCCIÓN	1
1.1. Generalidades del osteosarcoma.	1
1.2. Tratamiento y sobrevida en pacientes con tumores óseos.	2
1.3. Biología molecular y genes <i>driver</i> en osteosarcoma.	2
1.4. Estrategias <i>consensus</i> de priorización de genes.	3
1.5. Análisis de enriquecimiento funcional y ontología génica.	5
1.6. Modelos de aprendizaje de máquinas para predicción de fármacos anti-sarcoma	7
1.6.1. Modelo PTML para predicción de fármacos anti-sarcoma.	7
1.6.2. Modelo de multi-objetivos.	9
2. OBJETIVOS	13
2.1. Objetivo general	13
2.2. Objetivos específicos	13
3. RESULTADOS Y DISCUSIÓN	14
3.1. Uso de técnicas bioinformáticas para priorización de genes patogénicos en osteosarcoma	14
3.1.1. Métodos de priorización y estrategia <i>consensus</i>.	14
3.1.2. Ontología génica, redes de interacciones proteína – proteína y análisis de comunalidades.	16
3.2. Modelos de aprendizaje de máquinas y reposicionamiento de fármacos en osteosarcoma	25
3.2.1. Modelo de aprendizaje de máquinas teoría de la perturbación (PTML) en sarcomas.	25
3.2.2. Moldeos PTML generados en investigación oncológica.	26
3.2.3. Comparación entre modelos PTML y modelos ML.	29
3.2.4. Modelo multi-objetivo de predicción para fármacos anti-sarcoma.	31
3.2.5. Cribado virtual y reposicionamiento de fármacos para osteosarcoma.	32
4. CONCLUSIONES	38
5. REFERENCIAS	40
6. PUBLICACIONES (ANEXOS)	52

Artículos científicos publicados a partir del desarrollo de este proyecto de investigación:

Cabrera-Andrade, A., López-Cortés, A., Jaramillo-Koupermann, G., Paz-y-Miño, C., Pérez-Castillo, Y., Munteanu, C. R., González-Díaz, H., Pazos, A., Tejera, E. (2020). Gene Prioritization through Consensus Strategy, Enrichment Methodologies Analysis, and Networking for Osteosarcoma Pathogenesis. *International journal of molecular sciences*, 21(3), 1053. doi:10.3390/ijms21031053

Cabrera-Andrade, A., López-Cortés, A., Munteanu, C. R., Pazos, A., Pérez-Castillo, Y., Tejera, E., Arrasate, S., González-Díaz, H. (2020). Perturbation-Theory Machine Learning (PTML) Multilabel Model of the ChEMBL Dataset of Preclinical Assays for Antisarcoma Compounds. *ACS omega*, 5(42), 27211-27220. doi: 10.1021/acsomega.0c03356

Cabrera-Andrade, A., López-Cortés, A., Jaramillo-Koupermann, G., González-Díaz, H., Pazos, A., Munteanu, C. R., Pérez-Castillo, Y., Tejera, E. (2020). A Multi-Objective Approach for Anti-Osteosarcoma Cancer Agents Discovery through Drug Repurposing. *Pharmaceuticals*, 13(11), 409. doi:10.3390/ph13110409

Publicaciones adicionales con afiliación de la Universidad de A Coruña:

López-Cortés, A., Paz-y-Miño, C., **Cabrera-Andrade, A.**, Barigye, S. J., Munteanu, C. R., González-Díaz, H., Pazos, A., Pérez-Castillo, Y., Tejera, E. (2018). Gene prioritization, communality analysis, networking and metabolic integrated pathway to better understand breast cancer pathogenesis. *Scientific reports*, 8(1), 1-15. doi: 10.1038/s41598-018-35149-1

López-Cortés, A., Paz-y-Miño, C., Guerrero, S., **Cabrera-Andrade, A.**, Barigye, S. J., Munteanu, C. R., González-Díaz, H., Pazos, A., Pérez-Castillo Y, Tejera, E. (2020). OncoOmics approaches to reveal essential genes in breast cancer: a panoramic view from pathogenesis to precision medicine. *Scientific reports*, 10(1), 1-21. doi: 10.1038/s41598-020-62279-2

López-Cortés, A., **Cabrera-Andrade, A.**, Vázquez-Naya, J. M., Pazos, A., González-Díaz, H., Paz-y-Miño, C., Guerrero, S., Pérez-Castillo, Y., Tejera, E., Munteanu, C.R. (2020). Prediction of breast cancer proteins involved in immunotherapy, metastasis, and RNA-binding using molecular descriptors and artificial neural networks. *Scientific Reports*, 10(1), 1-13. doi: 10.1038/s41598-020-65584-y

1. INTRODUCCIÓN

1.1. Generalidades del osteosarcoma.

El osteosarcoma (OS) es una enfermedad genética rara que representa el 20% de todos los tipos de neoplasias malignas y benignas en el hueso. Es el tipo de cáncer primario de hueso más prevalente en niños y adolescentes de 0 - 19 años de edad, y afecta de manera similar en ambos sexos (1). Según la OMS (2), su incidencia anual es de 3.1 por cada millón de habitantes para población en general, y del 4.4 por cada millón para individuos menores a 25 años de edad. Además, su incidencia presenta una distribución bimodal, con un alto número de casos diagnosticados durante la adolescencia, y un incremento en el diagnóstico para personas de edad avanzada (3).

El sitio común para el desarrollo de tumores OS es en la metáfisis, y principalmente en los huesos largos como fémur, radio, cúbito, tibia y peroné. A nivel histológico, pueden subdividirse como convencionales, de bajo grado central, periosteal, parosteal, telangiectásicos, condroblásticos y de células pequeñas (4). Todos estos tipos están compuestos por osteoblastos malignos, y a pesar de toda esta diversidad histológica, una característica compartida es la alta producción de tejido osteoide inmaduro. En la mayoría de los casos su presentación clínica es asintomática por lo que su diagnóstico es tardío. En individuos que presentan estadios tumorales avanzados, más del 90% presentan dolor en la zona afectada, y además se evidencia inflamación y disminución en el rango de movimiento (5).

La histopatología es la principal herramienta de diagnóstico en biopsias de tejido tumoral, sin embargo, existe un sub-diagnóstico en estadios tempranos (principalmente en tumores de bajo grado y en biopsias poco diferenciadas), y solo el 20% de pacientes con metástasis son clínicamente detectables (6). La agresividad de los tumores óseos primarios se inicia cuando los osteocitos pierden contacto con el tejido parental e ingresan en la microvasculatura. Esta agresividad, descrita como metástasis es la primera causa de muerte en OS, es característica en los tumores óseos y el 80% de pacientes poseen micro-metástasis indetectables al momento del diagnóstico. Solo el 30 % de pacientes con neoplasias metastásicas alcanzan una tasa sobrevivencia de 5 años (7-9) por lo que el desafío actual se enfoca en el desarrollo de técnicas para diagnóstico temprano y mejora en su tratamiento.

1.2. Tratamiento y sobrevida en pacientes con tumores óseos.

El tratamiento de los pacientes diagnosticados con OS no ha cambiado en las últimas décadas. La terapia sistémica actual de primera línea incluye ciclos de cisplatino, doxorubicina y metotrexato en altas dosis. Como segunda línea se integran algunos inhibidores de la tirosina quinasa, como sorafenib y everolimus, además de agentes antineoplásicos como etopósido, topotecán y ciclofosfamida (10). La quimioterapia neoadyuvante generalmente se administra durante un período de 10 semanas, seguida de la resección quirúrgica del área tumoral comprometida y radioterapia. Si el 90% o más del área del tumor presenta necrosis, se aplican ciclos adicionales de terapia posoperatoria para rechazar la micrometástasis (5, 11, 12).

A pesar de que los tratamientos actuales han mejorado la sobrevida de los pacientes a largo plazo del <20 al 70% cuando el tumor se encuentra localizado, el progreso en una terapia personalizada ha sido lenta en las últimas décadas y sobre todo en pacientes con metástasis en donde se observan en el $\leq 20\%$ de afectos tasas de supervivencia a 5 años (13, 14). Esta dificultad en la propuesta de nuevas terapias más eficaces y sensibles se debe principalmente a la alta heterogeneidad de este tipo de tumores. Además, otro factor importante es la variabilidad genética propia de las poblaciones estudiadas. La investigación oncológica fundamental, en donde se busca comprender procesos biológicos tumorales, como factores que inducen crecimiento celular descontrolado, inhibición de procesos apoptóticos relacionados con supresión tumoral y eventos de migración tumoral definida como metástasis principalmente, han sido desarrollados a partir de modelos celulares, bio-bancos y clasificaciones moleculares a partir de estudios genómicos en poblaciones caucásicas (15). En consecuencia, se crea un sesgo importante en la descripción de patrones genéticos ya que las variantes, biomarcadores moleculares e incluso los fármacos desarrollados, son específicos para este grupo poblacional.

1.3. Biología molecular y genes *driver* en osteosarcoma.

Varios subtipos histopatológicos de cáncer de hueso presentan un comportamiento biológico distinto y sobre todo una alta variabilidad molecular. Una de las principales características moleculares del OS es su alta inestabilidad cromosómica y aneuploidía. El uso de metodologías citogenéticas y moleculares convencionales como la aplicación de cariotipo, hibridación genómica comparativa (CGH), hibridación fluorescente in situ (FISH), PCR cuantitativa (qPCR) y análisis de polimorfismo de conformación de cadena simple han evidenciado una alta heterogeneidad tumoral (16). Posteriormente, tecnologías *ómicas* como secuenciación de siguiente generación (NGS), microarrays de expresión y de variación en el

número de copias, ha generado una descripción de genes *driver* inmersos en procesos tumorigénicos para este sarcoma. *Drivers* significativos para OS se tienen a *TP53*, *NOTCH1*, *MYC*, *FOS*, *BF2*, *WIF1*, *BRCA2*, *APC*, *PTCH1* y *PRKARIA* (17). Por otro lado, *driver-sinérgicos* definidos como promotores de desarrollo tumoral en conjunto con *drivers* oncológicos se tiene a *RBI*, *TWIST*, *PTEN* y *JUN* (18).

Esta descripción en la interacción de genes *driver* ayuda también a dilucidar las rutas de señalización que se da en una célula tumoral ósea, por lo tanto, se puede generar una visión más sistémica en función de rutas metabólicas implicadas en este proceso oncológico (19, 20). Por ejemplo, vías específicas para diferenciación a tejido óseo a partir de células mesenquimales como Hedgehog, Notch y WNT tienen implicaciones importantes en la tumorigénesis y desarrollo del OS. Los receptores NOTCH son uno de los principales receptores en inducir diferenciación de células mesenquimales a osteocitos, en donde su activación limita la diferenciación la condrogénesis (21). En muestras OS, se ha relacionado sobre-expresión de genes NOTCH (NOTCH1, NOTCH2 y NOTCH3) con fenotipos metastásicos (22) y en la actualidad existen pruebas pre-clínicas para su inhibición (RO4929097, inhibidor de la γ -secretasa para la vía Notch) (23). FGF es otra vía asociada con la osteogénesis controlando el desarrollo de la placa de crecimiento y proliferación de los condrocitos, vía AKT y MEK1 (24). Pérdida de heterocigosidad en FGFR2 se asocia con OS de alto grado (25), mientras que la sobre-expresión de los genes IGFR1 y VEGF se correlacionan con crecimiento, invasión, progresión y supervivencia baja (26, 27). Otro marcador importante asociado con una prognosis pobre es HER2 (ERBB2). Este receptor de membrana tiene implicaciones importantes en varios tipos de tumores cancerígenos y en OS. HER2 es un blanco interesante para el desarrollo de terapia personalizada, sin embargo existe mucha evidencia contradictoria sobre el nivel de expresión y su valor pronóstico (28) por lo que su valor predictivo debe ser evaluado.

1.4. Estrategias *consensus* de priorización de genes.

El desarrollo y aplicación de estrategias *consensus* mediante herramientas computacionales, que utilizan múltiples fuentes de datos, aumenta la confiabilidad de un proceso de toma de decisiones y permite mejorar la detección de genes relacionados con rasgos complejos o clínicos específicos fenotipos (29). Esta combinación de enfoques conceptualmente diferentes puede proporcionar herramientas de priorización con mayor eficiencia por lo que han sido utilizadas para la priorización de genes envueltos en la patogénesis de trastornos neurodegenerativos (30), preeclampsia (31), y para cáncer de mama (32). En este sentido, se propone el desarrollo de una estrategia de priorización, que será

integrada utilizando una estrategia consensus en donde se ponderarán aquellos genes inmersos en la patogénesis del OS.

Para el desarrollo de la estrategia de priorización consensus escogemos nueve métodos bioinformáticos que cumplen con dos criterios principales: disponibilidad en plataformas web y la entrada de un nombre/enfermedad para la priorización de genes. Los métodos escogidos fueron Biograph (33), Cipher (34), DisGeNET (35), Génie (36), GLAD4U (37), Guildify (38), Phenolizer (39), PolySearch (40), y SNPs3D (41). Cada método prioriza una lista de genes con puntuaciones distintas por lo que se desarrolla una estrategia de integración para generar una lista consensus de genes patogénicos para el OS.

En este sentido, la estrategia aplicada para integrar las puntuaciones de genes obtenidas en cada método independiente es similar a la descrita anteriormente (31, 32). Normalizamos cada gen (denotado como i) de la lista clasificada obtenida de cada método (denotado como j) ($GeneN_{i,j}$ que significa la puntuación normalizada del gen " i " en el método " j "). La puntuación final por gen ($ConsenScore_i$) se consideró como la puntuación media normalizada y el número de métodos que predicen el gen (denotado como n_i):

$$ConsenScore_i = \sqrt{\left(\frac{(n_i-1)}{9-1}\right) \left(\frac{1}{j} \sum_j GeneN_{i,j}\right)}$$

Esta ecuación se refiere a la media geométrica entre la puntuación media de cada gen derivada de cada método, y la puntuación normalizada según el número de métodos que predicen la asociación del gen y la enfermedad. Este enfoque de consenso conducirá a una gran lista final de genes clasificados según el $ConsenScore_i$ por lo que también incluimos una estrategia para generar un punto de corte dentro de esta información.

La validación de esta estrategia se realiza a partir de la identificación de genes específicos implicados en la patogénesis del OS. Así, se toma en consideración genes patógenos de OS definidos por una revisión de la literatura a partir de dos tipos de estudios: meta-análisis, basado en publicaciones e informes de casos para pacientes con OS (denominados genes G1), y descripción de genes en modelos animales y líneas celulares OS (denominadas genes G2). Con esta información no solo se valida la tasa de reconocimiento de verdaderos genes, sino que además se utiliza para genera un punto de corte dentro de toda la lista de genes priorizados (31, 32). Los genes patógenos de OS (definidos como G1 y G2) se utilizaron para calcular $I_i = \frac{TP_i}{FP_i+1} ConsenScore_i$, donde TP y FP son los verdaderos positivos y falsos positivos (hasta el

valor de clasificación del gen i) respectivamente. De acuerdo con lo anteriormente descrito (31, 32), el valor máximo de I_i puede tomarse como el compromiso máximo entre las tasas de TP y FP compensadas con el índice de clasificación de cada gen. La clasificación (“i”) en la que I_i es máxima representará un límite racional para la lista de consenso.

1.5. Análisis de enriquecimiento funcional y ontología génica.

La aplicación de análisis de enriquecimiento funcional ha demostrado ser un enfoque eficiente en la priorización de genes, ya que describe procesos celulares e interacciones metabólicas importantes para explicar procesos patogénicos de una enfermedad determinada (42). En este sentido, a partir del grupo de genes priorizados previamente se adiciona en un análisis de interacción proteína – proteína y de vías metabólicas para discutir posibles rutas celulares patogénicas, y además se analiza por ontología génica (GO) los procesos biológicos del grupo de genes priorizados para evidenciar procesos celulares envueltos en la etiología de este sarcoma. El análisis de ontología génica se aplica mediante el uso de David Bioinformatics Resource (43, 44), y se consideran todos los términos relacionados con procesos biológicos con una frecuencia menor al 0.01% utilizando la herramienta bioinformática Revigo (45). Con respecto al análisis de redes, las interacciones de proteínas de los miembros de la lista de consenso se evalúan a partir de la base de datos STRING (46), teniendo en cuenta interacciones con un límite de confianza de 0,9. A partir de esta información, se desarrolla una red de interacción que posteriormente se analiza utilizando el software Cytoscape (47).

Dado que las redes resultantes incluyen vías de señalización complejas, se incluye un análisis de comunalidades para especificar clústeres basales y grupos de genes con relevancia biológica. El análisis de la comunalidad se lleva a cabo utilizando el método de percolación de cliques por medio de Cfinder (48). El análisis de comunalidad proporciona una descripción de la topología de la red, que incluye la ubicación de subgráficos (cliques) altamente conectados y/o módulos superpuestos y que generalmente se corresponden con información biológica relevante. La selección del valor "k-clique" ($k = 1,2,3... n$) afecta el número de comunidades y también el número de genes en cada comunidad. Para determinar el mejor k-clique en el análisis de comunalidad utilizamos el índice “S” (31, 32): $S^k = \frac{|mean(N_g^k) - median(N_g^k)|}{N_c^k}$, donde N_g^k and N_c^k son el número de genes en cada comunidad y el número de comunidades para un valor de corte de k-clique definido. Si la distribución de genes entre las comunidades es cercana a una distribución gaussiana o constante, S^k tenderá hacia 0.

Valores más altos de k-cliques implican pocas comunidades, mientras que valores más bajos conducen a muchas comunidades. Así, para definir las mejores comunidades dentro de una k-clique, se utiliza el algoritmo particional K-means (49). Las variables utilizadas para el cálculo son $ConsenScore_i$ y $Degree_i$ y para cada comunidad dentro de un k-clique. El $Degree_i$ se refiere al índice de centralidad del grado del nodo calculado para cada gen de la red OS-PPI. A partir de las comunidades seleccionadas en este agrupamiento, se crea una subred más específica en donde se visualizan las interacciones de todos los miembros de las comunidades elegidas. Mediante esta estrategia, se reduce el espacio génico y se discuten clústeres patogénicos para el OS.

A partir de estos genes priorizados, se desarrolla una metodología de validación en donde comparamos los clústeres ponderados en el análisis de redes con resultados experimentales derivados del proyecto DRIVE y del portal OncoPPI. El proyecto DRIVE (50) describe un mapeo completo de genes con relevancia oncológica, analizados experimentalmente a partir de 398 líneas celulares cancerígenas en donde mediante un silenciamiento secuencial se asocian fenotipos específicos descritos en cáncer. En esta estrategia de validación, se filtran los resultados de ocho líneas celulares que presentan anotaciones patológicas relacionadas con el cáncer de hueso (A673, SAOS2, SJSA1, SKES1, SKNMC, SW1353, TC71 y U2OS) y se compara con aquellos genes priorizados obtenidos en este trabajo. Adicionalmente, considerando la información publicada en el portal Onco-PPI (<http://oncoppi.emory.edu/>) (51), se genera una red de interacción proteína-proteína centrada en el cáncer, considerando únicamente las interacciones descritas para los tipos de tumores óseos (etiquetados como OncoPPI). Mediante esta estrategia, se valida lo obtenido a partir de la estrategia de priorización consensus y se proponen posibles procesos celulares patogénicos para OS y discuten nuevas vías de señalización oncológica que explican el inicio y desarrollo de este sarcoma.

Por último, tomando en cuenta las redes de interacción proteína - proteína construidas, identificamos todos aquellos factores de transcripción (TF) descritos en la base de datos "The Human Transcription Factors" (52) y analizamos el grado de interacción de cada factor utilizando la información descrita en la Base de datos de regulación de la transcripción genética (GTRD) (53). Esta base de datos contiene información experimental de ensayos ChIP-seq para sitios de unión de factores de transcripción. Los datos se recopilan de forma sistemática y se procesan de manera uniforme mediante un flujo de trabajo especial (canalización) para una plataforma BioUML (<http://www.biouml.org>). Inspeccionando todos los genes diana descritos

para *Homo sapiens*, incluimos la información de todos los genes definidos como factores de transcripción y de todos los genes o +/-5000 pares de bases que contienen un meta cluster GTRD para cada factor. Utilizando este enfoque, analizamos el grado de interacción de todos los factores de transcripción dentro de nuestros genes de consenso propuestos y discutimos posibles mecanismos de regulación a nivel de vías metabólicas.

1.6. Modelos de aprendizaje de máquinas para predicción de fármacos anti-sarcoma

El desarrollo y la validación de nuevos compuestos terapéuticos es un proceso laborioso, que requiere mucho tiempo y recursos económicos, por lo que la propuesta de modelos teóricos que permitan predecir nuevos compuestos tiene un alto impacto dentro de la investigación oncológica. Así, el reposicionamiento de fármacos, en donde se exploran posibles usos novedosos de moléculas conocidas basándose en algoritmos de predicción, es un enfoque eficaz e innovador (54). El segundo componente en este trabajo es el desarrollo de modelos teóricos de predicción, que permitan postular nuevos agentes terapéuticos para el tratamiento del OS. Como primer acercamiento, proponemos un algoritmo en donde se combinan los principios de la teoría de la perturbación (PT) con aprendizaje de máquinas (ML) para desarrollar un modelo PTML de predicción para compuestos anti-sarcoma. Por otro lado, desarrollamos un algoritmo multiobjetivo para la reutilización de nuevos fármacos con posible actividad anti-osteosarcoma, tomando en cuenta moléculas con actividad biológica descritas para líneas celulares HOS, MG63, SAOS2 y U2OS en la base de datos de ChEMBL.

1.6.1. Modelo PTML para predicción de fármacos anti-sarcoma.

Muchos de los compuestos anticancerígenos utilizados hoy en día dentro de terapias oncológicas tienden a tener una alta citotoxicidad y una baja especificidad celular (55). Esto conduce a una menor eficacia dentro tratamientos quimioterapéuticos y una baja tasa de remisión en enfermedad oncológicas. Sin embargo, la descripción de nuevos marcadores moleculares y el desarrollo constante de ensayos preclínicos de fármacos han generado grandes cantidades de datos experimentales (56-59). El uso de estos datos puede conducir a su vez al diseño de fármacos más selectivos, que tengan en cuenta impulsores específicos basados en vías de señalización patógenas. Así, la base de datos química del Laboratorio Europeo de Biología Molecular (ChEMBL) (60, 61) contiene resultados experimentales para más de 37,900 ensayos preclínicos con distintos candidatos a fármacos anti-sarcoma. Estos ensayos cubren una serie de más de 34,900 compuestos químicos. Estas condiciones experimentales incluyen hasta 155 parámetros distintos de actividades biológicas, 36 dianas proteicas, 43 líneas celulares

y 17 organismos de ensayo. En general, esto forma un conjunto de datos grande y complejo susceptible de análisis a fin de extraer conocimientos útiles para el descubrimiento de fármacos.

En este contexto, utilizamos esta información para desarrollar un modelo que considera múltiples condiciones de ensayos al mismo tiempo. Las ideas de la teoría de la perturbación con métodos de aprendizaje automático (modelos PT + ML = PTML) resulta útiles para ajustar conjuntos de datos complejos con características de *big data* en el descubrimiento de fármacos, proteómica, nanotecnología, etc. (62-73). Es interesante notar que las metodologías quimioinformáticas han tenido éxito en el descubrimiento de nuevos candidatos a fármacos eficaces en el laboratorio (74, 75). Sin embargo, muchos algoritmos desarrollados hasta el momento se aplican mayoritariamente a la descripción de fármacos con actividad para tumores del tipo carcinoma, se centran en series homólogas de compuestos con un objetivo o línea celular única (76-83), y tienen un campo de aplicación estrecho ya que se enfocan en un solo conjunto de condiciones (una propiedad específica, proteína objetivo o línea celular).

Los modelos PTML inician con una función de referencia, que mide la probabilidad de que un fármaco esté activo en determinadas condiciones (proteína, línea celular, organismo, etc.), y posteriormente incluyen operadores PT (PTO) para tener en cuenta las perturbaciones (desviaciones) de las variables de entrada de cada fármaco con respecto a una población de fármacos ensayados en las mismas condiciones. Las medias móviles de varias condiciones (MMA) son PTOs similares a los operadores de media móvil de Box-Jenkins. Sin embargo, los MMA son PTO que explican las perturbaciones (cambios) en múltiples condiciones c_j al mismo tiempo, mientras que las medias móviles cuantifican los cambios en solo una condición. Mediante el uso del análisis discriminante lineal (LDA) (84), obtenemos una ecuación PTML-LDA de la siguiente manera:

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot f(v_{ij})_{ref} + \sum_{k=1}^{k_{max}} a_{kj} \cdot D_k + \sum_{k=1, j=0}^{k_{max}, j_{max}} a_{kj} \cdot \Delta D_k(c_j)$$

El modelo genera una puntuación de salida $f(v_{ij})_{calc}$ que se refiere a una función de puntuación para una actividad biológica v_{ij} en las condiciones de ensayo c_j . El algoritmo LDA incluye la métrica de distancia de Mahalanobis que permite inferir valores predictivos mediante un cálculo de probabilidad $p(f(v_{ij})=1)_{pred}$. Para la selección de variables, se detectan perturbaciones específicas dentro de las condiciones c_j que se ajustarán a las propiedades anticancerígenas mediante una estrategia progresiva. Condiciones tales como $c_1 =$ proteína diana, $c_2 =$ línea celular y $c_3 =$ organismo de ensayo fueron significativas, por lo que se toman

en consideración dentro de nuestro modelo. Mediante $p(f(v_{ij})=1)_{\text{pred}}$, se predice la actividad de cada compuesto aplicando la función $f(v_{ij})_{\text{pred}}=1$ cuando $p(f(v_{ij})=1)_{\text{pred}} > 0.5$ o $f(v_{ij})_{\text{pred}}=0$.

Los algoritmos ML se utilizan para establecer la relación entre las entradas y la variable de salida (85). En investigación previas (86-92) se han propuesto modelos similares a PTML para diferentes tipos de cánceres (con énfasis en los carcinomas) como los cánceres de vejiga, próstata, cerebro y mama. Además, Bediaga et al., (93) describe un algoritmo PTML para predecir compuestos anti-cáncer utilizando datos descritos para múltiples tipos de carcinomas al mismo tiempo. Speck-Planche *et al.*, es el único trabajo, hasta nuestro entender, que describe un modelo similar a PTML para la predicción del antisarcoma compuestos utilizando un enfoque de momento espectral (94). De todas formas, no hay reportes de modelos PTML para compuestos anti-sarcoma.

En este estudio llevamos a cabo una compilación, curación y preprocesamiento exhaustivos del conjunto de datos de ChEMBL para ensayos preclínicos de compuestos anti-sarcoma. A partir de estos datos, construimos el primer modelo PTML capaz de ajustar este complejo conjunto de datos con $>$ de 37,900 ensayos y $>$ 34,900 compuestos. Hasta donde sabemos, el estudio supera todos los esfuerzos anteriores en términos de simplicidad del modelo y número de casos, compuestos y líneas celulares consideradas.

1.6.2. Modelo de multi-objetivos.

El modelo multi-objetivo propuesto para reposicionamiento de fármacos con actividad anti-osteosarcoma se aborda a partir de la integración de soluciones potencialmente deseables, construidas a partir de la información de compuestos con actividad biológica para 4 líneas celulares distintas de OS. Para ello, se aplican técnicas computacionales que incluyen la relación cuantitativa estructura-actividad (QSAR) y el cribado virtual basado en ligandos (95-98). Varios de estos estudios se han centrado en la descripción de nuevos agentes terapéuticos, especialmente para el tratamiento de carcinomas (99-103). Sin embargo, muy pocos se han concentrado en tumores de origen mesenquimatoso (94, 104). Así, desarrollamos un modelo multiobjetivo para la predicción de fármacos con potencial actividad biológica frente al OS, uno de los cánceres más prevalentes en poblaciones pediátricas donde la quimioterapia actual no ha variado en las últimas décadas.

Los modelos de predicción individuales se desarrollan a partir de compuestos descritos en la base de datos química (Versión 25) del Laboratorio Europeo de Biología Molecular (ChEMBL) (60, 61) con actividad biológica descrita para las líneas celulares OS HOS

(ChEMBL614736), MG63 (ChEMBL614347), SAOS (ChEMBL614894) y U2OS (ChEMBL615023). Se incluyen todos los valores estándar evaluados por concentración inhibidora media máxima (IC50), porcentaje de inhibición del crecimiento celular a una concentración fija (GI50) y concentración efectiva media máxima (EC50). A partir de estas puntuaciones, se define una clase para cada compuesto, clasificando como activos (1) a todas aquellas drogas con valores estándar $<10 \mu\text{M}$ e inactivos (0) a aquellos con valores $> 10 \mu\text{M}$. Por otro lado, los compuestos que no muestran información sobre su actividad biológica, datos no concluyentes sobre su actividad o información incompleta sobre ChEMBL ID o SMILES canónicos son eliminados del análisis. Las estructuras químicas en formato SMILES de cada fármaco se codifican en el software JChem para Excel (18.8.0.253) (105) de ChemAxon y estandarizan en ChemAxon's Standardizer (106). Se normalizan además quimiotipos específicos como el nitro a una representación única, la aromatización de anillos, la curación de formas tautoméricas, el rayado de sales y pequeños fragmentos, y por último se identifican las estructuras duplicadas utilizando la herramienta EdiSDF dentro de ISIDA / QSPR paquete para posteriormente eliminarse de la data (107).

Los descriptores moleculares son usados para predecir propiedades biológicas y fisicoquímicas de moléculas, en donde se transforma una representación simbólica de una molécula en un número y se posibilita tratamientos matemáticos (108). Utilizando ISIDA Fragmentor 2017 (109, 110), en este trabajo se calculan los siguientes descriptores bidimensionales: secuencias de átomos y enlaces; fragmentos centrados en átomos basados en secuencias de átomos y enlaces; fragmentos centrados en átomos basados en secuencias de átomos y enlaces de longitud fija; y tripletes. Luego de esto, se emplea el algoritmo de relevancia máxima de redundancia mínima (mRMR) (111) para mantener 500 características principales en cada conjunto de datos. Para mRMR, la puntuación del Cociente de Información Mutua (MIQ) se utiliza como una métrica de clasificación de características. Este subconjunto de 500 descriptores moleculares seleccionados se emplea a continuación para el modelado QSAR.

La construcción de los algoritmos de predicción requiere una data balanceada. Con esto, se logra generar modelos que tengan una alta tasa de predicción y no un sobre aprendizaje que posteriormente se puede ver reflejado en un cribado poco robusto. El proceso de balanceo dentro de la data trabajada consiste en aplicar un agrupamiento jerárquico. Para esto, se clacula una medida de intervalo, distancia euclidiana y el método de agrupación de Ward, tanto en compuestos activos como inactivos para cada línea celular. Utilizando el software IBM SPSS Statistics v.25 (IBM Corp., Armonk, NY, EE. UU.), se generan dendrogramas que permiten

definir y visualizar todos los clústeres dentro de los datos. Una vez determinado el número de conglomerados, se procede con una extracción estratificada aleatoria de la misma cantidad de compuesto en ambas clases. Este procedimiento, ya descrito previamente (112), permite identificar datos representativos dentro del espacio de la diversidad química y seleccionar aquellos compuestos con características moleculares similares dentro de los clústeres definidos, por lo que el proceso de equilibrio se desarrolla homogéneamente.

La selección de características (o *Feature selection*) es una estrategia importante que se aplica en conjunto con el desarrollo de algoritmos de aprendizaje de máquinas. En nuestro caso, aplicamos algoritmos genéticos como selección de características considerando una población inicial de 50 cromosomas y 30 generaciones. La validación de la función de *fitness* en el algoritmo genético desarrolla aplicando una estrategia de validación cruzada y utilizando la tasa de clasificación equilibrada promedio (BRC por sus siglas en inglés) en 100 divisiones aleatorias (muestreo de arranque o *bootstrap*). Esto significa que en cada generación se evalúan 100 modelos y se extraen la exactitud promedio. Los modelos utilizados junto con el algoritmo genético son: support vector machine (SVM), random forest (RF), redes neurales (NN), árboles de decisión (DT), k-Nearest Neighbors (KNN), y XGBoost (113). Para las métricas de rendimiento de los modelos se calcula la precisión total (AC), la sensibilidad (SN), la especificidad (SP) y la tasa de clasificación equilibrada (BCR) como se describe previamente (114).

Para el ensamblaje del modelo multi-objetivo se toma en cuenta todos aquellos modelos generados a partir de los compuestos descritos con actividad biológica para cada línea celular con valores de AC > 80% para data externa, y se calcula un valor de deseabilidad global de la siguiente forma:

$$D_1 = (d(y_1)d(y_2)..d(y_k))^{1/k}$$

En donde y_k corresponde a las puntuaciones de deseabilidad de cada línea celular (k 1-4). La predicción de cada modelo para un compuesto determinado da como resultado una puntuación vinculada a la pertenencia a la clase, ya sea una puntuación de predicción para una clase activa o una puntuación para clase inactiva. Por lo tanto, en todos los casos se calcula la media geométrica de todas las puntuaciones obtenidas por compuesto y se utiliza esta medida como una puntuación de deseabilidad para cada modelo y_k . Dado que existen varias combinaciones posibles, se realiza una exploración exhaustiva para obtener el mejor modelo posible. Por lo tanto, exploramos la combinación de todos los modelos posibles en el cálculo

de cada $d(y_k)$ y consecuentemente D_1 con el fin de obtener el mejor rendimiento en métricas de reconocimiento temprano para el cribado virtual.

Además de los estadísticos calculados para evaluar la tasa de predicción de los modelos generados (AC, SN y SP), aplicamos un cribado virtual (VS) para analizar la tasa de predicción del modelo multiobjetivo propuesto utilizando fármacos con actividad terapéutica utilizados en el tratamiento de pacientes con OS. Así, desarrollamos una base de datos con los compuestos antitumorales utilizados en el tratamiento actual del osteosarcoma y compuestos validados en estudios clínicos para OS, y compuestos publicados en el portal web de ensayos clínicos del gobierno de EE. UU. (<https://clinicaltrials.gov/>). Como fármacos terapéuticos de primera línea se incluyen (8, 12, 19, 115): doxorubicina (ChEMBL53463), metotrexato (ChEMBL34259), ifosfamida (ChEMBL1024), etopósido (ChEMBL44657), sorafenib (ChEMBL1336), ciclofosfamida (ChEMBL1336), docetaxel (ChEMBL92), gemcitabina (ChEMBL888), dactinomicina (ChEMBL1554) y vincristina (ChEMBL90555). Además, como fármacos validados en ensayos clínicos: temsirolimus (ChEMBL1201182) (116, 117), ridaforolimus (ChEMBL2103839) (118), sirolimus (ChEMBL413) (119) y pazopanib (ChEMBL477772) (120). Por otro lado, como compuestos inactivos se consideran moléculas retiradas en el proceso de equilibrio de datos (descritos anteriormente) y compuestos ChEMBL comunes para las cuatro líneas celulares que no mostraron actividad biológica (valores estándar $> 10 \mu\text{M}$). Además, se generan moléculas señuelo (*decoys*) basadas en los compuestos activos seleccionados empleando el servidor DUD-E 5 (121). En forma general, se incorporan alrededor de 50 moléculas inactivas para cada compuesto activo, que es la proporción utilizada en la base de datos DUD-E empleada ampliamente para validar los flujos de trabajo de cribado virtual. Por último, el desempeño de los modelos generados dentro de este cribado virtual es evaluado utilizando las métricas AUC (área bajo la curva de acumulación), la discriminación mejorada de Boltzmann de ROC (BEDROC) y la eficiencia de recuperación (EF) dentro del 1% de la lista seleccionada como se describe en trabajos previos (122, 123).

2. OBJETIVOS

2.1. Objetivo general

Desarrollar una estrategia de priorización de genes que describan la patogénesis del osteosarcoma y generar modelos de predicción de fármacos anti-sarcoma y anti-osteosarcoma con posible aplicación terapéutica.

2.2. Objetivos específicos

- Describir una estrategia *consensus* empleando diversos métodos de priorización para identificar grupos de genes asociados con la patogénesis del osteosarcoma.
- Desarrollar un algoritmo PTML para predicción de fármacos con actividad anti-sarcoma.
- Construir un algoritmo de predicción multi-objetivo basado en compuestos con actividad biológica de líneas celulares de osteosarcoma.
- Implementar un modelo de aprendizaje de máquinas para reposicionamiento de fármacos con actividad biológica frente a líneas celulares de cáncer óseo.
- Proponer nuevos fármacos con posible actividad terapéutica para tratamiento de osteosarcoma.

3. RESULTADOS Y DISCUSIÓN

3.1. Uso de técnicas bioinformáticas para priorización de genes patogénicos en osteosarcoma.

La etiología molecular del OS es compleja y heterogénea por lo que hasta el momento no se han podido describir procesos patogénicos específicos para este sarcoma. Este vacío en la descripción de genes involucrados en la patogénesis de la enfermedad, así como la identificación de dianas terapéuticas y biomarcadores, han impedido la propuesta de nuevas metodologías de diagnóstico y el desarrollo de terapias personalizadas. En este sentido, el presente trabajo se ha enfocado en el desarrollo y aplicación de metodologías bioinformáticas que permitan proponer nuevas vías metabólicas inmersas en el proceso de patogénesis de la enfermedad, y además la propuesta de nuevos fármacos que puedan ser utilizados como drogas terapéuticas para el tratamiento del OS.

3.1.1. Métodos de priorización y estrategia consensus.

La metodología de priorización de consenso desarrollada demostró mejorar la tasa de detección de genes patogénicos para OS al compararla con las herramientas bioinformáticas descritas previamente. Para comparar esta tasa de detección, identificamos 75 genes patógenos de OS de la literatura disponible, de los cuales 47 se clasificaron como G1 y 41 como G2. El número de genes patógenos detectados por las nueve herramientas de priorización fue menor que al comparar los detectados con nuestra estrategia de consenso (**Tabla 1**).

Tabla 1. Identificación (%) de genes patogénicos para OS por cada herramienta bioinformática.

Métodos	1%			5%			10%			20%		
	G1	G2	G1-2	G1	G2	G1-2	G1	G2	G1-2	G1	G2	G1-2
BioGraph	0	0	0	0	18.2	12.5	40	45.5	37.5	60	54.6	50
CIPHER	7.7	6.7	8.7	7.7	6.7	8.7	23.1	20	17.4	30.8	26.7	26.1
DisGeNET	9.5	16.7	10.8	21.4	30.6	21.5	42.9	58.3	46.2	57.1	77.8	64.6
Genie	37.8	36.1	35.3	62.2	61.1	57.4	75.6	69.4	70.6	86.7	75	80.9
GLAD4U	0	0	3.6	19.1	33.3	25	42.9	50	46.4	57.1	66.7	64.3
GUILDify	10.9	7.5	8.2	13	7.5	9.6	21.7	17.5	19.2	34.8	25	30.1
Phenolizer	33.3	36.6	30.1	57.8	61	53.4	62.2	61	56.2	77.8	75.6	72.6
PolySearch	0	0	0	11.1	14.3	7.1	11.1	28.6	14.3	11.1	28.6	14.3
SNPs3D	10	10.5	6.3	10	42.1	25	40	57.9	50	75	73.7	71.9
Consensus	66	61	60	87.2	80.5	81.3	89.4	82.9	84	93.6	85.4	88

Al comparar el número de genes patógenos detectados por todas las metodologías, nuestra lista de consenso identifica el porcentaje más alto de genes patogénicos definidos como G1 y G2. Específicamente, en el 1% superior de nuestro método de consenso (las primeras 158 posiciones), se detectaron el 60% de los genes patógenos (45 de 75), seguidos de las metodologías Genie (35,29%) y Phenolizer (30,14%). Además, en el 20% superior, el método de consenso sigue siendo el mejor en la detección de genes patógenos (88%), seguido de Genie, Phenolizer y SNPs3D con porcentajes del 80,88%, 72,60% y 71,88%, respectivamente. Por otro lado, la clasificación media de los genes patógenos detectados en el 1% superior de la lista es 49,3, lo que significa que 45 genes G1-G2 se encuentran en las 50 posiciones superiores. Esta media es superior a la calculada para el resto de metodologías de priorización, dado que el número de genes patógenos detectados es mayor. Sin embargo, es interesante notar que el número de genes y el promedio de clasificación son similares, lo que indica que la mayoría de estos genes patógenos se encuentran en las primeras posiciones. Estos resultados confirman que esta metodología sí mejora la detección y priorización de genes patógenos, como se había descrito previamente en otras patologías (31, 32).

La priorización inicial generó una cantidad de 15.809 genes por lo que luego de la identificación de genes patogénicos G1 y G2, generamos un punto de corte y reducimos esta lista a 553 genes. Los 10 mejores clasificados fueron *TP53*, *RB1*, *CHEK2*, *RUNX2*, *E2F1*, *MDM2*, *CDKN1A*, *JUN*, *CCNA2* y *CDKN2A*. De ellos, *TP53*, *RB1*, *CHEK2* y *MDM2* se clasificaron en las posiciones 1^a, 2^a, 3^a y 6^a, respectivamente y son genes centrales dentro del proceso patogénico del OS. Los primeros estudios centrados en la biología molecular del OS se llevaron a cabo en individuos con síndromes familiares que presentaban un a alta predisposición para desarrollar estos tumores óseos. La inactivación germinal de *RB1* y *TP53* se describió inicialmente en pacientes con retinoblastoma hereditario y síndrome de Li-Fraumeni respectivamente (124, 125), y posteriormente en sarcomas esporádicos (126, 127). Dado que estas proteínas supresoras son centrales dentro del control del ciclo celular, estudios posteriores describieron varias proteínas de interacción. *MDM2*, por ejemplo, es una proteína que se une a *RB1* e inactiva *TP53* (128). Su amplificación es un evento que ocurre en tumores óseos primarios (3–25%) y se sobre expresa en pacientes recurrentes y con metástasis (129, 130). *CHEK2* es otra proteína que forma parte de un punto de control del ciclo celular, funciona como estabilizador de *TP53* y muestra una frecuencia de mutaciones del 7% en pacientes con OS (131, 132).

3.1.2. *Ontología génica, redes de interacciones proteína – proteína y análisis de comunalidades.*

Los procesos biológicos derivados del análisis de GO de los 553 genes describen a TP53 como el principal transductor de señal, que media procesos asociados con el ciclo celular, respuesta al daño del ADN, replicación del ADN y la regulación de la señalización apoptótica intrínseca/extrínseca. Además, se describen procesos biológicos más específicos como por ejemplo la proliferación de fibroblastos, diferenciación y desarrollo de osteoblastos y proliferación y transición de células mesenquimales (**Tabla 2**).

Tabla 2. Procesos biológicos obtenidos por análisis de enriquecimiento en genes OS consensus.

Proceso biológico	Frecuencia	Log10 p-value (FDR)
regulation of signal transduction by p53 class mediator	0.01%	-22.8416
DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0.00%	-20.1656
positive regulation of smooth muscle cell proliferation	0.01%	-16.5544
positive regulation of fibroblast proliferation	0.01%	-16.5031
positive regulation of DNA replication	0.01%	-15.1965
positive regulation of pri-miRNA transcription from RNA polymerase II promoter	0.00%	-14.983
positive regulation of neuron apoptotic process	0.01%	-14.9393
cellular response to mechanical stimulus	0.01%	-13.3507
response to estradiol	0.01%	-11.7258
positive regulation of osteoblast differentiation	0.01%	-11.5058
SMAD protein signal transduction	0.01%	-11.1904
intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator	0.01%	-10.8356
extrinsic apoptotic signaling pathway in absence of ligand	0.01%	-10.0846
somatic stem cell population maintenance	0.01%	-9.6162
response to gamma radiation	0.01%	-9.4056
positive regulation of mesenchymal cell proliferation	0.01%	-9.2628
osteoblast development	0.00%	-9.1046
thymus development	0.01%	-8.2907
vascular endothelial growth factor receptor signaling pathway	0.01%	-7.6946
positive regulation of epithelial to mesenchymal transition	0.01%	-7.6126

De acuerdo con nuestros resultados, estudios previos han identificado procesos biológicos similares, donde los siguientes se consideran términos asociados con el OS: regulación del ciclo celular (mediada principalmente por RB1 y TP53), diferenciación de osteoblastos (mediada por RUNX2), daño del ADN, respuesta al estrés, procesos epigenéticos, mitosis, funciones de motilidad celular y miembros implicados en la proliferación de células OS (ponderando la vía de señalización NF κ B, y las proteínas NF κ BIE y RELA) (133-136). Tomados en conjunto, estos procesos sugieren que nuestra lista de consenso prioriza genes asociados con la osteogénesis, la diferenciación celular y la transición a tipos celulares óseos.

La información utilizada por STRING nos permitió definir el grado de interacción física de los 553 miembros de la lista de consenso. A partir de esta red se calculó el índice de centralidad de los nodos que posteriormente se utilizó como variable para evidenciar la tasa de contribución de los genes patógenos a un propósito biológico común. Teniendo esto en consideración, escogimos todos aquellos con información de interacción y generamos una red de interacción proteína-proteína OS (OS-PPI). A mayor centralidad de un nodo dentro de la red OS-PPI, mayor será la probabilidad de que contribuya a la patogénesis. Esta asociación se validó analizando los genes definidos como patógenos (G1-G2), en los que se observaron diferencias significativas con el resto de genes consenso ($p < 0,0001$).

El índice de centralidad obtenido a partir de los 503 nodos incluidos en la red de interacción proteína-proteína determinó que TP53 fue el nodo más central, seguido de AKT1, MYC, JUN, EP300, CREBBP, CCND1, CDKN1A, STAT3 y RB1. Además, este cálculo permitió definir grupos más específicos y priorizar comunidades de genes asociadas con la patogénesis del OS. Aplicando el análisis de percolación y K-meas, determinamos que el clique k-9 es el grupo que presenta la mejor distribución de genes entre todas las comunidades resultantes (**Figura 1A**), y las comunidades 4, 9, 13 (cluster 1), 5, 8 y 10 (cluster 2) como los grupos de genes más importantes dentro de nuestro estudio (**Figura 1B**).

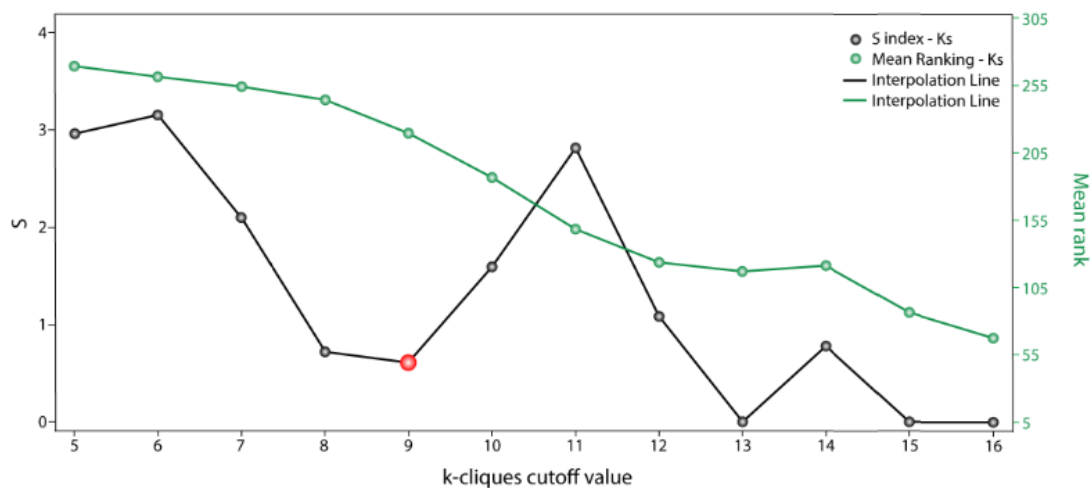
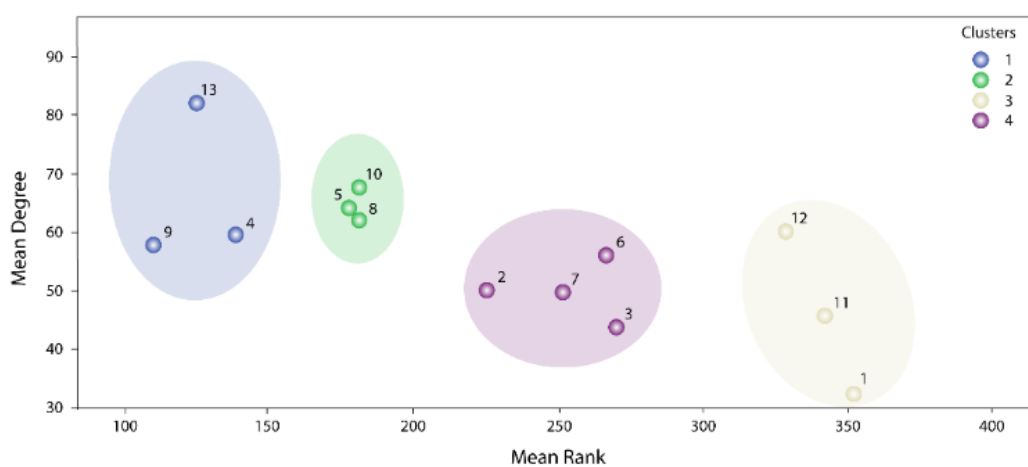
A**B**

Figura 1. K-cliques y análisis de comunalidades. **A)** Análisis de comunalidad por método de percolación de clic. Valores de S^k (puntos negros) y clasificaciones medias (puntos verdes) con respecto a cada valor de corte de k-click. **B)** Análisis de clusters para las comunidades en $k = 9$. Cada círculo de color representa grupos evaluados por K-Means.

El análisis de enriquecimiento en redes metabólicas da como resultado, casi en su totalidad, los mismos términos obtenidos de la lista de consenso inicial. Esto confirma que el gen filtrado a través del análisis de comunalidad comprendía casi los mismos procesos biológicos (**Tabla 3**).

Tabla 3. Análisis de enriquecimiento de rutas metabólicas en las comunidades de $k = 9$ y sus valores asociados.

Ruta metabólica	Valores de enriquecimiento	Comunidades en $k = 9$
p53 signaling pathway	0.603	2, 4, 9, 10
Cell cycle	0.595	2, 4, 7, 8, 9, 13
FoxO signaling pathway	0.578	2, 7, 8, 10, 11, 12, 13
Prolactin signaling pathway	0.574	2, 8, 10, 12
ErbB signaling pathway	0.565	2, 10, 11, 12, 13
Central carbon metabolism in cancer	0.564	2, 10, 11, 12, 13
TGF-beta signaling pathway	0.553	2, 6, 7, 8
Pathways in cancer	0.546	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
VEGF signaling pathway	0.536	2, 10, 11, 12
Adherens junction	0.534	2, 3, 6, 7, 8, 10, 11, 12
Proteoglycans in cancer	0.534	2, 10, 11, 12, 13
HIF-1 signaling pathway	0.532	2, 5, 6, 7, 8, 10, 11, 12, 13
Choline metabolism in cancer	0.526	2, 10, 11, 12
Thyroid hormone signaling pathway	0.524	1, 2, 3, 5, 6, 7, 10, 13
TNF signaling pathway	0.523	2, 5, 8, 13
NOD-like receptor signaling pathway	0.522	2, 8, 13
Osteoclast differentiation	0.52	2, 8, 11, 12, 13
Focal adhesion	0.518	2, 10, 11, 12, 13
Progesterone-mediated oocyte maturation	0.518	2
Apoptosis	0.515	2, 4, 5, 8, 9, 10, 13
Neurotrophin signaling pathway	0.515	2, 5, 10, 11, 12, 13
Fc epsilon RI signaling pathway	0.514	2, 10, 11, 12
MicroRNAs in cancer	0.508	2, 4, 8, 9, 10, 12, 13
mTOR signaling pathway	0.504	2, 10
B cell receptor signaling pathway	0.502	2, 5, 8, 10, 11, 12, 13

La vía de señalización P53 y el ciclo celular resultan en las primeras posiciones del análisis. Asimismo, FOXO mejora su ranking en este análisis de enriquecimiento. En diferentes tipos de cáncer, PI3K/AKT, Ras-MEK-ERK, IKK y AMPK son las vías de señalización más importantes que interactúan con FOXO (137). La ganancia de función de P13K y RAS, o la interrupción de PTEN, son eventos oncogénicos que promueven una pérdida de función en los factores de transcripción de *Forkhead Box* (FOXO) (138). Curiosamente, la pérdida de su expresión promueve una diferenciación osteogénica alterada, lo que sugiere que FOXO1 está implicado en la osteoblastogénesis y la osteoclastogénesis (139-141). Además, los miembros de FOXO tienen un papel importante en la decisión del destino celular, al desencadenar la expresión de ligandos de receptores de muerte como FASLG, ligando de apoptosis de TNF y

algunos miembros de la familia BCL-2 (BCL2L1, BNIP3, BCL2L11) (142, 143). La expresión de FOXO en los tumores OS es baja o incluso inexistente, lo que conduce a la progresión del tumor y la detención del ciclo celular (144). El hecho de que FOXO aumente su ponderación dentro de nuestro análisis de enriquecimiento demuestra su importancia como vía de señalización en la patogénia del OS. Además, la estrecha relación entre la vía de señalización de FOXO y el ciclo celular, eventos de diferenciación de osteoclastos y apoptosis a través de la vía de señalización de TNF, se evidencia en el análisis de enriquecimiento de redes aplicado a la lista de consenso y al clique $k = 9$.

Nuestra estrategia de consenso busca además especificar un grupo de genes que describen la etiología molecular del OS. En este sentido, el uso de todas las metodologías descritas anteriormente prioriza en gran medida los 47 genes dispuestos en las Comunidades 4, 5, 8, 9, 10 y 13 (**Tabla 4**).

Tabla 4. Distribución genética en las comunidades más relevantes en el clique $k = 9$.

Comms	Genes	Media <i>ConsenScore_i</i>	Media <i>Degree</i>	Pathogenic genes/ genes
9	<i>TP53, ATM, BRCA1, CHEK1, CDK2, ATR, BRCA2, RAD51, BLM</i>	0.802	57.78	0.333
13	<i>TP53, JUN, VEGFA, MYC, MMP2, BCL2, MMP9, NFKB1, IL6, FGF2, AKT1, TGFB1, CDH1</i>	0.776	81.85	0.692
4	<i>TP53, CDK4, ATM, BRCA1, CDK2, BRCA2, RAD51, MLH1, BLM</i>	0.751	59.33	0.444
5	<i>TP53, JUN, ATF2, CREBBP, SMARCB1, HMGB1, KAT2B, RELA, ARID1A, NR3C1, SMARCE1</i>	0.68	64	0.182
8	<i>NFKB1, SP1, CREBBP, CEBPB, CEBPD, STAT3, KLF4, EP300, RELA, PPARG, TGFB1</i>	0.675	62	0.273
10	<i>TP53, VEGFA, EGFR, PTK2, ERBB2, SHC1, PTEN, PIK3CA, HRAS, KRAS</i>	0.673	67.4	0.6

De estas seis comunidades, *BRCA1, AKT1, ATR, CDK4, HRAS, MYC, PIK3CA, RELA, STAT3* son genes validados por los datos tomados del proyecto DRIVE y la red Onco-PPI (19,1%), *RAD51, CDK2, CHEK1, SMARCB1, SMARCE1* están validados solo por DRIVE (10,6%), y *ATM, CDH1, EGFR, EP300, ERBB2, JUN, NFKB1, SHC1, TP53, SP1* por Onco-PPI (21,3%). La subred generada a partir de estas comunidades, denominada red de comunidades OS (OS-comms) refleja genes estrechamente interrelacionados a nivel de

interacción celular y también grupos de genes inmersos en importantes procesos oncológicos (**Figura 2**). Tamborero *et al.*, (145) a partir de los datos de secuenciación del exoma de 3205 tumores de la red de investigación Cancer Genome Atlas (TCGA), propone 291 genes onco-driver que actúan sobre 12 tipos diferentes de cáncer. Aunque en este estudio no se tomaron en cuenta datos de muestras de tumores óseos, sus resultados mostraron que varios miembros de la vía de señalización PI3K son onco-dirvers centrales; ATR-BRCA1 actúan como nodos reguladores de procesos de reparación asociados con TP53; CHEK1 y AKT como proteínas principales reguladoras del ciclo celular en función de CDK1A y CDK1B; y FOXO como principal activador *downstream* de vías oncológicas. Estos datos experimentales respaldan nuestros hallazgos, donde PIK3CA, AKT1, PTEN, HRAS y SHC1 son nodos altamente conectados dentro de nuestra red de comunidades OS.

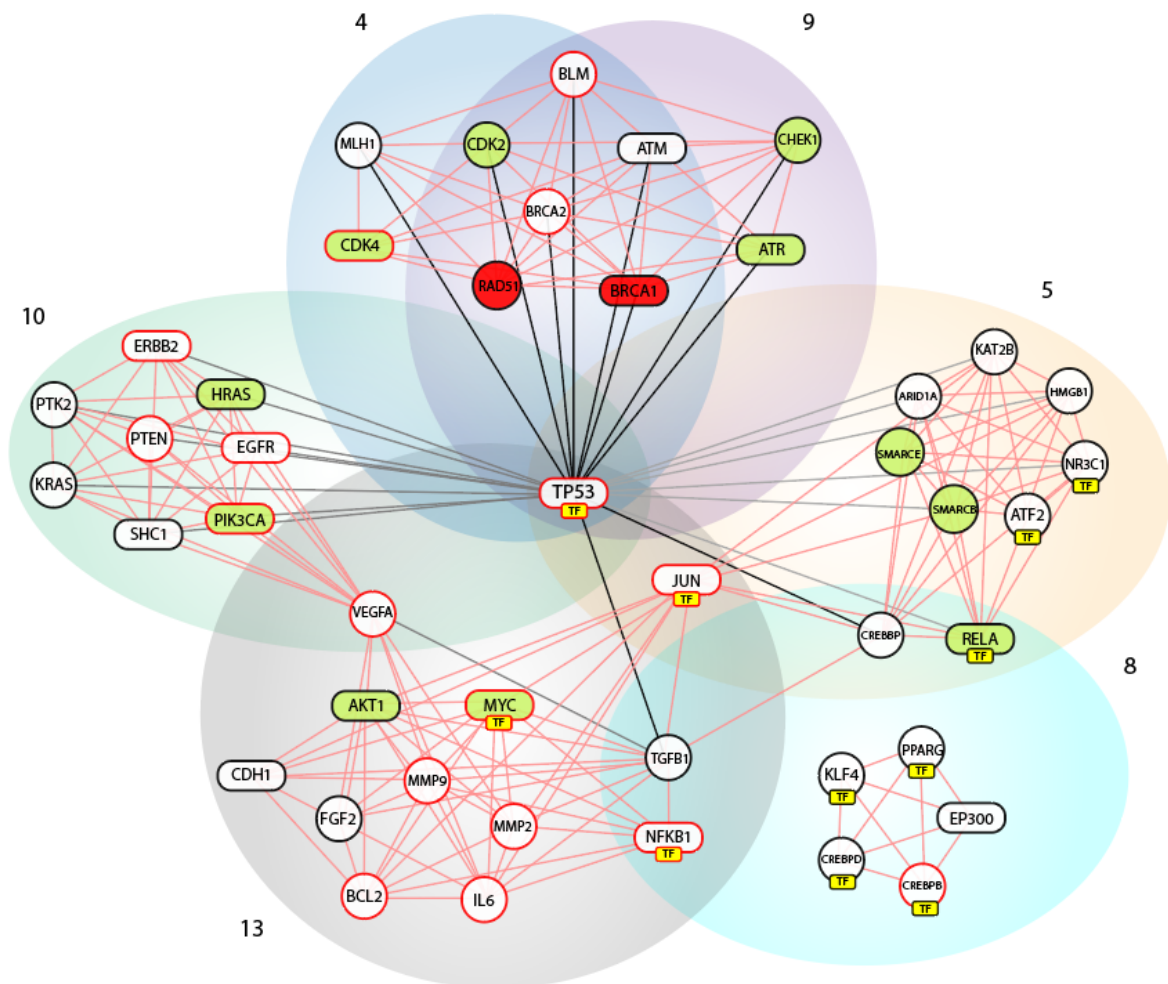


Figura 2. Validación de genes y análisis de redes del clique $k = 9$. Análisis de red de las comunidades 9, 13, 4, 5, 8 y 10 (red OS-comms). Los nodos pintados de rojo y verde se definen como genes esenciales y activos, respectivamente, según los resultados del proyecto DRIVE. Los nodos encerrados en rectángulos pertenecen a la red OncoPPI analizada. Los nodos con

bordes rojos son miembros de G1 y G2. Los recuadros amarillos (TF) apuntan a nodos identificados como factores de transcripción.

Nuestros hallazgos sugieren que PI3K/AKT y MAPK/ERK son las principales vías de señalización desreguladas para el OS. Varios estudios han demostrado que estas vías son responsables de controlar procesos celulares relacionados con la proliferación, el crecimiento, la diferenciación y la apoptosis (146). De hecho, la vía Ras/Raf/MEK/ERK está hiperactivada en el 30% de los cánceres humanos (147) y casi el 67% de los todos los casos con OS muestran una sobre activación de ERK (148). Las quinasas reguladas por señales extracelulares (ERK) promueven la proliferación celular, supervivencia celular y metástasis, particularmente por su activación *upstream* de EGFR y el receptor Ras acoplado a proteína G (149). La presencia de SHC1, EGFR, HRAS, PIK3CA, ERBB2 dentro de la comunidad 10 respalda este escenario para OS. Además, la alta conectividad de las metaloproteasas de matriz MMP2 y MMP9, en la comunidad 13 sugiere un evento de metástasis en la función de estas vías de señalización.

Aunque la invasión de células tumorales es un evento general en la carcinogénesis, la metástasis al pulmón es una de las principales características en pacientes con OS y una de las principales causas de mortalidad (150), por lo que este proceso distintivo de este sarcoma. Acontecimientos patogénicos como el desprendimiento celular de los tumores primarios, la remodelación de la matriz y la invasión de las células tumorales, la angiogénesis, la diseminación vascular y la proliferación en nuevos sitios, están implicados en la metástasis tumoral (151, 152). Reguladores *upstream* de la señalización MAP/ERK como IL6, VEGFA y FGFR1 demuestran un papel importante en este proceso (153-155) y son priorizados en nuestros resultados. Además, la Comunidad 13 muestra los genes MMP2 y MMP9 con un índice de centralidad alto. Esto se describe en ensayos sobre biopsias tumorales en donde se describe una alta expresión de MMP9 en muestras metastásicas de OS (156), lo que llevó a la especulación de que esta metaloproteinasas puede promover migración celular e invasión en tumores OS por componentes de degradación de la matriz extracelular. Esta evidencia sugiere que MMP2 y MMP9, junto con reguladores ascendentes de la señalización MAP/ERK como IL6, FGF2, VEGFA, EGFR y ERBB2, son nodos patógenos dependientes de la centralidad de PI3K/AKT y MAPK/ERK. Este hallazgo podría estar relacionado con aspectos de invasión y pronóstico, principalmente en tumores que presentan desregulación en estas dos vías de señalización.

Mientras que el grupo de genes dentro de la comunidad 13 permiten explicar eventos de migración e invasión tumoral, los genes agrupados en las comunidades 4, 5 y 9 describen procesos de recombinación homóloga (HR), reparación por escisión de bases y modificación

de cromatina. La respuesta al daño del ADN de las células implica principalmente mantener la integridad de los cromosomas y la estabilidad del genoma e implica el reconocimiento de las lesiones del ADN, seguido de una activación de los puntos de control en el ciclo celular que promueve las cascadas de señalización celular relacionadas con la reparación del ADN. Mientras que la vía ATM-CHEK2 es responsable del inicio de las respuestas celulares a las roturas de doble hebra (157), ATR-CHEK1 responde al estrés de la replicación del ADN mediante la fosforilación de varios sustratos en respuesta a agentes como los rayos UV y los rayos X entre otros (158). ATM, ATR y CHEK1 muestran un alto índice de centralidad en la red OS-comms, interactuando además con BRCA1 y RAD51 (descritos como genes esenciales) y con las quinasas dependientes de ciclina, CDK2 y CDK4 (descritas como activas según la validación de DRIVE). La activación de puntos de control por ATM controla principalmente G1/S, mientras que ATM y ATR contribuyen a establecer y mantener los puntos de control S y G2/M (159). Ya sea mediante la activación de ATR-CHEK1 o ATM-CHEK2, la señalización del daño del ADN promueve la inhibición de la actividad de CDK y, por tanto, la activación de los puntos de control G1/S, S y G2/M (160). En consecuencia, es probable que dichos nodos asociados con la reparación del ADN, como ATM, ATR, CHEK1, BLM, RAD51 y MLH1 (como se muestra en nuestro análisis de enriquecimiento de vías metabólicas), junto con los descritos anteriormente (BRCA1 y BRCA2) resultantes de secuenciación exómica (161), tienen importantes implicaciones con respecto a la desregulación del ciclo celular evidenciada en OS.

Si bien es cierto que los nodos descritos para las Comunidades 4 y 9 están relacionados principalmente con eventos de reparación y control del ciclo celular, el complejo de reparación por HR está involucrado en un evento distintivo de los sarcomas denominado mantenimiento alternativo de telómeros (ALT). Aún se desconocen varios detalles moleculares de este mecanismo; sin embargo, se describen dos fenotipos de telómeros distintivos para ALT en células humanas negativas a la telomerasa (células ALT): 1) la presencia de un ADN telomérico largo y heterogéneo y 2) el cuerpo PML (162), que juntos forman el cuerpo PML asociado con ALT (APB). El cuerpo de la PML es un núcleo compuesto por proteínas que se forman entre la cromatina y está relacionado con una amplia gama de procesos celulares, incluyéndose la formación de tumores, la senescencia celular y la reparación del ADN (163, 164). Numerosas líneas de evidencia sugieren fuertemente que la vía ALT depende de la HR ya que varias proteínas involucradas en la rotura de la doble hebra del ADN (DSB) están localizadas en APBs (165, 166). Es significativo que las proteínas localizadas en APB, como PML, helicasas de ADN de la familia RecQ (BLM, WRN y RECQL4), RAD51 y RAD52 (un miembro del complejo MNR), ocupan un lugar destacado en nuestra priorización. En este sentido, los

miembros que pertenecen a los complejos HR se describen como complejos de reparación en respuesta al daño del ADN. Son relevantes para la patogénia del OS no solo como factores inmersos en el control del ciclo celular, como se discutió previamente, sino que además porque están involucrados en procesos de estabilidad cromosómica dada por el mantenimiento de los telómeros (167-169). De acuerdo con la literatura, donde los tumores óseos se denominan muy heterogéneos, altamente mutables y genéticamente inestables, los miembros descritos en las Comunidades 4 y 9 (TP53, ATM, ATR, CHEK1, BLM, BRCA1, BRCA2, RAD51, MLH1, CDK2, CDK4) explican muchas de estas características clave dentro del OS y también pueden estar asociadas con características clínicas importantes como la agresividad del tumor, la metástasis y la supervivencia reducida en pacientes.

El uso de la base de datos GTRD nos permitió definir la frecuencia de interacción de cada TF con los 553 genes priorizados. Vale la pena señalar que más de la mitad de los factores priorizados (82,4%) interactuaron con más de la mitad de todos los genes al mismo tiempo. Esto sugiere que más del 80% de los genes definidos como TF regulan activamente los genes asociados con la patogénia del OS. El peso dado a cada uno de estos TF mediante análisis de interacción coloca a los siguientes genes en las primeras posiciones: *TP53*, *E2F1*, *JUN*, *RUNX2*, *FLII*, *YY1*, *HIF1A*, *MYC*, *TP63*, *ESR1*, *WT1*, *E2F4*, *ATF2*, *NFKB1*, *AR*, *SP1*, *STAT1*, *ERG*, *CEBPB* y *TFAP2A*. En comparación con la priorización total, los genes *E2F1*, *JUN*, *RUNX2*, *FLII*, *YY1*, *HIF1A*, *MYC*, *TP63*, *ESR1*, *WT1*, *E2F4*, *ATF2* y *NFKB1* mejoraron significativamente su clasificación. Durante la fase G1 del ciclo celular, RB1 suprime la función de los factores de transcripción E2F1, E2F2 y E2F3. La hipo-fosforilación secuencial de RB1 por quinasas dependientes de ciclina, CDK4 y CDK6, y CDK2, conduce a la liberación de E2F y la transcripción de genes necesarios para la progresión del ciclo celular, incluidas las ciclinas A, D y E (170). La ponderación mejorada de estos TF sugiere que estos eventos de desregulación en el ciclo celular son basales dentro de la patogénesis del OS. Aunque este escenario es común para todos los tipos de cáncer, sería necesario un estudio más profundo de los genes E2F1 y E2F4, y en función de los priorizados en las Comunidades 4 y 9 junto con TP53, para definir proteínas driver en tumores OS.

Muchos de los factores de transcripción priorizados se agruparon en las Comunidades 5, 8 y 13. Con TP53 como nodo central, JUN y MYC son proteínas clave en la patogénesis del OS que regulan vías de señalización asociadas con las rutas patogénicas PI3K/AKT y MAPK/ERK. Además, la priorización de TFs evidenció a NFKB1 como un nodo central en estas tres comunidades. El factor nuclear kappa B1 (NFB1) es un factor de transcripción pleiotrópico que contribuye con la tumorigénesis en varios tipos de cáncer. Funciona como un regulador clave

de una variedad de genes implicados en muchos eventos biológicos, incluyéndose supervivencia celular, diferenciación, apoptosis y autofagia (171). Al observar la red OS-comms, el alto grado de interacción de AKT con respecto a JUN-MYC, TGFB1, NFKB1 y BCL2 sugiere que este grupo es un grupo importante en la patogénesis de OS. Los términos de GO obtenidos concuerdan con estos hallazgos, ya que su activación promueve muchos tipos de señalización descendente, incluida la diferenciación de osteoblastos a través de TGFB1 y NFK1 o apoptosis a través de BCL2 (172, 173).

3.2. Modelos de aprendizaje de máquinas y reposicionamiento de fármacos en osteosarcoma.

Este análisis desarrollado sistémico desarrollado en la priorización de genes proporcionó una visión específica con eventos patogénicos para el OS, permitió especificar procesos biológicos importantes que se desarrollan dentro de la tumorigénesis y además priorizó genes clave que aclaran de mejor manera la etiología molecular de este sarcoma. El enfoque aplicado en este trabajo busca no solamente proponer nuevos biomarcadores, sino que además busca desarrollar modelos de inteligencia de máquinas que permitan proponer nuevas drogas terapéuticas. Así, la segunda fase de este trabajo describe y discute dos algoritmos de predicción: 1) modelo PTML construido a partir de drogas anti-sarcoma y 2) modelo de multiobjetivos construido a partir de fármacos con actividad biológica reportados para líneas celulares de OS y su aplicación en un procedimiento de reposicionamiento de fármacos.

3.2.1. Modelo de aprendizaje de máquinas teoría de la perturbación (PTML) en sarcomas.

El modelo PTML fue construido a partir de la información de > 37,0000 resultados de ensayos preclínicos para candidatos a fármacos contra el sarcoma descritos en la base de datos ChEMBL. La actividad biológica de cada fármaco fue definida por sus valores estándar, y los descriptores moleculares utilizados fueron $D_1 = \text{LogP}$ y $D_2 = \text{PSA}$, los mismos pre-calculados por este portal. El conjunto de datos final después de una curación consistió de 37,919 casos que comprendían 36 dianas proteicas, 43 líneas celulares y 17 organismos de ensayo. Así, el modelo PTML-LDA resultante resultó en la siguiente fórmula:

$$f(v_{ij})_{calc} = -11.8545 + 34.8028 \cdot f(v_{ij})_{ref} + 0.37 \cdot D_1 - 0.0128 \cdot D_2 - 0.3616 \cdot [D_1 - \langle D_1(c_j) \rangle] + 0.0191 \cdot [D_2 - \langle D_2(c_j) \rangle]$$

$$n = 34955 \quad \chi^2 = 16848.08 \quad p < 0.001$$

Los estadísticos obtenidos en nuestro modelo mostraron una alta especificidad (SP) y sensibilidad (SN) para la serie de entrenamiento (95,63 y 79,64, respectivamente). Además, se obtuvieron valores similares para SP (95,79) y SN (81,62) en los conjuntos de validación (Tabla . Además, el nivel de significancia fue menor <0.05 ($\chi^2 = 16848.08$), lo que indica que el modelo es capaz de realizar una partición estadísticamente significativa de ambas clases. También es interesante observar la alta precisión global (AC) obtenida en ambos conjuntos es mayor al 94%. Estos resultados sugieren que el modelo generado realiza una clasificación estadísticamente significativa de compuestos anti-sarcoma, por tanto, puede considerarse útil para modelos de clasificación con aplicación en química médica.

Tabla 5. Resultados estadísticos del modelo PTML.

Series	Parámetros estadísticos	Etadística predicha (%)	Conjuntos observados	Conjuntos predichos	
				$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
Entrenamiento	SP	95.63	$f(v_{ij})_{obs} = 0$	25647	1172
	SN	79.64	$f(v_{ij})_{obs} = 1$	330	1291
	AC	94.72	Total	25977	2463
Validación	SP	95.79	$f(v_{ij})_{obs} = 0$	8559	376
	SN	81.62	$f(v_{ij})_{obs} = 1$	100	444
	AC	94.98	Total	8659	820

3.2.2. Moldeos PTML generados en investigación oncológica.

Los parámetros ALOGP y PSA son ampliamente utilizados en química médica dado que están relacionados con la lipofilidad de los fármacos y, en consecuencia, con su capacidad para atravesar membranas biológicas o interactuar con bolsas hidrófobas de proteínas (174-176). El algoritmo PTML se ha aplicado previamente al estudio de múltiples ensayos preclínicos de fármacos contra el cáncer y principalmente sobre carcinomas. Por ejemplo, Speck-Planche *et al.*, describen modelos similares a PTML para cáncer de vejiga (87), colorectal (89), de mama (90), próstata (177) y para múltiples subtipos de carcinoma (93). Además, se han probado modelos similares a PTML en agentes anti-tumorales cerebrales (88). Curiosamente, Bediaga *et al.*, demuestran la aplicación de un PTML en varios tipos de carcinomas simultáneamente en donde obtienen valores de SN y SP similares a los obtenidos

en este trabajo (> 90%) (93). Todos estos modelos similares a PTML son capaces de explicar cambios en proteínas diana, líneas celulares, organismos, etc.; sin embargo, son modelos específicos para carcinomas, no para sarcomas.

Vale la pena señalar que Speck-Planche *et al.*, (94) parecen ser los únicos investigadores que han descrito un modelo anterior similar a PTML para sarcomas hasta el momento. En su estudio, el modelo de predicción en validación externa resultó en valores de AC (90,78) y SP (90,65) inferiores a los obtenidos en nuestro modelo (AC = 94,98 y SP = 95,79). Sin embargo, nuestro algoritmo PTML mostró una tasa menor de sensibilidad en los datos de validación externa (81,62%) al compararlo con el modelo obtenido por Speck-Planche *et al.* (91,74%). Incluso cuando nuestro modelo tenía un número mucho menor de variables y usaba una definición de corte más estricta para la clase de actividad (es decir, IC50 = 0.1 uM en lugar de 1 uM), estos aspectos por sí solos no pueden explicar la reducción de sensibilidad. El modelo PTML-LDA generado tiene características importantes que permiten su uso en investigaciones enfocadas al descubrimiento de fármacos. Una de las principales ventajas de nuestro modelo es la considerable reducción de variables de entrada para la construcción del algoritmo mediante la inclusión de PTOs. Esta reducción nos permitió trabajar en conjuntos de datos con una gran cantidad de información, definir valores de corte y calcular la probabilidad de pertenecer a una clase, ya sea una predicción para compuestos activos (1) o compuestos inactivos (0). De esta forma, los valores de SN o SP del modelo se pueden ajustar de acuerdo con los cortes delimitados. Un modelo de predicción ideal tiene una compensación razonable entre SN y SP. Esto significa que se logra una alta sensibilidad aceptando una SP relativamente baja y, a la inversa, se alcanza una SP alta comprometiendo SN. SN es sinónimo a tasa de verdaderos negativos, que está relacionada con la tasa de falsos positivos (178), por lo que una alta especificidad en un modelo de predicción para el descubrimiento de fármacos implica que es poco probable que se obtenga un resultado positivo en un fármaco que no tiene un efecto biológico deseado actividad. Por lo tanto, un resultado positivo en un modelo específico es bastante informativo en un escenario de descubrimiento de fármacos. Por otro lado, un atributo principal del modelo PTML es la posible combinación de varias condiciones experimentales para la predicción de nuevos compuestos. En este sentido, Speck-Planche *et al.*, (94) utilizaron alrededor de 3000 interacciones derivadas de 14 líneas celulares y solo consideraron ensayos de IC50 para su modelo. En contraste, en este trabajo modelamos 37919 casos de interacciones que comprenden 36 dianas proteicas, 43 líneas celulares y 17 organismos de ensayo. La tarea de modelado que tenemos es más compleja, no solo por el incremento en la diversidad química, sino que además por la alta heterogeneidad en las interacciones (es decir, tipos objetivo,

organismos). Los dos modelos no se pueden comparar en este escenario y nuestra reducción en la capacidad de detectar los casos positivos verdaderos (SN) podría ser una consecuencia de esta complejidad de datos y también de la estrategia de modelado.

En la construcción del modelo PTML implementamos un punto de corte (cut-off) estricto para la definición de activos o inactivos dentro del modelamiento, esto con el objetivo de generar un modelo eficiente al momento de predecir fármacos a ser probados en el laboratorio. Un valor restringido para la selección de verdaderos activos promueve una alta certeza en la predicción de compuestos activos para lograr una acción biológica deseada en múltiples condiciones de prueba (179, 180). Además, un límite estricto puede disminuir la tasa de falsos positivos previstos, por lo tanto, si se va a implementar el ensayo, necesita una mayor sensibilidad o una mayor especificidad; este valor puede modelarse dependiendo de las condiciones experimentales que se desee aplicar. Este valor de corte también influye en la precisión dentro de nuestro modelo. Al aumentar el rigor, el modelo mejoró sus valores de predicción para los compuestos activos (1) (**Figura 3**). Al observar estos resultados, nuestro algoritmo de predicción tiene en cuenta no solo varias condiciones experimentales, sino que también restringe la predicción de compuestos a aquellos que tienen verdadera actividad biológica.

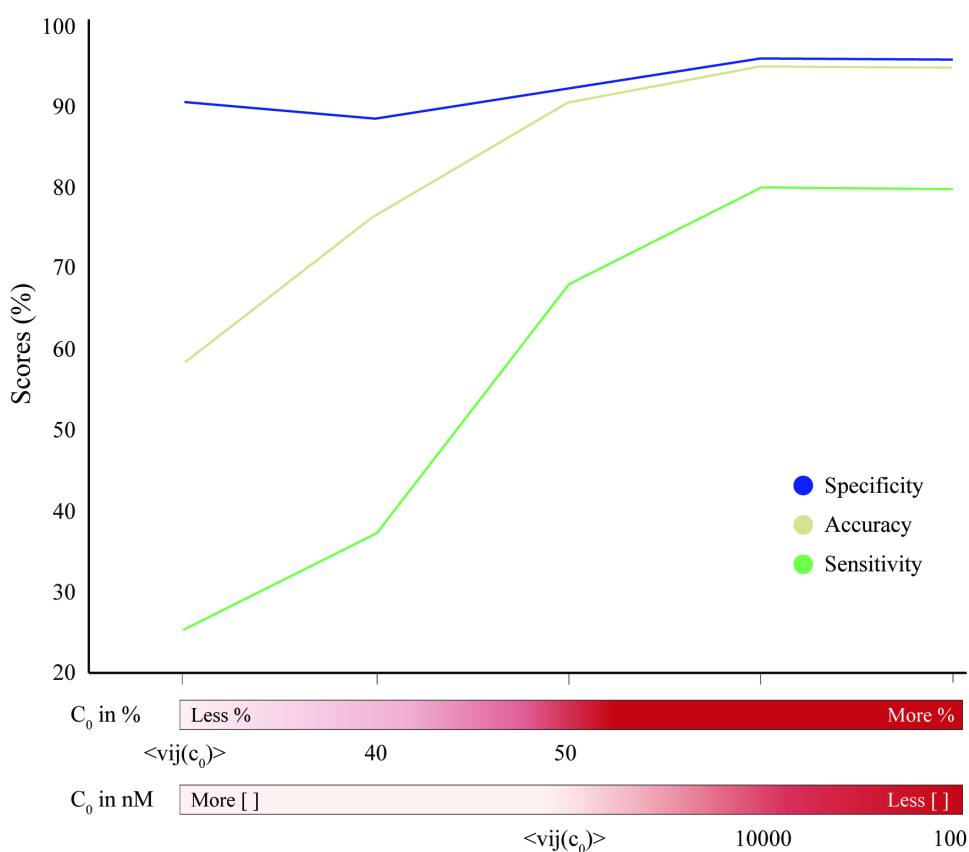


Figura 3. Variación de los valores de SP, SN y AC según los cut-offs implementados. La variación de estas puntuaciones en función de las actividades biológicas c_0 se incluye en el eje x. Se describen las actividades biológicas c_0 expresadas en porcentajes (P.E: inhibición, actividad, inhibición del crecimiento tumoral, etc.) y las expresadas en nM (PE: potencia, IC50, CC50, etc.). El modelo final se obtiene aplicando valores de corte de 50 para c_0 expresado en % y 100 para c_0 expresado en nM.

3.2.3. Comparación entre modelos PTML y modelos ML.

La mayoría de los métodos de aprendizaje ML multitarea o de etiquetas múltiples son útiles para predecir múltiples salidas categóricas para el mismo conjunto de variables continuas de entrada (181, 182). Sin embargo, nuestro problema es un poco diferente dado que se propone desarrollar un modelo ML con solo dos salidas posibles para el mismo conjunto de variables de entrada. Eso significa que nuestro algoritmo no es un modelo multitarea para un solo caso con un conjunto de variables de entrada que contienen múltiples variables continuas y múltiples categorías. En nuestro modelo, tenemos múltiples combinaciones de niveles o variables categóricas de entrada para el mismo conjunto de variables continuas de entrada. Por lo tanto, el modelo PTML es de etiquetas múltiples en las variables categóricas de entrada para el mismo conjunto de variables continuas de entrada. Para ilustrar este hecho, desarrollamos una comparación de nuestro modelo PTML-LDA con un algoritmo ML clásico usando múltiples variables categóricas de etiquetado (**Figura 4**). Así, calculamos 12 descriptores moleculares BCUT (183) con ChemAxon (<http://www.chemaxon.com>). Se calcularon los descriptores de carga clásicos no ponderados, así como los ponderados por las propiedades de carga y enlace de hidrógeno. Para el cálculo de los descriptores se utilizaron los valores propios más bajos y los tres más altos.

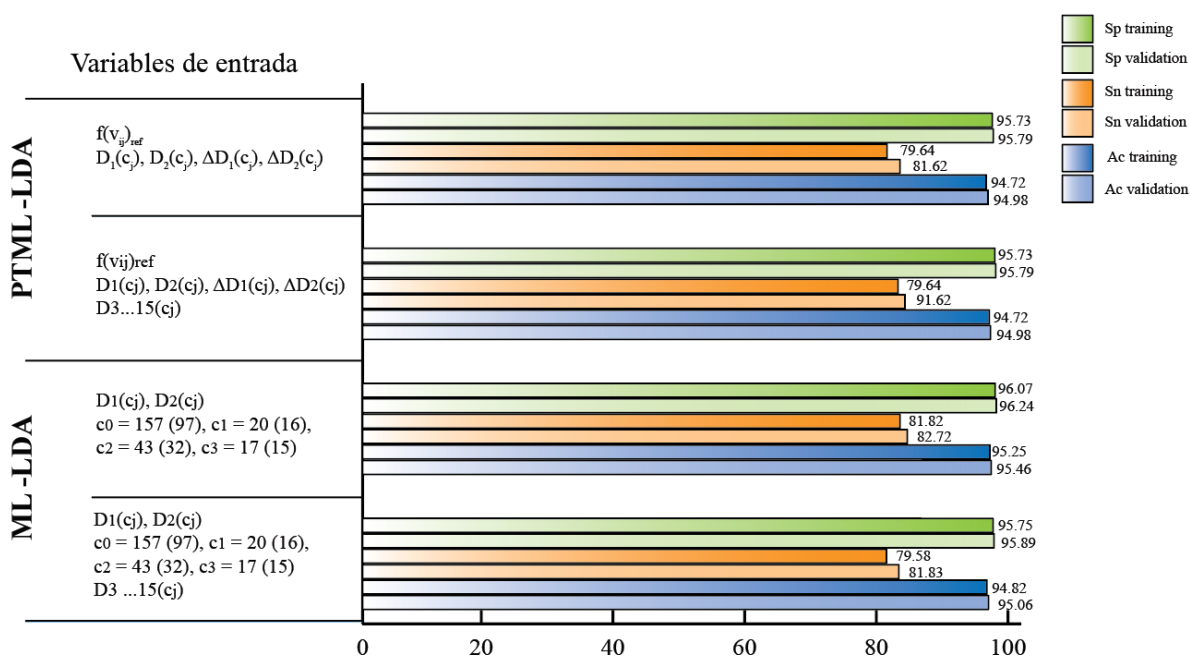


Figura 4. Comparación entre modelo PTML y modelo ML. Valores de predicción de los modelos PTML-LDA y ML-LDA utilizando diferentes tipos de variables de entrada: $f(v_{ij})_{pred}$ función de referencia, $D_1(c_j)$ y $D_2(c_j)$ son descriptores ALOGP y PSA respectivamente, $\Delta D_1(c_j)$ and $\Delta D_2(c_j)$ representan a las desviaciones de los descriptores moleculares de ALOGP y PSA respectivamente, $D_3...D_{15}(c_j)$ son los 12 descriptores moleculares de BCUT calculados a partir de ChemAxon. A diferencia del modelo PTML, el modelo ML se calcula con cada condición c_1 , c_2 y c_3 como un conjunto separado de variables categóricas.

El rendimiento del modelo PTML-LDA en comparación con un ML-LDA clásico demuestra valores similares basados en SP, SN y AS. De manera similar, al desarrollar redes neuronales (NN) los resultados de PTML-NN y ML-NN son bastante similares. Una de las ventajas de nuestro modelo PTML es la inclusión de PTOs, lo que reduce en gran medida el número de variables para generar el algoritmo. Así, aunque las estadísticas de todos los modelos generados son bastante similares, la metodología PTML permite reducir las variables de 164 variables en los métodos clásicos de ML a solo 5 en el modelo PTML (**Figura 5**).

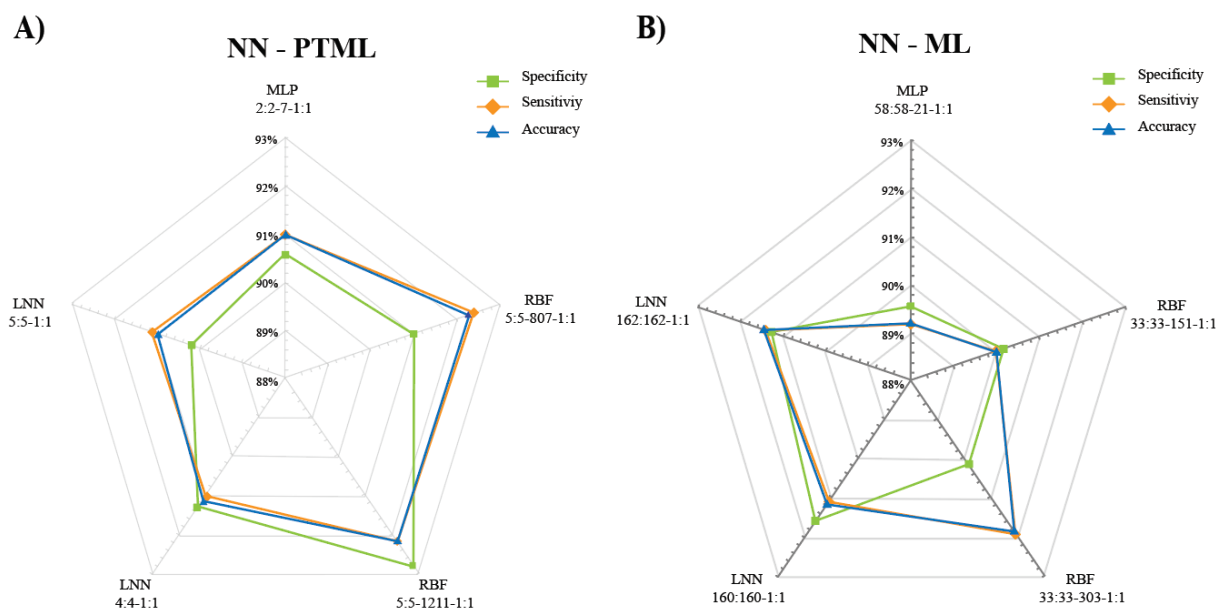


Figura 5. Comparación en el rendimiento de modelos NN-PTML y NN-ML. **A)** Valores de predicción entre los modelos de red neuronal-PTML (NN-PTML) y **B)** NN-ML. Los NN obtenidos fueron perceptrón multicapa (MLP), red neuronal lineal (LNN) y red de función de base radial (RBF).

3.2.4. Modelo multi-objetivo de predicción para fármacos anti-sarcoma.

El reposicionamiento de fármacos es una estrategia eficaz para encontrar nuevas relaciones fármaco-enfermedad para moléculas existentes por lo que utilizamos este enfoque como último punto de estudio en este trabajo. El reposicionamiento de fármacos ha ganado un interés considerable en los últimos años en comparación con estrategias *de novo*, que exigen más tiempo de investigación y horas experimentales en el caso del desarrollo de nuevos fármacos, y requiere una mayor inversión financiera. Por otro lado, el uso de fármacos ya probados demuestra ser altamente eficiente, de bajo costo y de bajo riesgo ya que el cribado se realiza en moléculas que han pasado todas las pruebas de seguridad clínica en la Fase I, Fase II y Fase III (184, 185). Así, desarrollamos un modelo de multi-objetivos en donde buscamos construir un algoritmo a partir de la información de compuestos con actividad biológica para líneas celulares específicas de cáncer de hueso. Luego de esto, aplicamos un procedimiento de reposicionamiento y proponemos nuevos fármacos con posible actividad terapéutica para el tratamiento del OS.

Varios estudios han demostrado que los modelos multiobjetivo tienen una mejor tasa de predicción durante el tiempo de detección, ya que abordan el problema con una perspectiva particular desde un conjunto de soluciones potencialmente deseables (98, 114, 186, 187). En

nuestro caso, cada una de estas posibles soluciones deseables se compone de cada algoritmo construido a partir de los compuestos descritos con actividad para las líneas celulares OS HOS, MG63, SAOS2 y U2OS. Al comparar el modelo multi-objetivo con cada modelo base desarrollado en un escenario de VS, obtenemos una mejora considerable en los valores de AUC y BEDROC en el VS, especialmente en los valores de EF al 1% (**Figura 6**).

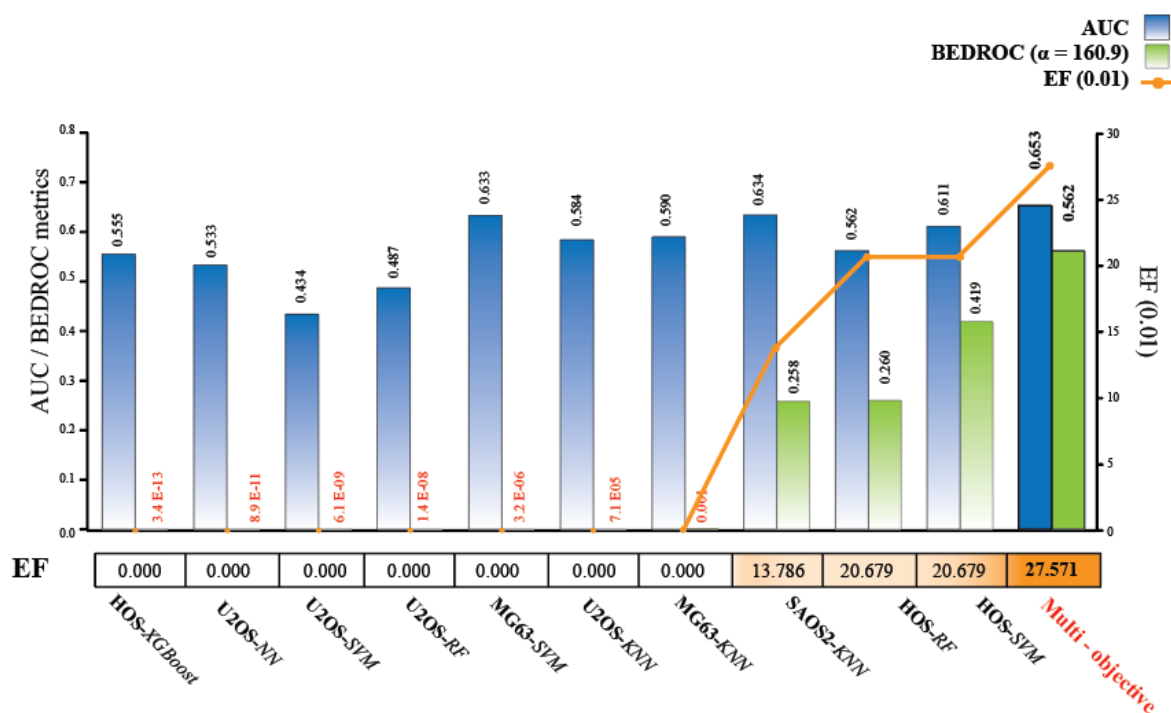


Figura 6. Resultados del desempeño de modelos base y modelos multiobjetivo en el cribado virtual (VS). Comparación de los valores AUC (barras negras) y BEDROC con $\alpha = 160,9$ de los modelos base y el algoritmo multiobjetivo.

Esto sugiere que nuestro algoritmo mejora la tasa de reconocimiento de moléculas descritas como terapéuticas para el tratamiento de la OS, especialmente dentro del 1% de los datos examinados. Específicamente, la EF obtenida indica que es posible recuperar en el primer 1% de una lista cribada casi 27 veces más compuestos multidireccionales de lo que se espera de una distribución uniforme de los activos en la base de datos de cribado virtual, algo que es no obtenido de los algoritmos generados por cada línea celular.

3.2.5. Cribado virtual y reposicionamiento de fármacos para osteosarcoma.

Dada la alta tasa de recuperación de compuestos activos obtenidos en nuestro modelo (EF 0,01 = 27,571), desarrollamos un cribado virtual sobre 2218 medicamentos aprobados por la FDA reportados en el DrugBank y consideramos los primeros 22 compuestos de mayor rango pertenecientes al 1% de los 2218 (**Tabla 6**).

Tabla 6. Fármacos repositionados por modelo multiobjetivo.

Score D1	Drug Bank ID	Fármaco	CTs en cáncer ^a	CT en pacientes OS ^b
0.8683	DB06287	Temsirolimus	67	10
0.8659	DB01229	Paclitaxel	1073	3
0.8618	DB00877	Sirolimus (Rapamicine)	126	10
0.8584	DB01590	Everolimus	195	2
0.8554	DB06772	Cabazitaxel	61	0
0.8506	DB01248	Docetaxel	567	8
0.8430	DB00602	Ivermectin	0	0
0.8416	DB01045	Rifampicin	1	0
0.8322	DB00864	Tacrolimus	183	0
0.8202	DB00337	Pimecrolimus	0	0
0.8016	DB00778	Roxithromycin	0	0
0.8003	DB00932	Tipranavir	0	0
0.7974	DB01211	Clarithromycin	65	0
0.7957	DB00595	Oxytetracycline	0	0
0.7949	DB00199	Erythromycin	3	0
0.7938	DB01319	Fosamprenavir	0	0
0.7921	DB01201	Rifapentine	0	0
0.7913	DB00254	Doxycycline	16	0
0.7876	DB00759	Tetracycline	2	0
0.7844	DB13179	Troleandomycin	0	0
0.7803	DB01017	Minocycline	8	0
0.7792	DB11431	Moxidectin	0	0

CT, Número de ensayos clínicos descritos en: ^apacientes con cáncer en general, y en ^bpacientes diagnosticados con osteosarcoma.

De estos 22 fármacos, 13 (59,1%) están inscritos en ensayos clínicos para pacientes con cáncer (revisados en <https://clinicaltrials.gov/>): temsirolimus, paclitaxel, sirolimus/rapamicina, everolimus, cabazitaxel, docetaxel, rifampicina, tacrolimus, claritromicina, erilitromicina, doxiciclina, tetraciclina y minociclina. Curiosamente, solo cinco de estos fármacos se incluyen en ensayos de pacientes con OS: temsirolimus, paclitaxel, sirolimus/rapamicina, everolimus y docetaxel. Los 10 fármacos restantes (ivermectina, pimecrolimus, roxitromicina, tipranavir, oxitetraciclina, fosamprenavir, rifapentina, troleandomicina y moxidectina) no están registrados en ningún ensayo clínico para pacientes con cáncer, sin embargo, sus mecanismos de acción son similares a varios agentes quimioterápicos utilizados en la práctica oncológica.

El cribado ponderó cuatro clases principales de fármacos en el primer 1% de la lista de cribado: antiinfecciosos para uso sistémico (antimicobacterianos, macrólidos, inhibidores de proteasa y tetraciclinas; que representan el 55%); agentes antineoplásicos/inmunomoduladores (inmunosupresores, inhibidores de proteína quinasa y taxanos; 32%); dermatológico /

inmunosupresor (agentes para la dermatitis, excluidos los corticosteroides; 4%); y antiparasitarios (un agente antinematodal y un endectocida de amplio espectro; 9%). Los dos primeros grupos representan más del 85% de todos los fármacos reposicionados (**Figura 7**).

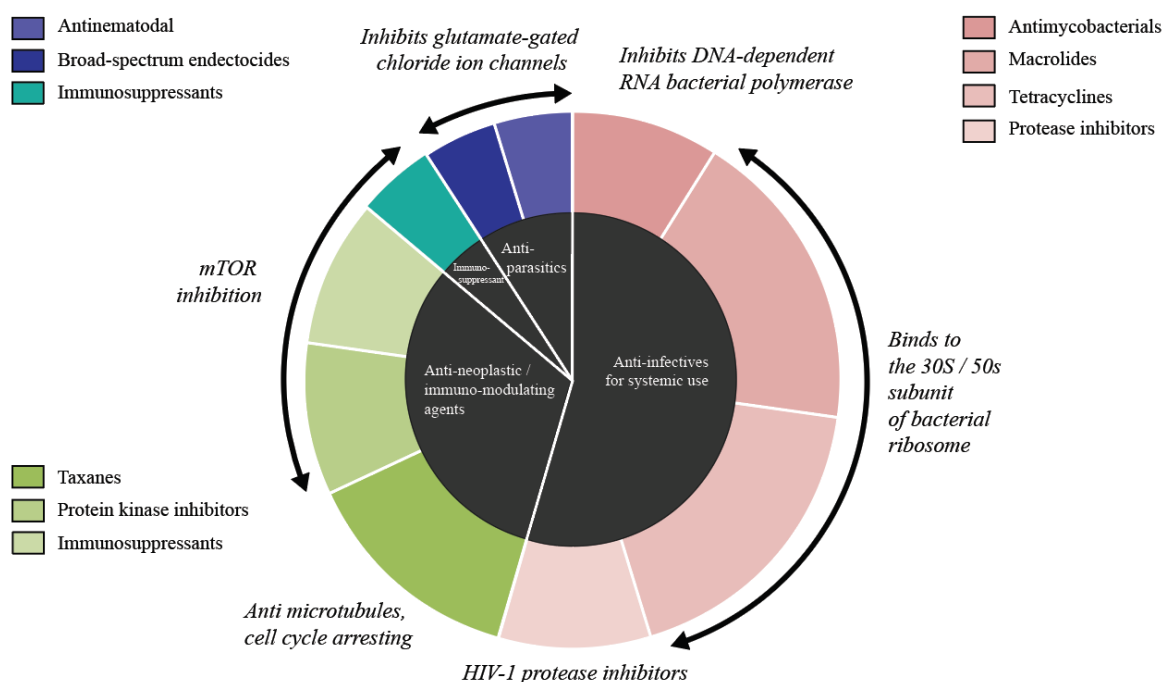


Figura 7. Fármacos reposicionados para el tratamiento del OS. El gráfico circular central (negro) muestra la distribución de las 4 clases principales de medicamentos reposicionadas en el primer 1% de la lista seleccionada, mientras que el gráfico circular exterior muestra los grupos que representan. Cada color representa un grupo específico obtenido del sistema de clasificación anatómico terapéutico químico (ATC). Los mecanismos de acción de los fármacos seleccionados también se incluyen en cursiva.

El mecanismo de acción de agentes antineoplásicos e inmunomoduladores inhibe principalmente la vía mTOR y la polimerización de los microtúbulos. En este grupo obtuvimos a los fármacos temsirolimus, paclitaxel, sirolimus (rapamicina), everolimus, cabazitaxel y docetaxel como los mejores puntuados. Por otro lado, los antibacterianos de amplio espectro descritos como fármacos que se unen a la subunidad 30S / 50s del ribosoma bacteriano, los inhibidores de la proteasa del VIH-1 y los antimicobacterianos, que inhiben la polimerasa bacteriana del ARN dependiente de ADN, se ponderaron en el cribado. También encontramos dos moléculas utilizadas para el tratamiento del VIH, descritas como inhibidores de la proteasa del VIH-1 (tipranavir y fosamprenavir) (**Figura 8**).

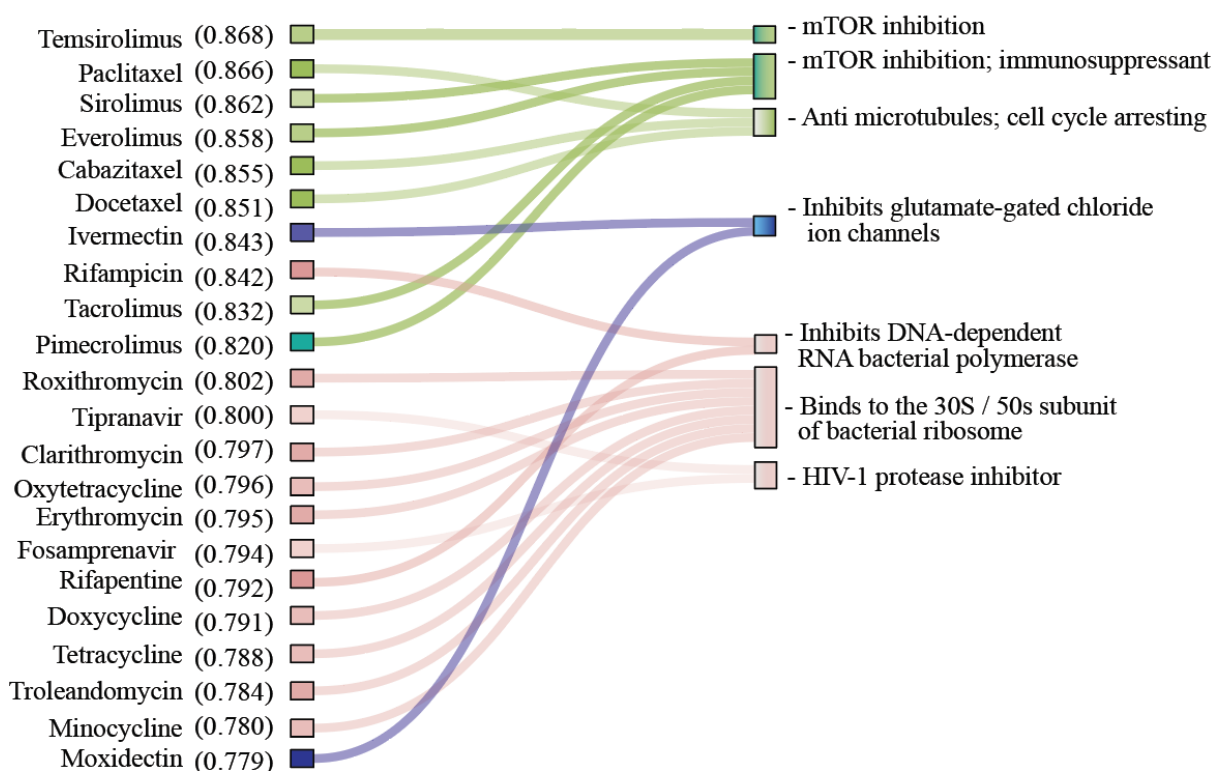


Figura 8. Correlación entre los fármacos mejor clasificados utilizando el modelo multiobjetivo y su mecanismo de acción. Se enumeran las primeras 22 posiciones (1%) de los 2218 compuestos de DrugBank examinados. Los fármacos y sus valores de deseabilidad obtenidos mediante el algoritmo de predicción se describen en la columna de la izquierda, mientras que su mecanismo de acción está en la columna de la derecha. Los colores representan los grupos de fármacos descritos en la figura anterior.

Es interesante notar que se han encontrado varios fármacos reposicionados en ensayos clínicos para pacientes con cáncer. De los agentes antineoplásicos e inmunomoduladores, solo el cabazitaxel aún no se ha estudiado en ensayos relacionados con sarcomas óseos. Además, los compuestos antibacterianos de amplio espectro como la claritromicina, la eritromicina, la doxiciclina y la tetraciclina son fármacos de primer nivel que están registrados en ensayos clínicos para carcinomas.

Las células cancerígenas se caracterizan por crecimiento no regulado, que conduce a una indiferenciación celular y la alteración de la función de los tejidos. La proliferación celular puede ser causada por una falla en los puntos de control en el ciclo celular o una interrupción en la vía de muerte celular denominada apoptosis. En este sentido, cualquier agente que afecte el metabolismo de las células cancerosas, ya sea reduciendo o inhibiendo la proliferación celular, o promoviendo apoptosis, es un objetivo potencial para el tratamiento del cáncer (188).

Varios agentes utilizados como tratamiento de primera línea para el OS como metotrexato, doxorubicina, etopósido, cisplatino e ifosfamida, inducen una alteración en estas funciones celulares. Esto se logra por medio de la interrupción en la síntesis de nucleótidos, por síntesis de ADN inhibiendo la topoisomerasa II o uniéndose al doble -cadena de ADN en donde se promueve apoptosis. Por otro lado, varios fármacos de segunda línea actúan sobre mTOR, una vía considerada patógena dentro del desarrollo y progresión de la OS (189, 190), y sobre la formación de microtúbulos, inhibiendo la progresión de la fase G1 a S del ciclo celular. En las seis primeras posiciones de nuestro cribado, encontramos fármacos quimioterapéuticos descritos como agentes terapéuticos para varios tipos de cáncer (temsirolimus, paclitaxel, sirolimus, everolimus, cabazitaxel y docetaxel). De hecho, estos compuestos pertenecen a una de las cuatro clases principales de fármacos que se encuentran en nuestro reposicionamiento llamados agentes antineoplásicos e inmunomoduladores. Su mecanismo de acción se asemeja a los descritos anteriormente como fármacos de segunda línea, que inhiben principalmente mTOR e interfieren con la despolimerización de los microtúbulos. Curiosamente, cabazitaxel es el único de estos seis compuestos mejor clasificados que no se informa en los ensayos clínicos de pacientes con OS. Esta molécula es un derivado semisintético de un taxoide natural que aumenta considerablemente la supervivencia global frente a la mitoxantrona después de un tratamiento previo con docetaxel en pacientes con cáncer de próstata metastásico resistente a la castración (191-193). Cabazitaxel induce la detención del ciclo celular al interactuar con la despolimerización de los microtúbulos por lo que se lo define como un agente desestabilizador de microtúbulos. Estos tipos de agentes muestran una alta actividad antineoplásica y se han informado en estudios anteriores sobre el reposicionamiento de fármacos (194). Aunque se usan comúnmente en oncología pediátrica (195), los taxanos estabilizadores de microtúbulos no se usan a menudo para tratar cánceres infantiles debido a su actividad limitada, incluso si se observa seguridad dentro de los ensayos (196). En este sentido, cabazitaxel puede ser un agente terapéutico importante para el tratamiento del OS, especialmente en pacientes que recaen después de una terapia basada en docetaxel.

Al analizar el mecanismo de acción de los compuestos cribados, es interesante observar que el 54,5% del total de compuestos previstos (12 de 22) se clasifican como antiinfecciosos para uso sistémico. Más concretamente, teniendo en cuenta el sistema de clasificación Anatómico Terapéutico Químico (ATC), nuestro protocolo de multi-objetivos ponderó varios macrólidos (roxitromicina, claritromicina, eritromicina y troleandomicina), tetraciclinas (oxitetraciclina, doxiciclina, tetraciclina y minociclina) e inhibidores de proteasa (tippernavir) e inhibidores de la proteasa (tippernavir). antimicobacterianos (rifampicina y rifapentina) como

posibles agentes anti-OS. Por un lado, estudios previos sobre la terapia del cáncer han señalado la importancia de los compuestos macrólidos y tetraciclina en el tratamiento del cáncer (197, 198). Algunos autores han sugerido que estos grupos de compuestos inhiben la acción de las metaloproteinasas de la matriz (MMP) para reducir el grado de invasión tumoral y metástasis (199). Otros han observado que estos fármacos actúan sobre la biogénesis mitocondrial (200, 201), interrumpiendo este proceso y aumentando así la eficacia de la quimioterapia o la radioterapia en las células tumorales. Por otro lado, se ha informado de la acción terapéutica de los inhibidores de la proteasa del VIH para el tratamiento del cáncer. Aunque no se espera que estas moléculas reaccionen de forma cruzada con péptidos humanos, los datos preclínicos sugieren que su actividad antitumoral puede estar relacionada en parte con la inhibición de endopeptidasas, como metaloproteasas y proteasomas (202). De nuestros fármacos reposicionados, la claritromicina, la eritromicina y la doxiciclina se encuentran actualmente en estudio como posibles agentes terapéuticos para la leucemia, el cáncer colorrectal, de próstata y de pulmón, entre otros (203-206), y están involucrados en ensayos clínicos de pacientes con cáncer. Tomando en cuenta nuestros hallazgos, estos agentes podrían demostrar actividad antitumoral en tumores óseos.

4. CONCLUSIONES

Las conclusiones se exponen en función de los objetivos propuestos: 1) desarrollo de estrategia consensus y priorización de genes patogénicos en osteosarcoma; 2) construcción de algoritmo de aprendizaje de máquinas teoría de la perturbación (PTML) para predicción de fármacos anti-sarcomas ; 3) construcción de modelo de multi-objetivos para predicción de fármacos anti-osteosarcoma; 4) reposicionamiento de fármacos con posible actividad terapéutica en osteosarcoma; 5) publicación de resultados en revistas indexadas.

- 1) La estrategia de consenso demostró ser eficaz al momento de especificar una amplia lista de genes obtenidos de varias herramientas de priorización bioinformática. Además, la combinación de esta estrategia con metodologías de ontología génica, análisis de redes y enriquecimiento, nos permitieron mostrar no solo explicar interacciones reales entre genes específicos, sino también para definir interacciones internas que explican los eventos celulares asociados con la patogénesis del OS. Los resultados obtenidos en la priorización nos permitieron explicar procesos metabólicos y proponer nuevos genes que pueden ser tomados en cuenta como dianas terapéuticas para la generación de fármacos dirigidos.
- 2) El modelo PTML-LDA construido es el primer algoritmo de aprendizaje de máquinas construido para predicción de fármacos con actividad biológica anti-sarcoma. La reducción en la cantidad de variables de entrada resulta en un modelo con alta simplicidad y e interpretabilidad, por lo que puede ser implementado en investigación médica oncológica.
- 3) El rendimiento de nuestro modelo de multi-objetivos mejora considerablemente la tasa de reconocimiento en un escenario de cribado virtual, desarrollado con fármacos de primera y segunda línea utilizados como tratamiento para el osteosarcoma. Específicamente, nuestro algoritmo de predicción puede recuperar casi 27 veces más el número de compuestos de múltiples objetivos en el primer 1% de la lista clasificada que lo que se espera de una distribución uniforme de los activos en la base de datos de cribado virtual. Considerando estos resultados, este modelo es de gran importancia dentro de investigación farmacológica en cáncer.
- 4) La falta de opciones terapéuticas para el tratamiento del osteosarcoma es una de las principales razones por las que desarrollamos un reposicionamiento de fármacos. Así, mediante esta estrategia, proponemos varios agentes antineoplásicos con posible acción terapéutica. Entre ellos, varios antibióticos de amplio espectro como la claritromicina, eritromicina y doxiciclina son importantes para ser validados en futuras investigaciones.
- 5) Todos los resultados obtenidos en este trabajo han sido publicados en revistas indexadas en SCOPUS.

En conclusión, la aplicación de estrategias teóricas para priorización, ontología génica y análisis de redes y enriquecimiento, resultan importantes al momento de explorar procesos biológicos que permiten explicar la patogénesis de una enfermedad tan heterogénea como es el osteosarcoma. Además, el desarrollo de modelos de predicción basados en aprendizaje de máquinas permite proponer nuevos fármacos en una enfermedad que mantiene su tratamiento farmacológico poco explorado durante décadas. Los resultados obtenidos en este trabajo son prometedores para futuros ensayos experimentales, preclínicos y clínicos por lo que la información generada es útil para futuros proyectos de investigación oncológica.

5. REFERENCIAS

1. Sadykova LR, Ntekim AI, Muyangwa-Semenova M, Rutland CS, Jeyapalan JN, Blatt N, et al. Epidemiology and Risk Factors of Osteosarcoma. *Cancer Invest.* 2020;38(5):259-69.
2. Organization WH. Global Health Observatory. Geneva: World Health Organization. 2020.
3. Ottaviani G, Jaffe N. The epidemiology of osteosarcoma. *Cancer Treat Res.* 2009;152:3-13.
4. Schaefer IM, Fletcher CDM. Recent advances in the diagnosis of soft tissue tumours. *Pathology.* 2018;50(1):37-48.
5. Misaghi A, Goldin A, Awad M, Kulidjian AA. Osteosarcoma: a comprehensive review. *SICOT J.* 2018;4:12.
6. PosthumaDeBoer J, Witlox MA, Kaspers GJ, van Royen BJ. Molecular alterations as target for therapy in metastatic osteosarcoma: a review of literature. *Clin Exp Metastasis.* 2011;28(5):493-503.
7. Moore DD, Luu HH. Osteosarcoma. *Orthopaedic oncology:* Springer; 2014. p. 65-92.
8. Durfee RA, Mohammed M, Luu HH. Review of Osteosarcoma and Current Management. *Rheumatol Ther.* 2016;3(2):221-43.
9. Meazza C, Scanagatta P. Metastatic osteosarcoma: a challenging multidisciplinary treatment. *Expert Rev Anticancer Ther.* 2016;16(5):543-56.
10. Network NCC. NCCN Clinical Practice Guidelines in Oncology (NCC Guidelines) - Bone Cancer 2020 [cited 2020. Available from: <https://www.nccn.org/>.
11. Jafari F, Javdansirat S, Sanaie S, Naseri A, Shamekh A, Rostamzadeh D, et al. Osteosarcoma: A comprehensive review of management and treatment strategies. *Ann Diagn Pathol.* 2020;49:151654.
12. Biermann JS, Chow W, Reed DR, Lucas D, Adkins DR, Agulnik M, et al. NCCN guidelines insights: bone cancer, version 2.2017. 2017;15(2):155-67.
13. Faisham WI, Mat Saad AZ, Alsaigh LN, Nor Azman MZ, Kamarul Imran M, Biswal BM, et al. Prognostic factors and survival rate of osteosarcoma: A single-institution study. *Asia Pac J Clin Oncol.* 2017;13(2):e104-e10.
14. Kager L, Tamamyran G, Bielack S. Novel insights and therapeutic interventions for pediatric osteosarcoma. *Future Oncol.* 2017;13(4):357-68.
15. Guerrero S, Lopez-Cortes A, Indacochea A, Garcia-Cardenas JM, Zambrano AK, Cabrera-Andrade A, et al. Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. *Sci Rep.* 2018;8(1):13978.
16. Martin JW, Squire JA, Zielenska M. The genetics of osteosarcoma. *Sarcoma.* 2012;2012:627254.
17. Piraino SW, Furney SJ. Identification of coding and non-coding mutational hotspots in cancer genomes. *BMC Genomics.* 2017;18(1):17.
18. Rickel K, Fang F, Tao J. Molecular genetics of osteosarcoma. *Bone.* 2017;102:69-79.
19. Serra M, Hattinger CM. The pharmacogenomics of osteosarcoma. *Pharmacogenomics J.* 2017;17(1):11-20.
20. Gianferante DM, Mirabello L, Savage SA. Germline and somatic genetics of osteosarcoma - connecting aetiology, biology and therapy. *Nat Rev Endocrinol.* 2017;13(8):480-91.
21. Long F, Ornitz DM. Development of the endochondral skeleton. *Cold Spring Harb Perspect Biol.* 2013;5(1):a008334.
22. Mu X, Isaac C, Greco N, Huard J, Weiss K. Notch Signaling is Associated with ALDH Activity and an Aggressive Metastatic Phenotype in Murine Osteosarcoma Cells. *Front Oncol.* 2013;3:143.

23. Kolb EA, Gorlick R, Keir ST, Maris JM, Lock R, Carol H, et al. Initial testing (stage 1) by the pediatric preclinical testing program of RO4929097, a gamma-secretase inhibitor targeting notch signaling. *Pediatr Blood Cancer*. 2012;58(5):815-8.
24. Toosi S, Behravan J. Osteogenesis and bone remodeling: A focus on growth factors and bioactive peptides. *Biofactors*. 2020;46(3):326-40.
25. Rettew AN, Getty PJ, Greenfield EM. Receptor tyrosine kinases in osteosarcoma: not just the usual suspects. *Adv Exp Med Biol*. 2014;804:47-66.
26. Cao Y, Roth M, Piperdi S, Montoya K, Sowers R, Rao P, et al. Insulin-like growth factor 1 receptor and response to anti-IGF1R antibody therapy in osteosarcoma. *PLoS One*. 2014;9(8):e106249.
27. Corre I, Verrecchia F, Crenn V, Redini F, Trichet V. The Osteosarcoma Microenvironment: A Complex But Targetable Ecosystem. *Cells*. 2020;9(4).
28. Tabak SA, Khalifa SE, Fathy Y. HER-2 Immunohistochemical Expression in Bone Sarcomas: A New Hope for Osteosarcoma Patients. *Open Access Maced J Med Sci*. 2018;6(9):1555-60.
29. Cisek P. Making decisions through a distributed consensus. *Curr Opin Neurobiol*. 2012;22(6):927-36.
30. Cruz-Monteagudo M, Borges F, Paz YMC, Cordeiro MN, Rebelo I, Perez-Castillo Y, et al. Efficient and biologically relevant consensus strategy for Parkinson's disease gene prioritization. *BMC Med Genomics*. 2016;9:12.
31. Tejera E, Cruz-Monteagudo M, Burgos G, Sanchez ME, Sanchez-Rodriguez A, Perez-Castillo Y, et al. Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis. *BMC Med Genomics*. 2017;10(1):50.
32. Lopez-Cortes A, Paz YMC, Cabrera-Andrade A, Barigye SJ, Munteanu CR, Gonzalez-Diaz H, et al. Gene prioritization, communality analysis, networking and metabolic integrated pathway to better understand breast cancer pathogenesis. *Sci Rep*. 2018;8(1):16679.
33. Liekens AM, De Knijf J, Daelemans W, Goethals B, De Rijk P, Del-Favero J. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol*. 2011;12(6):R57.
34. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol Syst Biol*. 2008;4:189.
35. Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)*. 2015;2015:bav028.
36. Fontaine JF, Priller F, Barbosa-Silva A, Andrade-Navarro MA. Genie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res*. 2011;39(Web Server issue):W455-61.
37. Jourquin J, Duncan D, Shi Z, Zhang B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics*. 2012;13 Suppl 8:S20.
38. Guney E, Garcia-Garcia J, Oliva B. GUILDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. *Bioinformatics*. 2014;30(12):1789-90.
39. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods*. 2015;12(9):841-3.
40. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res*. 2008;36(Web Server issue):W399-405.
41. Yue P, Melamud E, Moulton J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*. 2006;7:166.

42. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet.* 2012;13(8):523-36.
43. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4(1):44-57.
44. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37(1):1-13.
45. Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One.* 2011;6(7):e21800.
46. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43(Database issue):D447-52.
47. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-504.
48. Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature.* 2005;435(7043):814-8.
49. Zhou Z, Xiao Z, Deng W. Improved community structure discovery algorithm based on combined clique percolation method and K-means algorithm. *Peer-to-Peer Networking and Applications.* 2020;13(6):2224-33.
50. McDonald ER, 3rd, de Weck A, Schlabach MR, Billy E, Mavrakis KJ, Hoffman GR, et al. Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell.* 2017;170(3):577-92 e10.
51. Li Z, Ivanov AA, Su R, Gonzalez-Pecchi V, Qi Q, Liu S, et al. The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nat Commun.* 2017;8:14356.
52. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. *Cell.* 2018;175(2):598-9.
53. Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation-2019 update. *Nucleic Acids Res.* 2019;47(D1):D100-D5.
54. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov.* 2019;18(1):41-58.
55. Double J, Barrass N, Barnard ND, Navaratnam V. Toxicity testing in the development of anticancer drugs. *Lancet Oncol.* 2002;3(7):438-42.
56. Williams RJ, Walker I, Takle AK. Collaborative approaches to anticancer drug discovery and development: a Cancer Research UK perspective. *Drug Discov Today.* 2012;17(5-6):185-7.
57. Heinemann F, Huber T, Meisel C, Bundschus M, Leser U. Reflection of successful anticancer drug development processes in the literature. *Drug Discov Today.* 2016;21(11):1740-4.
58. Sun J, Wei Q, Zhou Y, Wang J, Liu Q, Xu H. A systematic analysis of FDA-approved anticancer drugs. *BMC Syst Biol.* 2017;11(Suppl 5):87.
59. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, et al. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* 2019;47(D1):D1056-D65.
60. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* 2019;47(D1):D930-D40.

61. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. *Nucleic Acids Res.* 2017;45(D1):D945-D54.
62. Casanola-Martin GM, Le-Thi-Thu H, Perez-Gimenez F, Marrero-Ponce Y, Merino-Sanjuan M, Abad C, et al. Multi-output Model with Box-Jenkins Operators of Quadratic Indices for Prediction of Malaria and Cancer Inhibitors Targeting Ubiquitin-Proteasome Pathway (UPP) Proteins. *Curr Protein Pept Sci.* 2016;17(3):220-7.
63. Romero-Duran FJ, Alonso N, Yanez M, Caamano O, Garcia-Mera X, Gonzalez-Diaz H. Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology.* 2016;103:270-8.
64. Kleandrova VV, Luan F, Gonzalez-Diaz H, Ruso JM, Speck-Planche A, Cordeiro MN. Computational tool for risk assessment of nanomaterials: novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environ Sci Technol.* 2014;48(24):14686-94.
65. Luan F, Kleandrova VV, Gonzalez-Diaz H, Ruso JM, Melo A, Speck-Planche A, et al. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale.* 2014;6(18):10623-30.
66. Alonso N, Caamano O, Romero-Duran FJ, Luan F, MN DSC, Yanez M, et al. Model for high-throughput screening of multitarget drugs in chemical neurosciences: synthesis, assay, and theoretic study of rasagiline carbamates. *ACS Chem Neurosci.* 2013;4(10):1393-403.
67. Gonzalez-Diaz H, Arrasate S, Gomez-SanJuan A, Sotomayor N, Lete E, Besada-Porto L, et al. General theory for multiple input-output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry. *Curr Top Med Chem.* 2013;13(14):1713-41.
68. Kleandrova VV, Ruso JM, Speck-Planche A, Dias Soeiro Cordeiro MN. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS Comb Sci.* 2016;18(8):490-8.
69. Speck-Planche A, Cordeiro MN. Simultaneous virtual prediction of anti-Escherichia coli activities and ADMET profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS Comb Sci.* 2014;16(2):78-84.
70. Speck-Planche A, Dias Soeiro Cordeiro MN. Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based in Silico Approach for the Design of Virtual Anti-Hepatitis C Leads. *ACS Comb Sci.* 2017;19(8):501-12.
71. Speck-Planche A, Cordeiro MN. Computer-aided discovery in antimicrobial research: In silico model for virtual screening of potent and safe anti-pseudomonas agents. *Comb Chem High Throughput Screen.* 2015;18(3):305-14.
72. Santana R, Zuluaga R, Ganan P, Arrasate S, Onieva E, Montemore MM, et al. PTML Model for Selection of Nanoparticles, Anticancer Drugs, and Vitamins in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Cotherapy. *Mol Pharm.* 2020;17(7):2612-27.
73. Santana R, Zuluaga R, Ganan P, Arrasate S, Onieva E, Gonzalez-Diaz H. Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models. *Nanoscale.* 2020;12(25):13471-83.
74. Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in cheminformatics and drug discovery. *Drug Discov Today.* 2018;23(8):1538-46.
75. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev.* 2019;11(1):31-9.

76. Wang J, Yun D, Yao J, Fu W, Huang F, Chen L, et al. Design, synthesis and QSAR study of novel isatin analogues inspired Michael acceptor as potential anticancer compounds. *Eur J Med Chem.* 2018;144:493-503.
77. Pogorzelska A, Slawinski J, Zolnowska B, Szafranski K, Kawiak A, Chojnacki J, et al. Novel 2-(2-alkylthiobenzenesulfonyl)-3-(phenylprop-2-ynylideneamino)guanidine derivatives as potent anticancer agents - Synthesis, molecular structure, QSAR studies and metabolic stability. *Eur J Med Chem.* 2017;138:357-70.
78. Slawinski J, Szafranski K, Pogorzelska A, Zolnowska B, Kawiak A, Macur K, et al. Novel 2-benzylthio-5-(1,3,4-oxadiazol-2-yl)benzenesulfonamides with anticancer activity: Synthesis, QSAR study, and metabolic stability. *Eur J Med Chem.* 2017;132:236-48.
79. Singh H, Kumar R, Singh S, Chaudhary K, Gautam A, Raghava GP. Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. *BMC Cancer.* 2016;16:77.
80. Toropov AA, Toropova AP, Benfenati E, Gini G, Leszczynska D, Leszczynski J. SMILES-based QSAR approaches for carcinogenicity and anticancer activity: comparison of correlation weights for identical SMILES attributes. *Anticancer Agents Med Chem.* 2011;11(10):974-82.
81. Gonzalez-Diaz H, Bonet I, Teran C, De Clercq E, Bello R, Garcia MM, et al. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur J Med Chem.* 2007;42(5):580-5.
82. Gonzalez-Diaz H, Vina D, Santana L, de Clercq E, Uriarte E. Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg Med Chem.* 2006;14(4):1095-107.
83. Gonzalez-Diaz H, Torres-Gomez LA, Guevara Y, Almeida MS, Molina R, Castanedo N, et al. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. *J Mol Model.* 2005;11(2):116-23.
84. Tharwat A, Gaber T, Ibrahim A, Hassanien AEJAc. Linear discriminant analysis: A detailed tutorial. 2017;30(2):169-90.
85. Ortega-Tenezaca B, Quevedo-Tumaili V, Bediaga H, Collados J, Arrasate S, Madariaga G, et al. PTML Multi-Label Algorithms: Models, Software, and Applications. *Curr Top Med Chem.* 2020;20(25):2326-37.
86. Speck-Planche A, Cordeiro M. Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol Divers.* 2017;21(3):511-23.
87. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. *Anticancer Agents Med Chem.* 2013;13(5):791-800.
88. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anticancer Agents Med Chem.* 2012;12(6):678-85.
89. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorg Med Chem.* 2012;20(15):4848-55.
90. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *Eur J Pharm Sci.* 2012;47(1):273-9.
91. Cordeiro MN, Speck-Planche A. Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. *Curr Top Med Chem.* 2012;12(24):2703-4.
92. Wei DQ, Selvaraj G, Kaushik AC. Computational Perspective on the Current State of the Methods and New Challenges in Cancer Drug Discovery. *Curr Pharm Des.* 2018;24(32):3725-6.

93. Bediaga H, Arrasate S, Gonzalez-Diaz H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb Sci.* 2018;20(11):621-32.
94. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Fragment-based QSAR model toward the selection of versatile anti-sarcoma leads. *Eur J Med Chem.* 2011;46(12):5910-6.
95. Jarada TN, Rokne JG, Alhaji R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J Cheminform.* 2020;12(1):46.
96. Park K. A review of computational drug repurposing. *Transl Clin Pharmacol.* 2019;27(2):59-63.
97. Sadeghi SS, Keyvanpour MR. An Analytical Review of Computational Drug Repurposing. *IEEE/ACM Trans Comput Biol Bioinform.* 2019;PP.
98. Cruz-Monteagudo M, Schurer S, Tejera E, Perez-Castillo Y, Medina-Franco JL, Sanchez-Rodriguez A, et al. Systemic QSAR and phenotypic virtual screening: chasing butterflies in drug discovery. *Drug Discov Today.* 2017;22(7):994-1007.
99. Nagamalla L, Kumar JVS. In silico screening of FDA approved drugs on AXL kinase and validation for breast cancer cell line. *J Biomol Struct Dyn.* 2020:1-15.
100. Issa NT, Stathias V, Schurer S, Dakshanamurthy S. Machine and deep learning approaches for cancer drug repurposing. *Semin Cancer Biol.* 2020.
101. Koudijs KKM, Terwisscha van Scheltinga AGT, Bohringer S, Schimmel KJM, Guchelaar HJ. Personalised drug repositioning for Clear Cell Renal Cell Carcinoma using gene expression. *Sci Rep.* 2018;8(1):5250.
102. Wei GG, Gao L, Tang ZY, Lin P, Liang LB, Zeng JJ, et al. Drug repositioning in head and neck squamous cell carcinoma: An integrated pathway analysis based on connectivity map and differential gene expression. *Pathol Res Pract.* 2019;215(6):152378.
103. Lopez-Cortes A, Paz YMC, Guerrero S, Cabrera-Andrade A, Barigye SJ, Munteanu CR, et al. OncoOmics approaches to reveal essential genes in breast cancer: a panoramic view from pathogenesis to precision medicine. *Sci Rep.* 2020;10(1):5285.
104. Li X, Yan ML, Yu Q. Identification of candidate drugs for the treatment of metastatic osteosarcoma through a subpathway analysis method. *Oncol Lett.* 2017;13(6):4378-84.
105. ChemAxon. JChem for Office 2018 [Available from: <https://chemaxon.com/>].
106. ChemAxon. Chemaxon Standardizer 2018 [Available from: <https://www.chemaxon.com>].
107. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, et al. ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors. 2008;4(3):191.
108. Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov Today.* 2016;21(8):1291-302.
109. Ruggiu F, Marcou G, Varnek A, Horvath D. ISIDA Property-Labelled Fragment Descriptors. *Mol Inform.* 2010;29(12):855-68.
110. Varnek A, Fourches D, Hoonakker F, Solov'ev VP. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J Comput Aided Mol Des.* 2005;19(9-10):693-703.
111. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell.* 2005;27(8):1226-38.
112. Potter T, Matter H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J Med Chem.* 1998;41(4):478-88.

113. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
114. Perez-Castillo Y, Sanchez-Rodriguez A, Tejera E, Cruz-Monteagudo M, Borges F, Cordeiro M, et al. A desirability-based multi objective approach for the virtual screening discovery of broad-spectrum anti-gastric cancer agents. *PLoS One*. 2018;13(2):e0192176.
115. Hattinger CM, Vella S, Tavanti E, Fanelli M, Picci P, Serra M. Pharmacogenomics of second-line drugs used for treatment of unresponsive or relapsed osteosarcoma patients. *Pharmacogenomics*. 2016;17(18):2097-114.
116. Schwartz GK, Tap WD, Qin LX, Livingston MB, Undevia SD, Chmielowski B, et al. Cixutumumab and temsirolimus for patients with bone and soft-tissue sarcoma: a multicentre, open-label, phase 2 trial. *Lancet Oncol*. 2013;14(4):371-82.
117. Trucco MM, Meyer CF, Thornton KA, Shah P, Chen AR, Wilky BA, et al. A phase II study of temsirolimus and liposomal doxorubicin for patients with recurrent and refractory bone and soft tissue sarcomas. *Clin Sarcoma Res*. 2018;8:21.
118. Demetri GD, Chawla SP, Ray-Coquard I, Le Cesne A, Staddon AP, Milhem MM, et al. Results of an international randomized phase III trial of the mammalian target of rapamycin inhibitor ridaforolimus versus placebo to control metastatic sarcomas in patients after benefit from prior chemotherapy. *J Clin Oncol*. 2013;31(19):2485-92.
119. Qayed M, Cash T, Tighiouart M, MacDonald TJ, Goldsmith KC, Tanos R, et al. A phase I study of sirolimus in combination with metronomic therapy (CHOAnome) in children with recurrent or refractory solid and brain tumors. *Pediatr Blood Cancer*. 2020;67(4):e28134.
120. van der Graaf WT, Blay JY, Chawla SP, Kim DW, Bui-Nguyen B, Casali PG, et al. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet*. 2012;379(9829):1879-86.
121. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*. 2012;55(14):6582-94.
122. Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model*. 2007;47(2):488-508.
123. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes? *J Comput Aided Mol Des*. 2008;22(3-4):213-28.
124. Srivastava S, Wang S, Tong YA, Hao ZM, Chang EH. Dominant negative effect of a germ-line mutant p53: a step fostering tumorigenesis. *Cancer Res*. 1993;53(19):4452-5.
125. Friend SH, Bernards R, Rogelj S, Weinberg RA, Rapaport JM, Albert DM, et al. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*. 1986;323(6089):643-6.
126. Tsuchiya T, Sekine K, Hinohara S, Namiki T, Nobori T, Kaneko Y. Analysis of the p16INK4, p14ARF, p15, TP53, and MDM2 genes and their prognostic implications in osteosarcoma and Ewing sarcoma. *Cancer Genet Cytogenet*. 2000;120(2):91-8.
127. Gokgoz N, Wunder JS, Mousses S, Eskandarian S, Bell RS, Andrulis IL. Comparison of p53 mutations in patients with localized osteosarcoma and metastatic osteosarcoma. *Cancer*. 2001;92(8):2181-9.
128. Vassilev LT, Vu BT, Graves B, Carvajal D, Podlaski F, Filipovic Z, et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*. 2004;303(5659):844-8.

129. Stratton MR, Moss S, Warren W, Patterson H, Clark J, Fisher C, et al. Mutation of the p53 gene in human soft tissue sarcomas: association with abnormalities of the RB1 gene. *Oncogene*. 1990;5(9):1297-301.
130. Moll UM, Petrenko O. The MDM2-p53 interaction. *Mol Cancer Res*. 2003;1(14):1001-8.
131. Tamura K, Utsunomiya J, Iwama T, Furuyama J, Takagawa T, Takeda N, et al. Mechanism of carcinogenesis in familial tumors. *Int J Clin Oncol*. 2004;9(4):232-45.
132. Sandberg AA, Bridge JA. Updates on the cytogenetics and molecular genetics of bone and soft tissue tumors: osteosarcoma and related tumors. *Cancer Genet Cytogenet*. 2003;145(1):1-30.
133. Kuijjer ML, Hogendoorn PC, Cleton-Jansen AM. Genome-wide analyses on high-grade osteosarcoma: making sense of a genomically most unstable tumor. *Int J Cancer*. 2013;133(11):2512-21.
134. Poos K, Smida J, Maugg D, Eckstein G, Baumhoer D, Nathrath M, et al. Genomic heterogeneity of osteosarcoma - shift from single candidates to functional modules. *PLoS One*. 2015;10(4):e0123082.
135. Sun L, Li J, Yan B. Gene expression profiling analysis of osteosarcoma cell lines. *Mol Med Rep*. 2015;12(3):4266-72.
136. Shi Z, Zhou H, Pan B, Lu L, Wei Z, Shi L, et al. Exploring the key genes and pathways of osteosarcoma with pulmonary metastasis using a gene expression microarray. *Mol Med Rep*. 2017;16(5):7423-31.
137. Farhan M, Wang H, Gaur U, Little PJ, Xu J, Zheng W. FOXO Signaling Pathways as Therapeutic Targets in Cancer. *Int J Biol Sci*. 2017;13(7):815-27.
138. Shaw RJ, Cantley LC. Ras, PI(3)K and mTOR signalling controls tumour cell growth. *Nature*. 2006;441(7092):424-30.
139. Siqueira MF, Flowers S, Bhattacharya R, Faibish D, Behl Y, Kotton DN, et al. FOXO1 modulates osteoblast differentiation. *Bone*. 2011;48(5):1043-51.
140. Kim HN, Iyer S, Ring R, Almeida M. The Role of FoxOs in Bone Health and Disease. *Curr Top Dev Biol*. 2018;127:149-63.
141. Tan P, Guan H, Xie L, Mi B, Fang Z, Li J, et al. FOXO1 inhibits osteoclastogenesis partially by antagonizing MYC. *Sci Rep*. 2015;5:16835.
142. Coomans de Brachene A, Demoulin JB. FOXO transcription factors in cancer development and therapy. *Cell Mol Life Sci*. 2016;73(6):1159-72.
143. Moriishi T, Kawai Y, Komori H, Rokutanda S, Eguchi Y, Tsujimoto Y, et al. Bcl2 deficiency activates FoxO through Akt inactivation and accelerates osteoblast differentiation. *PLoS One*. 2014;9(1):e86629.
144. Guan H, Tan P, Xie L, Mi B, Fang Z, Li J, et al. FOXO1 inhibits osteosarcoma oncogenesis via Wnt/beta-catenin pathway suppression. *Oncogenesis*. 2015;4:e166.
145. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep*. 2013;3:2650.
146. Xu Y, Li N, Xiang R, Sun P. Emerging roles of the p38 MAPK and PI3K/AKT/mTOR pathways in oncogene-induced senescence. *Trends Biochem Sci*. 2014;39(6):268-76.
147. De Luca A, Maiello MR, D'Alessio A, Pergameno M, Normanno N. The RAS/RAF/MEK/ERK and the PI3K/AKT signalling pathways: role in cancer pathogenesis and implications for therapeutic approaches. *Expert Opin Ther Targets*. 2012;16 Suppl 2:S17-27.
148. Han G, Wang Y, Bi W. C-Myc overexpression promotes osteosarcoma cell invasion via activation of MEK-ERK pathway. *Oncol Res*. 2012;20(4):149-56.
149. Samatar AA, Poulikakos PI. Targeting RAS-ERK signalling in cancer: promises and challenges. *Nat Rev Drug Discov*. 2014;13(12):928-42.

150. Daw NC, Chou AJ, Jaffe N, Rao BN, Billups CA, Rodriguez-Galindo C, et al. Recurrent osteosarcoma with a single pulmonary metastasis: a multi-institutional review. *Br J Cancer*. 2015;112(2):278-82.
151. Diepenbruck M, Christofori G. Epithelial-mesenchymal transition (EMT) and metastasis: yes, no, maybe? *Curr Opin Cell Biol*. 2016;43:7-13.
152. Jiang WG, Sanders AJ, Katoh M, Ungefroren H, Gieseler F, Prince M, et al. Tissue invasion and metastasis: Molecular, biological and clinical perspectives. *Semin Cancer Biol*. 2015;35 Suppl:S244-S75.
153. Gross AC, Cam H, Phelps DA, Saraf AJ, Bid HK, Cam M, et al. IL-6 and CXCL8 mediate osteosarcoma-lung interactions critical to metastasis. *JCI Insight*. 2018;3(16).
154. Weekes D, Kashima TG, Zandueta C, Perurena N, Thomas DP, Sunters A, et al. Regulation of osteosarcoma cell lung metastasis by the c-Fos/AP-1 target FGFR1. *Oncogene*. 2016;35(22):2948.
155. Liang C, Li F, Wang L, Zhang ZK, Wang C, He B, et al. Tumor cell-targeted delivery of CRISPR/Cas9 by aptamer-functionalized lipopolymer for therapeutic genome editing of VEGFA in osteosarcoma. *Biomaterials*. 2017;147:68-85.
156. Liu Y, Wang Y, Teng Z, Chen J, Li Y, Chen Z, et al. Matrix metalloproteinase 9 expression and survival of patients with osteosarcoma: a meta-analysis. *Eur J Cancer Care (Engl)*. 2017;26(1).
157. Guleria A, Chandna S. ATM kinase: Much more than a DNA damage responsive protein. *DNA Repair (Amst)*. 2016;39:1-20.
158. Awasthi P, Foiani M, Kumar A. ATM and ATR signaling at a glance. *J Cell Sci*. 2015;128(23):4255-62.
159. Ray A, Blevins C, Wani G, Wani AA. ATR- and ATM-Mediated DNA Damage Response Is Dependent on Excision Repair Assembly during G1 but Not in S Phase of Cell Cycle. *PLoS One*. 2016;11(7):e0159344.
160. Blackford AN, Jackson SP. ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response. *Mol Cell*. 2017;66(6):801-17.
161. Kovac M, Blattmann C, Ribí S, Smida J, Mueller NS, Engert F, et al. Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. *Nat Commun*. 2015;6:8940.
162. Nabetani A, Ishikawa F. Alternative lengthening of telomeres pathway: recombination-mediated telomere maintenance mechanism in human cells. *J Biochem*. 2011;149(1):5-14.
163. di Masi A, Cilli D, Berardinelli F, Talarico A, Pallavicini I, Pennisi R, et al. PML nuclear body disruption impairs DNA double-strand break sensing and repair in APL. *Cell Death Dis*. 2016;7:e2308.
164. Lallemand-Breitenbach V, de Thé H. PML nuclear bodies: from architecture to function. *Curr Opin Cell Biol*. 2018;52:154-61.
165. Voisset E, Moravcsik E, Stratford EW, Jaye A, Palgrave CJ, Hills RK, et al. Pml nuclear body disruption cooperates in APL pathogenesis and impairs DNA damage repair pathways in mice. *Blood*. 2018;131(6):636-48.
166. Dilley RL, Verma P, Cho NW, Winters HD, Wondisford AR, Greenberg RA. Break-induced telomere synthesis underlies alternative telomere maintenance. *Nature*. 2016;539(7627):54-8.
167. Kim JY, Brosnan-Cashman JA, An S, Kim SJ, Song KB, Kim MS, et al. Alternative Lengthening of Telomeres in Primary Pancreatic Neuroendocrine Tumors Is Associated with Aggressive Clinical Behavior and Poor Survival. *Clin Cancer Res*. 2017;23(6):1598-606.
168. Singhi AD, Liu TC, Roncaioli JL, Cao D, Zeh HJ, Zureikat AH, et al. Alternative Lengthening of Telomeres and Loss of DAXX/ATRAX Expression Predicts Metastatic

- Disease and Poor Survival in Patients with Pancreatic Neuroendocrine Tumors. *Clin Cancer Res.* 2017;23(2):600-9.
169. Fogli A, Demattei MV, Corset L, Vaurs-Barriere C, Chautard E, Biau J, et al. Detection of the alternative lengthening of telomeres pathway in malignant gliomas for improved molecular diagnosis. *J Neurooncol.* 2017;135(2):381-90.
 170. Chong JL, Wenzel PL, Saenz-Robles MT, Nair V, Ferrey A, Hagan JP, et al. E2f1-3 switch from activators in progenitor cells to repressors in differentiating cells. *Nature.* 2009;462(7275):930-4.
 171. Pires BRB, Silva R, Ferreira GM, Abdelhay E. NF-kappaB: Two Sides of the Same Coin. *Genes (Basel).* 2018;9(1).
 172. Seoane J, Gomis RR. TGF-beta Family Signaling in Tumor Suppression and Cancer Progression. *Cold Spring Harb Perspect Biol.* 2017;9(12).
 173. Pohl T, Gugasyan R, Grumont RJ, Strasser A, Metcalf D, Tarlinton D, et al. The combined absence of NF-kappa B1 and c-Rel reveals that overlapping roles for these transcription factors in the B cell lineage are restricted to the activation and function of mature cells. *Proc Natl Acad Sci U S A.* 2002;99(7):4514-9.
 174. Eurtivong C, Reynisson J. The Development of a Weighted Index to Optimise Compound Libraries for High Throughput Screening. *Mol Inform.* 2019;38(3):e1800068.
 175. Oyewole RO, Oyebamiji AK, Semire B. Theoretical calculations of molecular descriptors for anticancer activities of 1, 2, 3-triazole-pyrimidine derivatives against gastric cancer cell line (MGC-803): DFT, QSAR and docking approaches. *Heliyon.* 2020;6(5):e03926.
 176. Ciura K, Belka M, Kawczak P, Baczek T, Markuszewski MJ, Nowakowska J. Combined computational-experimental approach to predict blood-brain barrier (BBB) permeation based on "green" salting-out thin layer chromatography supported by simple molecular descriptors. *J Pharm Biomed Anal.* 2017;143:214-21.
 177. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg Med Chem.* 2011;19(21):6239-44.
 178. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov.* 2019;18(6):463-77.
 179. Lin X, Li X, Lin X. A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules.* 2020;25(6).
 180. Chen J, Zhang L. A survey and systematic assessment of computational methods for drug response prediction. *Brief Bioinform.* 2021;22(1):232-46.
 181. Yuan H, Paskov I, Paskov H, Gonzalez AJ, Leslie CS. Multitask learning improves prediction of cancer drug sensitivity. *Sci Rep.* 2016;6:31619.
 182. Nikolova O, Moser R, Kemp C, Gonen M, Margolin AA. Modeling gene-wise dependencies improves the identification of drug response biomarkers in cancer studies. *Bioinformatics.* 2017;33(9):1362-9.
 183. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem.* 2011;32(7):1466-74.
 184. Xue H, Li J, Xie H, Wang Y. Review of Drug Repositioning Approaches and Resources. *Int J Biol Sci.* 2018;14(10):1232-44.
 185. Langedijk J, Mantel-Teeuwisse AK, Slijkerman DS, Schutjens MH. Drug repositioning and repurposing: terminology and definitions in literature. *Drug Discov Today.* 2015;20(8):1027-34.
 186. Bolognesi ML, Cavalli A. Multitarget Drug Discovery and Polypharmacology. *ChemMedChem.* 2016;11(12):1190-2.

187. Ma XH, Shi Z, Tan C, Jiang Y, Go ML, Low BC, et al. In-silico approaches to multi-target drug discovery : computer aided multi-target drug design, multi-target virtual screening. *Pharm Res.* 2010;27(5):739-49.
188. Parvathaneni V, Kulkarni NS, Muth A, Gupta V. Drug repurposing: a promising tool to accelerate the drug discovery process. *Drug Discov Today.* 2019;24(10):2076-85.
189. Ding L, Congwei L, Bei Q, Tao Y, Ruiguo W, Heze Y, et al. mTOR: An attractive therapeutic target for osteosarcoma? *Oncotarget.* 2016;7(31):50805-13.
190. Bishop MW, Janeway KA. Emerging concepts for PI3K/mTOR inhibition as a potential treatment for osteosarcoma. *F1000Res.* 2016;5.
191. de Bono JS, Oudard S, Ozguroglu M, Hansen S, Machiels JP, Kocak I, et al. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: a randomised open-label trial. *Lancet.* 2010;376(9747):1147-54.
192. de Wit R, de Bono J, Sternberg CN, Fizazi K, Tombal B, Wulfing C, et al. Cabazitaxel versus Abiraterone or Enzalutamide in Metastatic Prostate Cancer. *N Engl J Med.* 2019;381(26):2506-18.
193. Oudard S, Fizazi K, Sengelov L, Daugaard G, Saad F, Hansen S, et al. Cabazitaxel Versus Docetaxel As First-Line Therapy for Patients With Metastatic Castration-Resistant Prostate Cancer: A Randomized Phase III Trial-FIRSTANA. *J Clin Oncol.* 2017;35(28):3189-97.
194. Lo YC, Senese S, France B, Gholkar AA, Damoiseaux R, Torres JZ. Computational Cell Cycle Profiling of Cancer Cells for Prioritizing FDA-Approved Drugs with Repurposing Potential. *Sci Rep.* 2017;7(1):11261.
195. Reynolds CP, Kang MH, Maris JM, Kolb EA, Gorlick R, Wu J, et al. Initial testing (stage 1) of the anti-microtubule agents cabazitaxel and docetaxel, by the pediatric preclinical testing program. *Pediatr Blood Cancer.* 2015;62(11):1897-905.
196. Amoroso L, Castel V, Bisogno G, Casanova M, Marquez-Vega C, Chisholm JC, et al. Phase II results from a phase I/II study to assess the safety and efficacy of weekly nab-paclitaxel in paediatric patients with recurrent or refractory solid tumours: A collaboration with the European Innovative Therapies for Children with Cancer Network. *Eur J Cancer.* 2020;135:89-97.
197. Hussain A, Dar MS, Bano N, Hossain MM, Basit R, Bhat AQ, et al. Identification of dinactin, a macrolide antibiotic, as a natural product-based small molecule targeting Wnt/beta-catenin signaling pathway in cancer cells. *Cancer Chemother Pharmacol.* 2019;84(3):551-9.
198. Gupta A, Okesli-Armlovich A, Morgens D, Bassik MC, Khosla C. A genome-wide analysis of targets of macrolide antibiotics in mammalian cells. *J Biol Chem.* 2020;295(7):2057-67.
199. Bahrami F, Morris DL, Pourgholami MH. Tetracyclines: drugs with huge therapeutic potential. *Mini Rev Med Chem.* 2012;12(1):44-52.
200. Fiorillo M, Toth F, Sotgia F, Lisanti MP. Doxycycline, Azithromycin and Vitamin C (DAV): A potent combination therapy for targeting mitochondria and eradicating cancer stem cells (CSCs). *Aging (Albany NY).* 2019;11(8):2202-16.
201. Lamb R, Ozsvari B, Lisanti CL, Tanowitz HB, Howell A, Martinez-Outschoorn UE, et al. Antibiotics that target mitochondria effectively eradicate cancer stem cells, across multiple tumor types: treating cancer like an infectious disease. *Oncotarget.* 2015;6(7):4569-84.
202. Maksimovic-Ivanic D, Fagone P, McCubrey J, Bendtzen K, Mijatovic S, Nicoletti F. HIV-protease inhibitors for the treatment of cancer: Repositioning HIV protease inhibitors while developing more potent NO-hybridized derivatives? *Int J Cancer.* 2017;140(8):1713-26.

203. Petroni G, Stefanini M, Pillozzi S, Crociani O, Becchetti A, Arcangeli A. Data describing the effects of the Macrolide Antibiotic Clarithromycin on preclinical mouse models of Colorectal Cancer. *Data Brief*. 2019;26:104406.
204. Van Nuffel AM, Sukhatme V, Pantziarka P, Meheus L, Sukhatme VP, Bouche G. Repurposing Drugs in Oncology (ReDO)-clarithromycin as an anti-cancer agent. *Ecancermedicalsecience*. 2015;9:513.
205. de Jong J, Hellemans P, De Wilde S, Patricia D, Masterson T, Manikhas G, et al. A drug-drug interaction study of ibrutinib with moderate/strong CYP3A inhibitors in patients with B-cell malignancies. *Leuk Lymphoma*. 2018;59(12):2888-95.
206. Markowska A, Kaysiewicz J, Markowska J, Huczynski A. Doxycycline, salinomycin, monensin and ivermectin repositioned as cancer drugs. *Bioorg Med Chem Lett*. 2019;29(13):1549-54.

6. PUBLICACIONES (ANEXOS)

A continuación, se presentan las 3 publicaciones principales generadas en este trabajo de titulación. Además, se anexan 3 publicaciones con afiliación de la Universidade da Coruña que fueron realizadas en el período de estudios de esta investigación.



Article

Gene Prioritization through Consensus Strategy, Enrichment Methodologies Analysis, and Networking for Osteosarcoma Pathogenesis

Alejandro Cabrera-Andrade ^{1,2,3,*} , Andrés López-Cortés ^{3,4} ,
Gabriela Jaramillo-Koupermann ⁵ , César Paz-y-Miño ⁴ , Yunierkis Pérez-Castillo ^{1,6},
Cristian R. Munteanu ^{3,7,8} , Humbert González-Díaz ^{9,10} , Alejandro Pazos ^{3,7,8} and
Eduardo Tejera ^{1,11,*}

- ¹ Grupo de Bio-Quimioinformática, Universidad de Las Américas, Quito 170125, Ecuador; yunierkis.perez@udla.edu.ec
 - ² Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de Las Américas, Quito 170125, Ecuador
 - ³ RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, 15071 A Coruña, Spain; aalc84@gmail.com (A.L.-C.); c.munteanu@udc.es (C.R.M.); apazos@udc.es (A.P.)
 - ⁴ Centro de Investigación Genética y Genómica, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Quito 170129, Ecuador; cesar.pazymino@ute.edu.ec
 - ⁵ Laboratorio de Biología Molecular, Subproceso de Anatomía Patológica, Hospital de Especialidades Eugenio Espejo, Quito 170403, Ecuador; gaby_jaramillok@yahoo.com
 - ⁶ Escuela de Ciencias Físicas y Matemáticas, Universidad de Las Américas, Quito 170125, Ecuador
 - ⁷ Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), 15006 A Coruña, Spain
 - ⁸ Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Campus de Elviña s/n, 15071 A Coruña, Spain
 - ⁹ Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940 Leioa, Spain
 - ¹⁰ IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain; humberto.gonzalezdiaz@ehu.es
 - ¹¹ Facultad de Ingeniería y Ciencias Agropecuarias, Universidad de Las Américas, Quito 170125, Ecuador
- * Correspondence: raul.cabrera@udla.edu.ec (A.C.-A.); eduardo.tejera@udla.edu.ec (E.T.);
Tel.: +593-2398-1000 (ext. 2717) (A.C.-A.); +593-2398-1000 (ext. 713) (E.T.)

Received: 10 December 2019; Accepted: 30 January 2020; Published: 5 February 2020



Abstract: Osteosarcoma is the most common subtype of primary bone cancer, affecting mostly adolescents. In recent years, several studies have focused on elucidating the molecular mechanisms of this sarcoma; however, its molecular etiology has still not been determined with precision. Therefore, we applied a consensus strategy with the use of several bioinformatics tools to prioritize genes involved in its pathogenesis. Subsequently, we assessed the physical interactions of the previously selected genes and applied a communality analysis to this protein–protein interaction network. The consensus strategy prioritized a total list of 553 genes. Our enrichment analysis validates several studies that describe the signaling pathways PI3K/AKT and MAPK/ERK as pathogenic. The gene ontology described TP53 as a principal signal transducer that chiefly mediates processes associated with cell cycle and DNA damage response. It is interesting to note that the communality analysis clusters several members involved in metastasis events, such as *MMP2* and *MMP9*, and genes associated with DNA repair complexes, like *ATM*, *ATR*, *CHEK1*, and *RAD51*. In this study, we have identified well-known pathogenic genes for osteosarcoma and prioritized genes that need to be further explored.

Keywords: gene prioritization; osteosarcoma; communality analysis; pathogenesis; early recognition

1. Introduction

In recent years, high-throughput technologies have focused on studying the molecular etiology of osteosarcoma (OS) worldwide [1–5]. Valuable information has been gained about whole genetic groups that describe cellular and molecular changes in OS [6,7]. Despite this, there has not been an agreement about specific driver genes for OS etiology, nor have new biomarkers been proposed to be used as therapeutic targets.

OS tumors are characterized by being heterogeneous and showing high rates of somatic structural variations. Their heterogeneity is closely related to their high rates of mutations, which are comparable to breast tumors and leukemia [8–10]. Moreover, cytogenetic abnormalities in OS tumors, including chromosomal segment loss, rearrangement, and amplification with karyotypic complexity in the absence of recurrent clonal translocations, have been described [11,12]. This acute chromosomal instability and widespread deregulation in cell signaling pathways could be the main limitations for the description of specific gene drivers associated with OS. It is therefore necessary to develop an integrative study focused on the biology of systems described for this tumor.

The use of prioritization strategies, through computational tools that use multiple heterogeneous data sources, allows for the improvement in gene detection related to complex traits or specific clinical phenotypes [13,14]. In addition, applying the functional enrichment analysis has proven to be a very efficient approach in gene prioritization because it describes important metabolic interactions that aid in explaining the pathogenesis of a given disease [15,16]. Thus, we used several bioinformatics tools in order to prioritize genes that describe oncological signaling pathways for OS and also applied a consensus strategy with the aim to specify and postulate new pathogenic mechanisms that explain the onset and development of this sarcoma. In Figure 1, we summarize the general workflow to prioritize genes associated with the pathogenesis of OS.

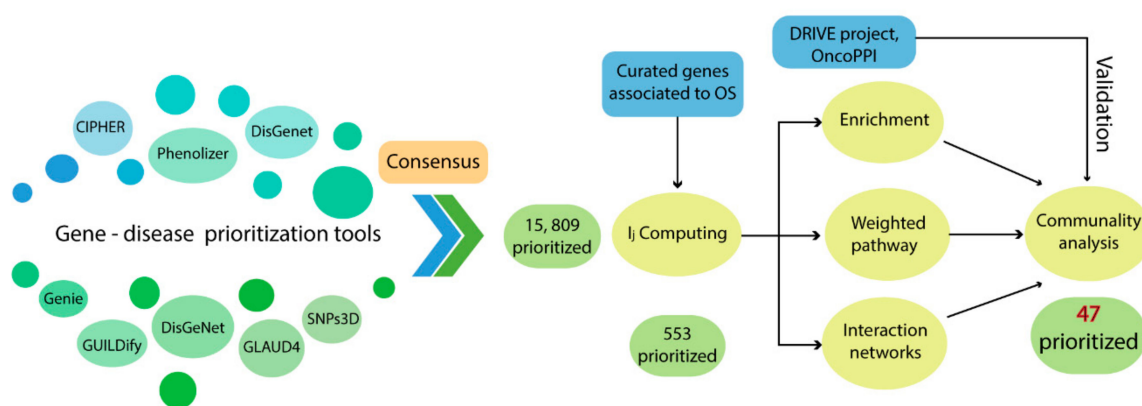


Figure 1. General workflow to gene prioritization.

2. Results

2.1. Consensus Prioritization

We chose nine bioinformatics methods that fulfilled two main criteria: full availability in web service platform and only requiring the disease name (or OMIN code, 259,500 for OS) for gene prioritization. In total, the combination of all methodologies resulted in 15,809 genes.

The validation strategy for gene prioritization was performed from the identification of specific genes involved in the OS pathogenesis. For this, we took into consideration pathogenic OS genes defined by a literature review of two types of studies: meta-analysis, based on publications and case reports for OS patients (named as G1 genes), and gene description in animal models and OS cell lines (named as G2 genes). Thereby, we identified 75 pathogenic OS genes from the available literature, of which 47 were classified as G1 and 41 as G2 (Table S1).

The number of pathogenic genes detected by the nine prioritization tools was lower than our consensus strategy (Table 1). By comparing the number of pathogenic genes detected by all methodologies, our consensus list identifies the highest percentage of those defined as G1 and G2. Specifically, in the top 1% of our consensus method (the first 158 positions), 60% of pathogenic genes (45 of 75) were detected, followed by Genie (35.29%) and Phenolizer (30.14%) methodologies. Furthermore, in the top 20%, the consensus method remains the best at detecting pathogenic genes (88%), followed by Genie, Phenolizer, and SNPs3D with percentages of 80.88%, 72.60%, and 71.88%, respectively.

Table 1. Identification (in %) of pathogenic genes in each osteosarcoma (OS) approach.

Methods	1%			5%			10%			20%		
	G1	G2	G1-2	G1	G2	G1-2	G1	G2	G1-2	G1	G2	G1-2
BioGraph	0	0	0	0	18.2	12.5	40	45.5	37.5	60	54.6	50
CIPHER	7.7	6.7	8.7	7.7	6.7	8.7	23.1	20	17.4	30.8	26.7	26.1
DisGeNET	9.5	16.7	10.8	21.4	30.6	21.5	42.9	58.3	46.2	57.1	77.8	64.6
Genie	37.8	36.1	35.3	62.2	61.1	57.4	75.6	69.4	70.6	86.7	75	80.9
GLAD4U	0	0	3.6	19.1	33.3	25	42.9	50	46.4	57.1	66.7	64.3
GUILDify	10.9	7.5	8.2	13	7.5	9.6	21.7	17.5	19.2	34.8	25	30.1
Phenolizer	33.3	36.6	30.1	57.8	61	53.4	62.2	61	56.2	77.8	75.6	72.6
PolySearch	0	0	0	11.1	14.3	7.1	11.1	28.6	14.3	11.1	28.6	14.3
SNPs3D	10	10.5	6.3	10	42.1	25	40	57.9	50	75	73.7	71.9
Consensus	66	61	60	87.2	80.5	81.3	89.4	82.9	84	93.6	85.4	88

On the other hand, the mean ranking of the pathogenic genes detected in the top 1% of the list is 49.3 (Table 2), which means that 45 G1–G2 genes are located in the top 50 positions. This mean is higher than that calculated for the other prioritization methodologies, given that the number of pathogenic genes detected is greater. However, it is interesting to note that the number of genes and the ranking average are similar, which indicates that the majority of these pathogenic genes are found in the top positions.

Table 2. Rank of pathogenic genes in each OS approach.

Methods	1%			5%			10%			20%		
	G1	G2	G1-2	G1	G2	G1-2	G1	G2	G1-2	G1	G2	G1-2
BioGraph	-	-	-	-	3.5	3.5	7	6	6.3	9.5	7.3	8
CIPHER	2	7	4.5	2	7	4.5	41.3	43	32.8	58	59	57.7
DisGeNET	5.3	4.2	4.7	12.1	10	11.1	23.9	23.6	25.2	31.6	31.4	33.7
Genie	17	14.6	16.5	44	41.6	42.6	88.2	75	91.3	148.5	113.2	151.9
GLAD4U	-	1	1	4	4.2	4	8.6	6.6	8.2	13.3	10.2	13
GUILDify	15.8	8.3	16.7	42.6	8.3	43.3	366.5	536.4	491.2	873.8	973.9	972.1
Phenolizer	44.3	28	36.4	150.4	120.9	148	200.9	120.9	182.5	477.5	429.2	513.2
PolySearch	-	-	-	2	2	2	2	2.5	2.5	2	2.5	2.5
SNPs3D	1.5	1.5	1.5	1.5	6.4	4	17.8	10.9	14.4	27.1	16.2	21.6
Consensus	54.5	41.6	49.3	126.1	108.2	128	152.9	131.2	157.7	241.4	174.7	239.3

This initial prioritization generated an initial amount of 15,809 genes, so a rational cut-off was applied. The maximum variation between *li* and the gene ranking was 0.7609, corresponding with a ranking value of 553. Therefore, this cut-off reduces a list of 15,809 members to a consensus of 553 genes (Table S2), which corresponds to 3.5% of the total. The rate of pathogenic detection of the consensus was 87.2% for G1 (41 out of 47), 80.5% for G2 (33 out of 41), and 81.3% for G1 and G2 (61 out of 75), higher than the other methods in the top 5% onwards.

2.2. Enrichment Analysis of OS Related Genes and the Protein–Protein Interaction Network

A gene ontology (GO) analysis and pathway enrichment analysis was applied in order to describe biological functions from the consensus genes by using the David Bioinformatics Resource [17,18]. The GO analysis of these 553 consensus genes resulted in 263 terms related to biological processes

(Table S3), adjusted to an FDR p -value < 0.01. Using Revigo [19] and only considering terms with a frequency lower than 0.01%, we narrowed our list down to 92 (Table S4). Some of these specific biological processes are listed in Table 3.

Table 3. Some biological processes by enrichment analysis in OS consensus genes.

BP ID	Name	Frequency	Log10 p-Value (FDR)
GO:1901796	regulation of signal transduction by p53 class mediator	0.01%	−22.8416
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0.00%	−20.1656
GO:0048661	positive regulation of smooth muscle cell proliferation	0.01%	−16.5544
GO:0048146	positive regulation of fibroblast proliferation	0.01%	−16.5031
GO:0045740	positive regulation of DNA replication	0.01%	−15.1965
GO:1902895	positive regulation of pri-miRNA transcription from RNA polymerase II promoter	0.00%	−14.983
GO:0043525	positive regulation of neuron apoptotic process	0.01%	−14.9393
GO:0071260	cellular response to mechanical stimulus	0.01%	−13.3507
GO:0032355	response to estradiol	0.01%	−11.7258
GO:0045669	positive regulation of osteoblast differentiation	0.01%	−11.5058
GO:0060395	SMAD protein signal transduction	0.01%	−11.1904
GO:0042771	intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator	0.01%	−10.8356
GO:0097192	extrinsic apoptotic signaling pathway in absence of ligand	0.01%	−10.0846
GO:0035019	somatic stem cell population maintenance	0.01%	−9.6162
GO:0010332	response to gamma radiation	0.01%	−9.4056
GO:0002053	positive regulation of mesenchymal cell proliferation	0.01%	−9.2628
GO:0002076	osteoblast development	0.00%	−9.1046
GO:0048538	thymus development	0.01%	−8.2907
GO:0048010	vascular endothelial growth factor receptor signaling pathway	0.01%	−7.6946
GO:0010718	positive regulation of epithelial to mesenchymal transition	0.01%	−7.6126

Likewise, the enriched metabolic pathways considered in KEGG and Reactome databases are shown in Tables S5 and S6. A partial list of the prioritized metabolic pathways with an FDR p < 0.01 is presented in Table 4.

The enriched biological processes of the 553 genes describe terms associated with positive DNA replication, cellular proliferation, and apoptotic events, in which TP53 is one of the most relevant signal transducers. In addition, more specific sarcoma-related terms are listed, such as smooth muscle cell and fibroblast proliferation, osteoblast differentiation and development, and positive regulation of mesenchymal cell proliferation.

The pathway enrichment analysis showed pathways in cancer and the cell cycle in general. The enrichment from the KEGG database showed widely described signaling pathways in cancer in the top positions, for instance, FOXO, PI3K/AKT, TP53, MAPK, neurotrophin, and cell cycle. Moreover, the Reactome database lists events mainly related to cell cycle regulation such as cyclin D-associated

events in G1, G0 and early G1; Cyclin A: Cdk2-associated events at S phase entry, and Cyclin A/B1 associated events during G2/M transition.

Table 4. Pathways enrichment analysis using KEGG and Reactome databases in OS consensus genes.

Pathway ID	Pathway Name	% Genes	FDR
	KEGG Database		
hsa05200	Pathways in cancer	26.22	1.33×10^{-8}
hsa04110	Cell cycle	11.93	3.96×10^{-45}
hsa04068	FoxO signaling pathway	10.85	4.50×10^{-35}
hsa04151	PI3K-Akt signaling pathway	15.55	1.98×10^{-29}
hsa05206	MicroRNAs in cancer	14.1	3.07×10^{-29}
hsa04115	p53 signaling pathway	7.23	2.38×10^{-29}
hsa05205	Proteoglycans in cancer	11.57	1.63×10^{-27}
hsa04210	Apoptosis	6.69	6.13×10^{-26}
hsa04668	TNF signaling pathway	8.32	2.89×10^{-25}
hsa04510	Focal adhesion	10.85	4.08×10^{-23}
hsa04380	Osteoclast differentiation	8.68	1.21×10^{-22}
hsa04010	MAPK signaling pathway	11.75	8.86×10^{-22}
hsa04722	Neurotrophin signaling pathway	7.78	1.68×10^{-19}
hsa04012	ErbB signaling pathway	6.69	2.67×10^{-19}
hsa04917	Prolactin signaling pathway	5.79	3.57×10^{-17}
hsa04914	Progesterone-mediated oocyte maturation	6.33	3.70×10^{-17}
hsa04014	Ras signaling pathway	9.76	3.87×10^{-16}
hsa04550	Signaling pathways regulating pluripotency of stem cells	7.41	7.26×10^{-15}
hsa04919	Thyroid hormone signaling pathway	6.69	9.79×10^{-15}
hsa04350	TGF-beta signaling pathway	5.79	1.28×10^{-14}
	REACTOME Database		
R-HSA-69231	Cyclin D associated events in G1	4.7	5.00×10^{-21}
R-HSA-1538133	G0 and Early G1	3.44	1.13×10^{-15}
R-HSA-69656	Cyclin A:Cdk2-associated events at S phase entry	2.35	1.10×10^{-10}
R-HSA-69273	Cyclin A/B1 associated events during G2/M transition	2.71	1.42×10^{-10}
R-HSA-2173796	SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription	3.07	4.22×10^{-10}
R-HSA-1257604	PIP3 activates AKT signaling	4.34	5.16×10^{-9}
R-HSA-5674400	Constitutive Signaling by AKT1 E17K in cancer	2.53	4.01×10^{-8}
R-HSA-2219530	Constitutive Signaling by Aberrant PI3K in cancer	3.62	6.77×10^{-8}
R-HSA-69202	Cyclin E associated events during G1/S transition	1.99	9.93×10^{-8}
R-HSA-1912408	Pre-NOTCH Transcription and Translation	2.53	4.36×10^{-7}

2.3. Protein–Protein Interaction Analysis

We evaluated the physical interactions of the members of the consensus list by including the protein interactions described for *Homo sapiens* from the STRING database [20]. The protein–protein interaction (PPI) generated an osteosarcoma–PPI network (OS–PPI) of 505 nodes from the 553 consensus genes (91.3%). The node degrees of the 58 pathogenic genes (named as G1 and G2) detected in this network were higher than the non-pathogenic ones (39.05 and 19.25, respectively), showing statistical differences when applying the non-parametric Mann–Whitney *U*-test ($p < 0.001$). Therefore, a higher node degree given by this interaction signifies a greater probability of association with pathogenesis within the prioritized genes.

2.4. Community Analysis and Weight of Enriched Pathway

The community analysis was carried out using the clique percolation method. The clustering data through the community analysis was obtained with Cfinder [21], which defined “k-cliques” based on the interaction degree of each node from the OS–PPI network and the extent to which different communities overlapped in said network. The clique percolation method allowed us to detect 14 k-cliques and 86 possible communities with a composition of between 17 and 465 genes. The early minimum in S^k variation with respect to k-parameters (Figure 2) revealed that $k = 8$ and $k = 9$ have similar gene distributions within communities (S^k index 0.719 and 0.609, respectively). Both k-cliques are suitable for further analysis; however, we chose $k = 9$ because it had a better *Mean_rank* (218.89)

than $k = 8$ (243.95). Moreover, $k = 9$ is composed of 13 communities and 245 genes (44.3% of the 553 OS genes).

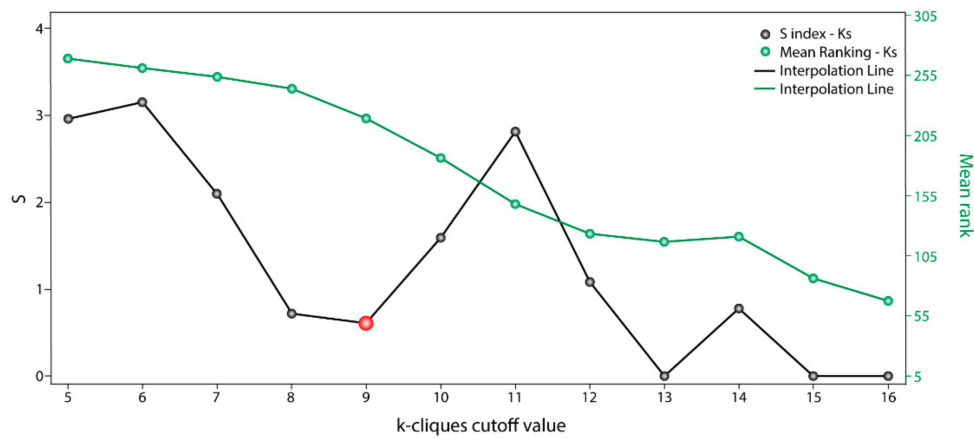


Figure 2. S^k scoring with respect to each k-clique cutoff value. Community analysis by clique percolation method. Values of S^k (black points) and mean rankings (green points) with respect to each k-clique cutoff value.

In order to weigh the metabolic pathways obtained in the enrichment analysis, we ranked these terms within each k-clique by means of a pathway enrichment analysis. The pathway enrichment analysis of genes in the 13 communities for $k = 9$ (Table S7) is consistent with the results obtained in the enrichment analysis (Table 4). As shown in Table 5, P53, cell cycle, and FOXO continue to hold the top positions and ErbB, TGFB and VEGF improved their statistical significance within this k-clique.

Table 5. Pathways enrichment analysis of $k = 9$ communities and their associated weights.

Pathway Name	PathScore _m	Community
p53 signaling pathway	0.603	2, 4, 9, 10
Cell cycle	0.595	2, 4, 7, 8, 9, 13
FoxO signaling pathway	0.578	2, 7, 8, 10, 11, 12, 13
Prolactin signaling pathway	0.574	2, 8, 10, 12
ErbB signaling pathway	0.565	2, 10, 11, 12, 13
Central carbon metabolism in cancer	0.564	2, 10, 11, 12, 13
TGF-beta signaling pathway	0.553	2, 6, 7, 8
Pathways in cancer	0.546	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
VEGF signaling pathway	0.536	2, 10, 11, 12
Adherens junction	0.534	2, 3, 6, 7, 8, 10, 11, 12
Proteoglycans in cancer	0.534	2, 10, 11, 12, 13
HIF-1 signaling pathway	0.532	2, 5, 6, 7, 8, 10, 11, 12, 13
Choline metabolism in cancer	0.526	2, 10, 11, 12
Thyroid hormone signaling pathway	0.524	1, 2, 3, 5, 6, 7, 10, 13
TNF signaling pathway	0.523	2, 5, 8, 13
NOD-like receptor signaling pathway	0.522	2, 8, 13
Osteoclast differentiation	0.52	2, 8, 11, 12, 13
Focal adhesion	0.518	2, 10, 11, 12, 13
Progesterone-mediated oocyte maturation	0.518	2
Apoptosis	0.515	2, 4, 5, 8, 9, 10, 13
Neurotrophin signaling pathway	0.515	2, 5, 10, 11, 12, 13
Fc epsilon RI signaling pathway	0.514	2, 10, 11, 12
MicroRNAs in cancer	0.508	2, 4, 8, 9, 10, 12, 13
mTOR signaling pathway	0.504	2, 10
B cell receptor signaling pathway	0.502	2, 5, 8, 10, 11, 12, 13

To be more selective about which communities would be most relevant within these 13, we used a clustering analysis. The K-means clustering analysis revealed four main community groups (Figure 3). Cluster 1 (Communities 4, 9 and 13) had the highest average values of $ConsenScore_i$, $Degree_i$ and $PathScore_m$, followed by Cluster 2 (Communities 5, 8 and 10) with regards to $ConsenScore_i$ and $Degree_i$. Therefore, these six communities were chosen for further analysis.

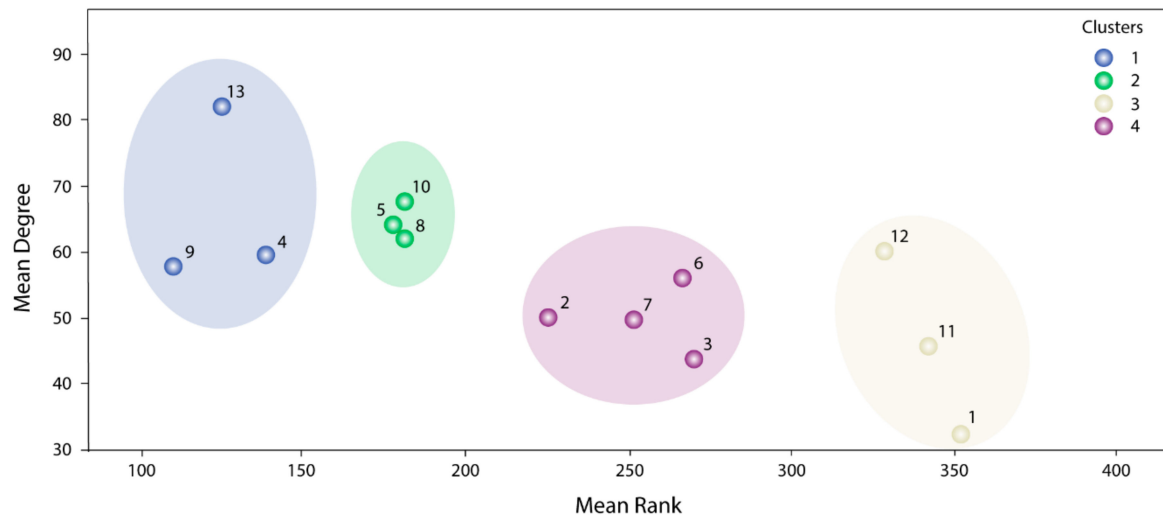


Figure 3. Clustering analysis for the $k = 9$ communities. Blue circles represent Cluster 1, purple circles Cluster 2, yellow circles Cluster 3 and purple circles represent Cluster 3.

Communities 4, 5, 8, 9, 10, and 13 have groups from 9 to 13 genes and, in total, contain 47 prioritized genes. The genetic distribution among the communities is almost specific and only Communities 4 and 9 present a high similarity (77%) regarding gene composition (Table 6). Only *TP53* is shared in five of the six communities, which denotes its centrality in this prioritization.

Table 6. Gene distribution in the most relevant communities in $k = 9$ -clique.

Comms	Genes	Mean $ConsenScore_i$	Mean $Degree$	Mean $PathScore_m$	Pathogenic Genes/Genes
9	<i>TP53, ATM, BRCA1, CHEK1, CDK2, ATR, BRCA2, RAD51, BLM</i>	0.802	57.78	0.656	0.333
13	<i>TP53, JUN, VEGFA, MYC, MMP2, BCL2, MMP9, NFKBL, IL6, FGF2, AKT1, TGFB1, CDH1</i>	0.776	81.85	0.598	0.692
4	<i>TP53, CDK4, ATM, BRCA1, CDK2, BRCA2, RAD51, MLH1, BLM</i>	0.751	59.33	0.656	0.444
5	<i>TP53, JUN, ATF2, CREBBP, SMARCB1, HMGB1, KAT2B, RELA, ARID1A, NR3C1, SMARCE1</i>	0.68	64	0.594	0.182
8	<i>NFKB1, SP1, CREBBP, CEBPB, CEBPD, STAT3, KLF4, EP300, RELA, PPARG, TGFB1</i>	0.675	62	0.612	0.273
10	<i>TP53, VEGFA, EGFR, PTK2, ERBB2, SHCL, PTEN, PIK3CA, HRAS, KRAS</i>	0.673	67.4	0.599	0.6

Genes in Communities 8, 10, and 13 are highly relevant for the signaling pathways PI3K/AKT and ERBB/MAPK (*PIK3CA, PTK2, HRAS, KRAS, SCH1, AKT*). In Community 13, the matrix metalloproteases

MMP2 and *MMP2* are prioritized, which together with *FGF2*, reflects processes related to cell migration. Since *AKT* is a central protein in cellular signaling, several downstream effectors are described in Communities 5 and 8. The genes *ARID1A*, *SMARCE1* and *SMARCB1*, specific to Community 5, are mainly associated with chromatin remodeling.

Given the close metabolic relationship between Communities 5, 8, 10, and 13, it is not surprising that *JUN*, *NFKB1*, *VEGFA*, *TGFB1*, *CREBBP*, and *RELA* are shared among them. However, Communities 4 and 9 are isolated from the rest of the clusters and only have *TP53* in common. The genetic composition of both communities is specific to one biological process: DNA repair. *ATM*, *CHEK1*, *ATR*, *BRCA1*, *BRCA2*, *RAD51*, *BLM*, and *MLH1* belong to DNA repair complexes associated with cellular response to DNA damage stimuli, DNA repair, and double-strand break repair via homologous recombination. Altogether, the genetic distribution of these communities is in accordance with the GO analysis obtained from our consensus list (Table 3).

The 47 genes grouped into the six communities defined above represent the most important prioritized members within this study, so we developed a sub-network based on these results (OS-comms network). The centrality index calculated in this sub-network was significantly correlated with the node degree ($Degree_i$) of the same genes in the original OS-PPI network ($r = 0.317$, $p = 0.03$).

2.5. Gene Validation

As a validation strategy, we compare our consensus list with the DRIVE project (deep RNAi interrogation of visibility effects in cancer) [22] and with the cancer-focused protein-protein interaction network (OncoPPI) [23] data. The data generated by the DRIVE project described 83.5% of our 553 consensus genes (Table S8). Of these 461 genes, 20 were determined as essential, 70 as active and 371 as inert. On the other hand, the OncoPPI network recognized 92 of our prioritized genes (16.6%) and its centrality index showed a significant correlation with the same gene in our OS-PPI network ($r = 0.445$, $p < 0.001$) (Table S9).

As shown in Figure 4A, both DRIVE and OncoPPI genes are present in the OS-comms network. From the DRIVE analysis, *BRCA1* and *RAD51* were identified as essential and *ATR*, *CDK2*, *CDK4*, *CHEK1*, *SMARCB1*, *SMARCE1*, *RELA*, *AKT1*, *MYC*, *HRAS* as active. On the other hand, 17 OncoPPI genes (36.18% of 47 in OS-comms network) were present in this network. Upon correlating the centrality indices between the OncoPPI network and the OS-comms network, we obtained a statistical correlation ($r = 0.512$, $p = 0.036$).

We can notice that in Figure 4B, several prioritized genes are actually transcription factors (TFs). Because of this, we chose to perform a second prioritization focused only on TFs and without using PPI networks. The PPI could bias toward physical interactions and reduce the relevance of regulatory mechanism, as presented in TFs. We identified 125 TFs from the initial 553 genes already prioritized. The $TFscore_i$ was evaluated for all TFs (Table S2). The top 20 more relevant TFs are *TP53*, *E2F1*, *JUN*, *RUNX2*, *FLI1*, *YY1*, *HIF1A*, *MYC*, *TP63*, *ESR1*, *WT1*, *E2F4*, *ATF2*, *NFKB1*, *AR*, *SP1*, *STAT1*, *ERG*, *CEBPB*, *TFAP2A*.

From the 125 TFs, 4% were identified as essential and 9.1% as active when compared to DRIVE genes. Additionally, 19 TFs were present in the OncoPPI network. Regarding community analysis, 27.6% were TFs and were mainly present within Communities 5, 8 and 13 (Figure 4B).

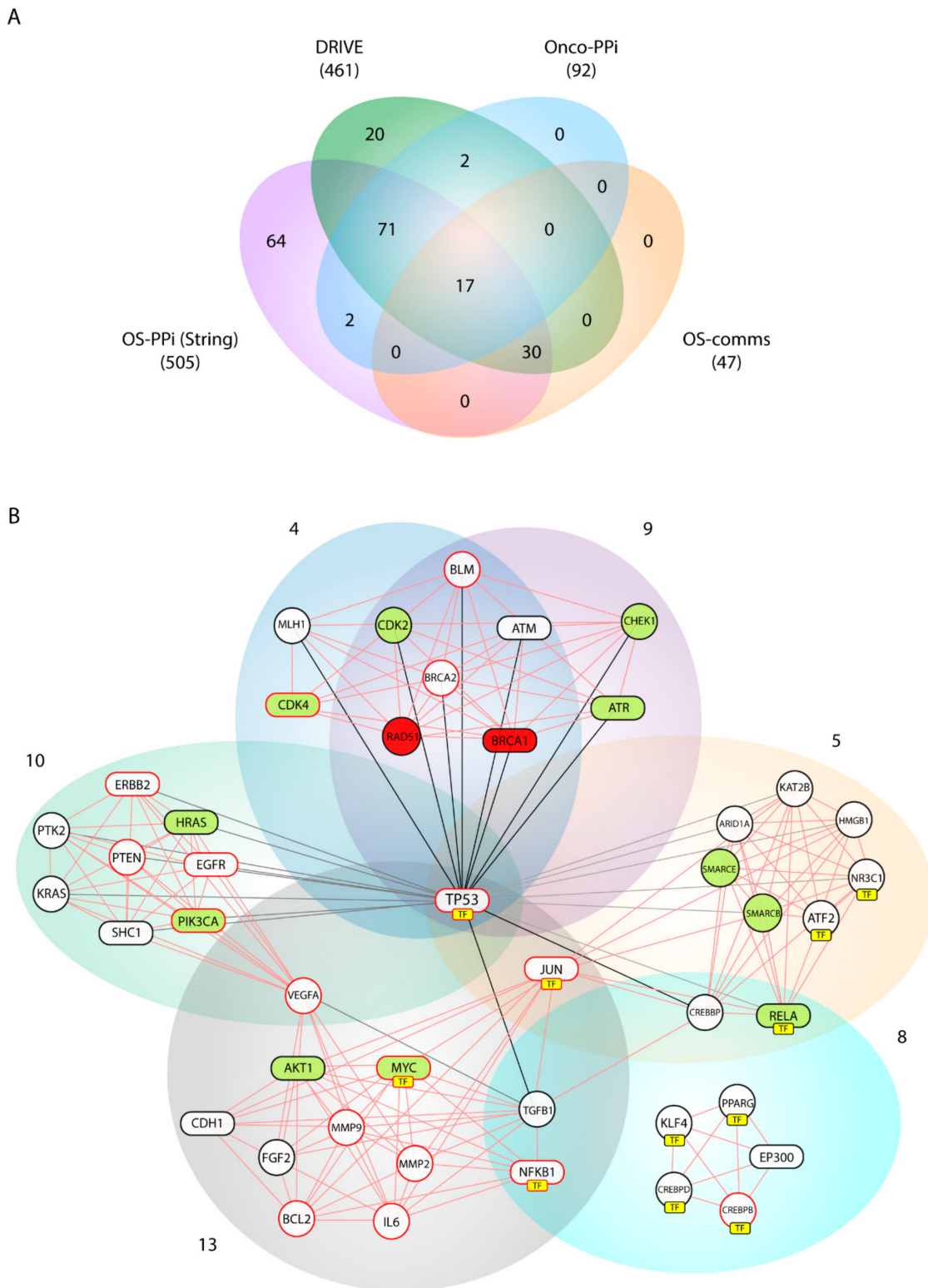


Figure 4. Gene validation and network analysis of the $k = 9$ -clique. **(A)** Comparison of prioritized genes from STRING (OS-PPI), DRIVE Project, OncoPPI network, and Cfinder analysis; **(B)** Network analysis from Communities 9, 13, 4, 5, 8, and 10 (OS-comms network). Red and green painted nodes are defined as essential and active genes, respectively, based on the results from the DRIVE project. Nodes enclosed in rectangles belong to the analyzed OncoPPI network. Nodes with red borders are members of G1 and G2. Yellow boxes (TF) point to nodes identified as transcription factors.

3. Discussion

As shown in Table 1, the detection rate of our consensus prioritization strategy was higher than all the bioinformatics tools employed in this analysis. Moreover, the mean rank of the pathogenic genes detected in the top 1% of the list was 49.3. Table 2 indicates that, on average, the 45 G1–G2 genes were located in the top 50 positions. These results confirm that this methodology does indeed improve the detection and prioritization of pathogenic genes, as had been previously described in other pathologies [24,25].

As a first approach, the prioritization strategy resulted in a consensus list of 553 genes and the 10 top-ranked genes were *TP53*, *RB1*, *CHEK2*, *RUNX2*, *E2F1*, *MDM2*, *CDKN1A*, *JUN*, *CCNA2* and *CDKN2A*. *TP53*, *RB1*, *CHEK2*, and *MDM2* were ranked in 1st, 2nd, 3rd, and 6th positions, respectively, and also the arrangement of the pathogenic genes in this list shows a distribution in the top positions. So far, the gene ranking along this prioritization reflects a proper gene weighting based mainly on this consensus strategy. These genes had been previously described in OS pathogenesis. Early studies focused on the molecular biology of OS were carried out on individuals with familial syndromes, which predisposed them to this tumor. Germline inactivation of *RB1* and *TP53* were initially described in patients with hereditary retinoblastoma and Li–Fraumeni syndrome, respectively [26,27], and subsequently in sporadic sarcomas [28,29]. Given that these two suppressors are central proteins in controlling the cell cycle, later studies briefly described many others that interacted with them. Mouse double minute 2 (*MDM2*), for example, is a protein that binds to *RB1* and inactivates *TP53* [30]. Its amplification is an event that occurs in primary OS (3–25%) and it is overexpressed in metastases and recurrences [31,32]. *CHEK2* is another protein that is part of a DNA damage checkpoint, works as a stabilizer of *TP53*, and shows a 7% frequency of mutations in OS patients [33,34].

The biological processes derived from the GO analysis of the 553 genes describe *TP53* as a principal signal transducer that mediates processes associated with cell cycle, DNA damage response, DNA replication and intrinsic/extrinsic apoptotic signaling regulation. Additionally, more specific biological processes were described, for instance, fibroblast proliferation, osteoblast differentiation and development, and mesenchymal cell proliferation and transition. In accordance with our results, previous studies have identified similar biological processes related to OS, where the following are considered OS-associated terms: cell cycle regulation (mainly mediated by *RB1* and *TP53*), osteoblast differentiation (mediated by *RUNX2*), DNA damage, stress response, epigenetic processes, mitosis, cell motility functions, and members involved in OS cell proliferation (weighting *NFKB* signaling, *NFKBIE*, and *RELA* members) [3,35–37]. Taken together, these processes suggest that the consensus list is evidence of the genes associated with osteogenesis, cell differentiation, and transition to bone cell types. In addition, the terms derived from the pathway enrichment analysis (Table 4) are in accordance with these biological processes.

The information used by STRING allowed us to define the degree of physical interaction of the consensus list members and calculate their centrality index. This centrality index was used as a variable to evidence the contribution rate of the pathogenic genes to a common biological purpose. Thus, the greater the centrality for a node within the OS–PPI network, the greater the probability of its contributing to pathogenesis. This association was validated by analyzing the genes defined as pathogenic (G1–G2), in which significant differences were observed in comparison with the rest of the consensus genes ($p < 0.0001$). The centrality index calculated from the 503 nodes included in the protein–protein interaction network determined *TP53* as the most central node, followed by *AKT1*, *MYC*, *JUN*, *EP300*, *CREBBP*, *CCND1*, *CDKN1A*, *STAT3*, and *RB1*. Furthermore, this degree allowed for the definition of more specific clusters and prioritization of gene communities associated with OS pathogenesis. Thus, $k=9$ was determined as the clique with the best gene distribution among all the resulting communities (S^k index 0.719) and Communities 4, 5, 8, 9, 10, and 13 as the most important groups of genes within our study.

The pathway enrichment analysis for the $k=9$ -clique results in, almost in its entirety, the same terms obtained from the initial consensus list. This confirms that the gene filtered through the communality

analysis comprised almost the same biological processes. Considering the $PathScore_m$ (Table 5), the P53 signaling pathway and cell cycle are in the top positions. FOXO also increases its significance in this enrichment analysis. In different cancer types, PI3K/AKT, Ras-MEK-ERK, IKK, and AMPK are the most important signaling pathways interacting with FOXO [38]. The gain of function of P13K and RAS, or PTEN disruption, are oncogenic events that promote a loss of function in the Forkhead Box transcription factors (FOXO) [39]. Interestingly, loss of its expression promotes impaired osteogenic differentiation, suggesting that *FOXO1* is involved in osteoblastogenesis and osteoclastogenesis [40–42]. Moreover, FOXO members have an important role in cell fate decision, via triggering the expression of death receptor ligands like FASLG, TNF apoptosis ligand, and some BCL-2 family members (*BCL2L1*, *BNIP3*, *BCL2L11*) [43–46]. FOXO expression in OS tumors is low or even lacking altogether, leading to tumor progression and cell cycle arrest [47]. The fact that FOXO enhances its weight within our enrichment analysis demonstrates its importance as a signaling pathway in the pathogenesis of OS. Furthermore, the close relationship between the FOXO signaling pathway and cell cycle, events of osteoclast differentiation and apoptosis via the TNF signaling pathway, is evidenced in the pathway enrichment analysis applied to the consensus list and the $k = 9$ clique.

Our consensus strategy seeks to specify a group of genes that describe the molecular etiology of OS. In this sense, the use of all the strategies previously described prioritizes to a great extent the 47 genes arranged in Communities 4, 5, 8, 9, 10, and 13. From these six communities, *BRCA1*, *AKT1*, *ATR*, *CDK4*, *HRAS*, *MYC*, *PIK3CA*, *RELA*, *STAT3* are genes validated by DRIVE and Onco-PPI (19.1%), *RAD51*, *CDK2*, *CHEK1*, *SMARCB1*, *SMARCE1* are validated only by DRIVE (10.6%), and *ATM*, *CDH1*, *EGFR*, *EP300*, *ERBB2*, *JUN*, *NFKB1*, *SHC1*, *TP53*, *SP1* by Onco-PPI (21.3%). The sub-network generated from these communities (OS-comms network) reflects closely interrelated genes at the cellular interaction level (Figure 4B) and also groups of genes immersed in important oncological processes. Tamborero et al. [48], from exome sequencing data of 3205 tumors in the Cancer Genome Atlas (TCGA) research network, proposed 291 high-confidence cancer driver genes acting on 12 different cancer types. Although in this study, data from samples of bone tumors were not taken into account, their results showed the members of the PI3K signaling pathway as central onco-drivers, *ATR*-*BRCA1* as regulatory nodes of repair processes associated with *TP53*, *CHEK1* and *AKT* as the main regulators of cell cycle in function of *CDK1A*, and *CDK1B* and activators for downstream pathways such as FOXO. This experimental data support our findings, where *PIK3CA*, *AKT1*, *PTEN*, *HRAS* and *SHC1* were nodes highly connected within our OS-comms network. Nodes that connect to Communities 10 and 13 describe genes representative of our weighted tumorigenic pathways, PI3K/AKT and MAPK/ERK.

The findings reported here suggest that PI3K/AKT and MAPK/ERK are the main signaling pathways deregulated for OS. Several reports have shown that these pathways are responsible for controlling cellular processes related to proliferation, growth, differentiation, and apoptosis [49,50]. In fact, the Ras/Raf/MEK/ERK pathway is hyperactivated in 30% of human cancers [51] and nearly 67% of OS shows aberrant ERK activation [52]. The extra cellular-signal-regulated kinases (ERK) promote cell proliferation, cell survival, and metastasis, particularly by its upstream activation from EGFR and the G protein-coupled receptor Ras [53]. The presence of *SHC1*, *EGFR*, *HRAS*, *PIK3CA*, *ERBB2* within Community 10 support this scenario for OS. In addition, the high connectivity of the matrix metalloproteases, *MMP2* and *MMP9*, in Community 13 suggests a metastasis event in the function of these signaling pathways.

Although the invasion of tumor cells is a general characteristic in carcinogenesis, metastasis to the lung is one of the main characteristics in patients with OS and one of the major causes of mortality [54,55], so this event is a hallmark for this sarcoma. Pathogenic events, including cellular detachment from primary tumors, matrix remodeling and invasion from tumor cells, angiogenesis, vascular dissemination, and proliferation at new sites, are involved in tumor metastasis [56,57]. Upstream regulators of MAPK/ERK signaling such as *IL6*, *VEGFA*, and *FGFR1* demonstrate an important role in this process [58–62] and are prioritized in our results. In addition, Community 13 shows the

MMP2 and *MMP9* genes with a high centrality index. A high expression of *MMP9* was observed in metastatic OS samples [63,64], leading to speculation that this metalloproteinase can promote cell migration and invasion in OS by degradation components of the extracellular matrix. This evidence suggests that *MMP2* and *MMP9*, together with upstream regulators of MAP/ERK signaling such as *IL6*, *FGF2*, *VEGFA*, *EGFR* and *ERBB2*, are pathogenic nodes dependent on the centrality of PI3K/AKT and MAPK/ERK. This finding could be related to aspects of invasiveness and prognosis, mainly in tumors that present deregulation in these two signaling pathways.

In addition to evidencing the previous findings, Communities 4, 5 and 9 include genes widely described in processes of homologous recombination (HR), base excision repair, and chromatin modification. Cells DNA damage response principally involves maintaining chromosome integrity and genome stability and implies recognition of DNA lesions, followed by an activation of the checkpoints in the cell cycle that promotes cellular signaling cascades related to DNA repair. While the ATM-CHEK2 pathway is responsible for the initiation of cellular responses to double-strand breaks [65,66], ATR-CHEK1 responds to DNA replication stress by means of the phosphorylation of several substrates in response to agents such as UV and X-ray among others [67]. *ATM*, *ATR*, and *CHEK1* show a high centrality index in the OS-comms network, interacting in addition to *BRCA1* and *RAD51*, described as essential genes, and with the cyclin-dependent kinases, *CDK2* and *CDK4*, described as active ones according to the DRIVE validation. Checkpoint activation by ATM mainly controls G1/S, whereas ATM and ATR contribute to establishing and maintaining the S and G2/M checkpoints [68]. Either by activation of ATR-CHEK1 or ATM-CHEK2, DNA damage signaling promotes inhibition of CDK activity and therefore the activation of G1/S, intra-S, and G2/M checkpoints [69]. Consequently, it is likely that such nodes associated with DNA repair, such as *ATM*, *ATR*, *CHEK1*, *BLM*, *RAD51* and *MLH1* (as shown in our pathway enrichment analysis), together with those previously described (*BRCA1* and *BRCA2*) from exome sequencing [70], have important implications regarding the deregulation of the cell cycle evidenced in OS.

While it is true that the nodes described for Communities 4 and 9 are mainly related to repair and cell cycle control events, the HR repair complex is involved in a hallmark event for sarcomas, such as alternative telomere maintenance (ALT). Several molecular details of this mechanism still remain unknown; however, two distinctive telomere phenotypes are described for ALT in human telomerase-negative cells (ALT cells) such as long and heterogeneous telomere DNA and promyelocytic leukemia (PML) body [71], together forming the ALT-associated promyelocytic leukemia body (APB). The PML body is a nuclear made up of proteins which form amongst the chromatin and is related to a wide range of cellular processes including tumors formation, cellular senescence, and DNA repair [72,73]. Numerous lines of evidence strongly suggest that the ALT pathway is dependent on HR since several proteins involved in DNA double-strand break (DSB) are localized at APBs [74–77]. It is significant that proteins localized at APBs, such as PML, DNA helicases of the RecQ family (*BLM*, *WRN* and *RECQL4*), *RAD51* and *RAD52* (a member of the MNR complex), rank highly in our prioritization. In this sense, the members belonging to HR complexes are described as repair complexes in response to DNA damage. They are relevant to the pathogenesis of the OS, not only as factors immersed in cell cycle control, as previously discussed, but also because they are involved in processes of chromosome stability given by telomere maintenance [78–81]. Consistent with the literature, where bone tumors are termed as highly heterogeneous, highly mutable, and genetically unstable, members described in Communities 4 and 9 (*TP53*, *ATM*, *ATR*, *CHEK1*, *BLM*, *BRCA1*, *BRCA2*, *RAD51*, *MLH1*, *CDK2*, *CDK4*) explain many of these key features within OS, and can also be associated with important clinical characteristics such as tumor aggressiveness, metastasis, and poor survival.

The use of the GTRD database allowed us to define the frequency of interaction of each TF with the 553 prioritized genes. It is worth noticing that more than half of the prioritized factors (103, 82.4%) interacted with more than half of all genes at the same time. This suggests that more than 80% of the genes defined as TFs actively regulated the genes associated with the pathogenesis of OS. The weight given to each one of these TFs through interaction analysis places the following genes on the top

positions: *TP53*, *E2F1*, *JUN*, *RUNX2*, *FLI1*, *YY1*, *HIF1A*, *MYC*, *TP63*, *ESR1*, *WT1*, *E2F4*, *ATF2*, *NFKB1*, *AR*, *SP1*, *STAT1*, *ERG*, *CEBPB*, and *TFAP2A*. When compared to total prioritization, genes *E2F1*, *JUN*, *RUNX2*, *FLI1*, *YY1*, *HIF1A*, *MYC*, *TP63*, *ESR1*, *WT1*, *E2F4*, *ATF2*, and *NFKB1* significantly improved their ranking. During the G1 phase of the cell cycle, RB1 suppresses the function of the E2F1, E2F2, and E2F3 TFs. Sequential hypo phosphorylation of RB1 by cyclin-dependent kinases, CDK4 and CDK6, and CDK2, led in the release of E2F and transcription of genes necessary for cell cycle progression, including cyclins A, D, and E [82]. The improved score of these TFs suggests that these deregulation events in the cell cycle are basal within the pathogenesis of the OS. Although this scenario is common for all types of cancer, a deeper study of the *E2F1* and *E2F4* genes, and depending on those prioritized in Communities 4 and 9 along with *TP53*, would be necessary to define driver proteins in OS tumors.

We identified *TP53*, *JUN*, *MYC*, *ATF2*, *NFKB1*, *SP1*, *CEBPB*, *STAT3*, *KLF4*, *RELA*, *NR3C1*, *CEBPD*, and *PPARG* as TFs (13 or the 47 nodes) in the OS-comms network. The new ranking calculated for each of them improved significantly when compared to the ranking of all OS genes (Table S2). This suggests that their degree of regulation within this network is very significant and shows evidence of its importance as regulatory proteins within each prioritized cluster.

TFs were grouped over Communities 5, 8 and 13. With *TP53* as the central node, *JUN* and *MYC* are key factors in the pathogenesis of the OS that regulate signaling associated with the pathogenic pathways PI3K/AKT and MAPK/ERK. Furthermore, the prioritization of TFs evidenced *NFKB1* as a central node in these three communities. Nuclear factor-kappa B1 (*NFκB1*) is a pleiotropic transcription factor that contributes to tumorigenesis in many types of cancer. It works as a key regulator of a variety of genes implicated in many biological events including cell survival, differentiation, apoptosis, and autophagy [83]. When observing the OS-comms network, the high degree interaction of *AKT* with respect to *JUN*-*MYC*, *TGFB1*, *NFKB1*, and *BCL2* suggests this cluster as an important group in the OS pathogenesis. GO terms listed in Table 3 are in accordance with these findings since its activation promotes many types of downstream signaling including osteoblast differentiation via *TGFB1* and *NFK1* or apoptosis via *BCL2* [84,85].

In conclusion, the use of a consensus strategy proved to be efficient when specifying a broad list of genes obtained from several bioinformatics prioritization tools. In addition, the combination of these strategies with a network enrichment analysis allowed us to show not only real interactions between specific genes but also to define internal interactions that explained cellular events associated with OS pathogenesis. Our results validate several studies that describe the signaling pathways PI3K/AKT and MAPK/ERK as oncological for OS. Nevertheless, given its centrality at the cellular signaling level, its deregulation can influence downstream specific pathways, such as FOXO, and promote tumorigenic scenarios like osteoblast undifferentiation via *TGFB1* and *NFK1*, apoptosis via *BCL2*, and migration and metastasis mediated mainly by *MMP2* and *MMP9*.

What is more, the gene composition of Communities 4 and 9, and more specifically to their *ATM*, *ATR*, *CHEK1*, and *RAD51* genes, suggest that the HR repair complex is an important group of genes within the pathogenesis of the OS. Its deregulation can influence tumorigenic events characteristic of this sarcoma as generalized disruption in the cell cycle and ALT mechanisms. Hence, it is necessary to experimentally validate these results, taking into account not only the patient's age group but also genetic factors that can influence the molecular behavior of these bone tumors, such as racial and ethnic factors. It should also be interesting to study genetic variants of the transcription factors identified and their relationship with possible disease prevalence.

4. Materials and Methods

4.1. Prioritization Methods and Consensus Strategy

The bioinformatics methods used in this study were for gene-disease prioritization Biograph [86], Cipher [87], DisGeNET [88], Génie [89], GLAD4U [90], Guildify [91], Phenolizer [92], PolySearch [93], and SNPs3D [94]. We chose these nine bioinformatics methods because (1) they are fully available

on web service platforms and (2) they only required the disease name (or OMIN code, 259,500 for OS) for gene prioritization. With the disease name and/or the OMIM code, a list of prioritized genes was obtained from each method. Each of these methods follows several different strategies for gene prioritization, and as a final output, they also provide different scores for each gene.

The strategy applied to integrate the gene scores obtained in each independent method is similar to that previously described [24,25]. Thus, we normalized each gene (denoted as i) from the ranked list obtained from each method (denoted as j) ($GeneN_{i,j}$ which means, the normalized score of the gene “ i ” in the method “ j ”). The final score by gene ($ConsenScore_i$) was considered as the average normalized score and the number of methods which predict the gene (denoted as n_i) are

$$ConsenScore_i = \sqrt{\left(\frac{(n_i-1)}{9-1}\right)\left(\frac{1}{j} \sum_j GeneN_{i,j}\right)} \quad (1)$$

This equation refers to the geometric mean between the average score of each gene derived from each method, and the normalized score according to the number of methods that predict the association of the gene and the disease. This consensus approach will lead to a big final list of genes ranked according to the $ConsenScore_i$. In order to reduce this list, we needed to follow some rational strategy.

From a manual observation and curation of the scientific literature, we create a list of genes that are highly probable to be involved in OS pathogenesis (Table S1). For this, we took into consideration pathogenic OS genes defined by a literature review of two types of studies: meta-analysis, based on publications and case reports for OS patients (named as G1 genes), and gene description in animal models and OS cell lines (named as G2 genes). Thereby, we identified 75 pathogenic OS genes from the available literature, of which 47 were classified as G1 and 41 as G2. These manually curated genes were used for (1) validation of the prioritized genes (and networks) and (2) to reduce the initial list of consensus genes.

The pathogenic OS genes (defined as G1 and G2 in Table S1) were used to calculate $I_i = \frac{TP_i}{FP_i+1} ConsenScore_i$, where TP and FP are the true and false positive values (up to the ranking value of the gene i), respectively. According to that which has been previously described [24,25], the maximum value of I_i can be taken as the maximum compromise between the TP and FP rates compensated with the ranking index of each gene. The ranking (“ i ”), at which “ I_i ” is maximal, will represent a rational cut-off for the consensus list.

We applied another prioritization methodology to demonstrate the degree of interaction of all transcription factors in our consensus genes. We used the “The Human Transcription Factors” database [95] to identify the TFs from the 553 initially prioritized genes. The second prioritization of only TFs was carried out considering the $ConsenScore_i$ (Equation (1)) for each TFs and the interaction degree of each TF using the information described in the GTRD (Gene Transcription Regulation Database) [96]. This database contains experimental information from ChIP-seq experiments of TF binding sites. Data were systematically collected and uniformly processed using a special workflow (pipeline) for a BioUML platform (<http://www.biouml.org>). By inspecting all the target genes described for *Homo sapiens*, we downloaded the information of all the genes defined as TF and all the genes or +/-5000bp that contain a GTRD meta cluster for this TF.

Thus, the $TFscore_i$ was calculated as

$$TFscore_i = \sqrt{\left(\frac{(t_i-1)}{553-1}\right) ConsenScore_i} \quad (2)$$

where t_i is the number of genes that are regulated by the transcription factor “ i ”. The general conception is that a transcription factor will be more relevant if it has a higher value in the consensus score and regulate many of the prioritized genes.

4.2. Protein–Protein Interaction Network Analysis

The protein interactions of the members of the consensus list were revised from the STRING database, only taking into consideration interactions with a confidence cut-off of 0.9. With this information, we generated a OS–PPI network with zero node addition. Network visualization and analysis were carried out through the Cytoscape software [97].

4.3. Communalities and Pathway Enrichment Analysis

The communalities analysis on OS–PPI was carried out using the clique percolation method with Cfinder [21]. The communalities analysis provides a topology description of the network including the location of highly connected sub-graphs (cliques) and/or overlapping modules that usually correspond with relevant biological information. The selection of the value “k-cliques” ($k = 1, 2, 3 \dots n$) will affect the number of community and also the number of genes in each community. In general, higher values of k-cliques imply few communities while lower values lead to many communities. In the OS–PPI network, both extremes (too small or too high k-cliques values) result in an unbalanced distribution of the genes across communities. This means some of the communities will have a big amount of genes while others will have a very small number.

In order to determine the best k-clique in the communalities analysis, we used the index “S” [24,25]: $S^k = \frac{|\text{mean}(N_g^k) - \text{median}(N_g^k)|}{N_c^k}$, where N_g^k and N_c^k are the number of genes in each community and the number of communities for a defined k-clique cut-off value. If the distribution of genes across communities is close to a Gaussian distribution, or constant, S^k will tend toward 0. Once k is defined in k-clique, a number of communities will be identified.

Additionally, we applied the partition algorithm K-means in order to define our best communities within a k-clique. The variables used for the clustering were the means of *ConsenScore_i*, *Degree_i*, and *PathScore_m* for each community within the k-clique. The *Degree_i* variable refers to the node’s degree centrality index calculated for each gene from the OS–PPI network and the *PathScore_m* is outlined below. From communities selected in this clustering, we created a sub-network to visualize the interactions of all the members of the chosen communities.

For the pathway enrichment analysis, we used a *PathRankScore_m*, *PathGeneScore_m*, and *PathScore_m* as described previously [24]: (1) Each community “k” was weighted as $W_k = \sum \text{ConsenScore}_i^k / N_k$, where ConsenScore_i^k is the *ConsenScore_i* of the gene “i” in the community “k” and N_k is the number of communities; (2) Each pathway “m” was weighted as $\text{PathRankScore}_m = \sum W_k^m / N_k^m$, where W_k^m is the weight (W_k) of each community connected with the pathway “m” and N_k^m is the number of communities connected with the pathway “m”, and (3) A second weight to the pathway “m”, *PathGeneScore_m*, considered all the genes included in each pathway: $\text{PathGeneScore}_m = \sqrt{\text{ConsenScore}_i^m \frac{n_m}{N_m}}$, where N_m is the total number of genes in the pathway “m”, while n_m is the number of those genes that are also found in the protein–protein interaction network. The average of the *ConsenScore_i* of all genes presents in the pathway “m” is $\langle \text{ConsenScore}_i^m \rangle$. The geometrical mean between *PathGeneScore_m* and the normalized *PathRankScore_m* refers to the final score associated with the pathway “m” (*PathScore_m*).

4.4. Gene Validation with the OncoPPI OS Network and the DRIVE Project

Besides the genes in the G1 and G2 groups, we also used the information in the DRIVE project. It is a project that describes a comprehensive mapping of cancer genes obtained from a larger-scale gene knockdown experiment in 398 cancer cell lines. We filtrated the results of eight cell lines, all of which had pathological annotations related to bone cancer (A673, SAOS2, SJS1, SKES1, SKNMC, SW1353, TC71, and U2OS). Subsequently, all essential genes that showed a *Sensitivity Value* of ≤ -3 in $>50\%$ of the chosen cell lines, active genes that showed values of ≤ -3 in 1–49%, and inert ones showed values of ≤ -3 for 0% of cancer cells [22] were compared with our results.

Additionally, from Onco-PPI Portal (<http://oncoppi.emory.edu/>) [23], a cancer-focused protein-protein interaction network was generated by only considering the interactions described for bone tumor types (labeled as OncoPPI). This network was comprised of 171 genes and 442 interactions. The Spearman correlation of *Degree_i* between the OncoPPI, OS-PPI, and the sub-network from the identified communities were calculated.

Supplementary Materials: Supplementary Materials can be found at <http://www.mdpi.com/1422-0067/21/3/1053/s1>. Table S1. Description of pathogenic OS genes. Table S2. Consensus OS gene list. Table S3. Biological processes by enrichment analysis in OS consensus gene list. Table S4. Biological processes of the consensus gene list using Revigo. Table S5. The enrichment analysis of the KEGG pathways of the consensus gene list. Table S6. The enrichment analysis of the Reactome pathways of the consensus gene list. Table S7. Pathway enrichment analysis of $k = 9$ of communities and their associated weights. Table S8. Classification of essential, active and inert genes based on the DRIVE project. Table S9. OncoPPI nodes present in the OS integrated network.

Author Contributions: E.T. and A.C.-A. conceived the project and wrote the manuscript. E.T. designed the algorithm. A.C.-A. and A.L.-C. implemented the algorithm and performed the data analysis. G.J.-K., C.P.-y.-M. and Y.P.-C. made substantial contributions in the discussion of the article. C.R.M., H.G.-D. and A.P. helped with study design and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by “Collaborative Project in Genomic Data Integration (CICLOGEN)”, grant number PI17/01826, “Consolidation and Structuring of Competitive Research Units—Competitive Reference Groups”, grant number ED431C 2018/49 and “Accreditation, Structuring, and Improvement of Consolidated Research Units and Singular Centers”, grant number ED431G/01, funded by the Ministry of Education, University and Vocational Training of the Xunta de Galicia endowed with EU FEDER funds.

Acknowledgments: This work was supported by Universidad de Las Américas (Quito, Ecuador), Hospital de Especialidades Eugenio Espejo (Quito, Ecuador), University of Coruna (Coruña, Spain) and University of the Basque Country (Bilbao, Spain).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

OS	Osteosarcoma
TF	Transcription factor
GO	Gene ontology
PPI	Protein-protein interaction
OS-PPI	Osteosarcoma protein-protein interaction network
OS-comms network	Sub-network based on the communality analysis
OncoPPI	Protein-protein interaction network obtained from the DRIVE project
HR	Homologous recombination
ALT	Alternative telomere maintenance

References

1. Man, T.K.; Lu, X.Y.; Jaeweon, K.; Perlaky, L.; Harris, C.P.; Shah, S.; Ladanyi, M.; Gorlick, R.; Lau, C.C.; Rao, P.H. Genome-wide array comparative genomic hybridization analysis reveals distinct amplifications in osteosarcoma. *BMC Cancer* **2004**, *4*, 45. [[CrossRef](#)] [[PubMed](#)]
2. Savage, S.A.; Mirabello, L.; Wang, Z.; Gastier-Foster, J.M.; Gorlick, R.; Khanna, C.; Flanagan, A.M.; Tirabosco, R.; Andrulis, I.L.; Wunder, J.S.; et al. Genome-wide association study identifies two susceptibility loci for osteosarcoma. *Nat. Genet.* **2013**, *45*, 799–803. [[CrossRef](#)] [[PubMed](#)]
3. Kuijjer, M.L.; Hogendoorn, P.C.; Cleton-Jansen, A.M. Genome-wide analyses on high-grade osteosarcoma: Making sense of a genomically most unstable tumor. *Int. J. Cancer* **2013**, *133*, 2512–2521. [[CrossRef](#)] [[PubMed](#)]
4. Groisberg, R.; Roszik, J.; Conley, A.; Patel, S.R.; Subbiah, V. The Role of Next-Generation Sequencing in Sarcomas: Evolution From Light Microscope to Molecular Microscope. *Curr. Oncol. Rep.* **2017**, *19*, 78. [[CrossRef](#)]
5. Cote, G.M.; He, J.; Choy, E. Next-Generation Sequencing for Patients with Sarcoma: A Single Center Experience. *Oncologist* **2018**, *23*, 234–242. [[CrossRef](#)]

6. Joseph, C.G.; Hwang, H.; Jiao, Y.; Wood, L.D.; Kinde, I.; Wu, J.; Mandahl, N.; Luo, J.; Hruban, R.H.; Diaz, L.A., Jr.; et al. Exomic analysis of myxoid liposarcomas, synovial sarcomas, and osteosarcomas. *Genes Chromosom. Cancer* **2014**, *53*, 15–24. [[CrossRef](#)]
7. Bousquet, M.; Noirot, C.; Accadbled, F.; Sales de Gauzy, J.; Castex, M.P.; Brousset, P.; Gomez-Brouchet, A. Whole-exome sequencing in osteosarcoma reveals important heterogeneity of genetic alterations. *Ann. Oncol.* **2016**, *27*, 738–744. [[CrossRef](#)]
8. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A., Jr.; Kinzler, K.W. Cancer genome landscapes. *Science* **2013**, *339*, 1546–1558. [[CrossRef](#)]
9. Lawrence, M.S.; Stojanov, P.; Mermel, C.H.; Robinson, J.T.; Garraway, L.A.; Golub, T.R.; Meyerson, M.; Gabriel, S.B.; Lander, E.S.; Getz, G. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **2014**, *505*, 495–501. [[CrossRef](#)]
10. Chen, X.; Bahrami, A.; Pappo, A.; Easton, J.; Dalton, J.; Hedlund, E.; Ellison, D.; Shurtleff, S.; Wu, G.; Wei, L.; et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep.* **2014**, *7*, 104–112. [[CrossRef](#)]
11. Gianferante, D.M.; Mirabello, L.; Savage, S.A. Germline and somatic genetics of osteosarcoma-connecting aetiology, biology and therapy. *Nat. Rev. Endocrinol.* **2017**, *13*, 480–491. [[CrossRef](#)] [[PubMed](#)]
12. Rickel, K.; Fang, F.; Tao, J. Molecular genetics of osteosarcoma. *Bone* **2017**, *102*, 69–79. [[CrossRef](#)] [[PubMed](#)]
13. Tranchevent, L.C.; Capdevila, F.B.; Nitsch, D.; De Moor, B.; De Causmaecker, P.; Moreau, Y. A guide to web tools to prioritize candidate genes. *Brief. Bioinform.* **2011**, *12*, 22–32. [[CrossRef](#)] [[PubMed](#)]
14. Bornigen, D.; Tranchevent, L.C.; Bonachela-Capdevila, F.; Devriendt, K.; De Moor, B.; De Causmaecker, P.; Moreau, Y. An unbiased evaluation of gene prioritization tools. *Bioinformatics* **2012**, *28*, 3081–3088. [[CrossRef](#)] [[PubMed](#)]
15. Moreau, Y.; Tranchevent, L.C. Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nat. Rev. Genet.* **2012**, *13*, 523–536. [[CrossRef](#)] [[PubMed](#)]
16. Chen, J.; Aronow, B.J.; Jegga, A.G. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinform.* **2009**, *10*, 73. [[CrossRef](#)]
17. Huang da, W.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4*, 44–57. [[CrossRef](#)]
18. Huang da, W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37*, 1–13. [[CrossRef](#)]
19. Supek, F.; Bosnjak, M.; Skunca, N.; Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **2011**, *6*, e21800. [[CrossRef](#)]
20. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P.; et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447–D452. [[CrossRef](#)]
21. Palla, G.; Derenyi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, *435*, 814–818. [[CrossRef](#)] [[PubMed](#)]
22. McDonald, E.R., 3rd; de Weck, A.; Schlabach, M.R.; Billy, E.; Mavrakis, K.J.; Hoffman, G.R.; Belur, D.; Castelletti, D.; Frias, E.; Gampa, K.; et al. Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **2017**, *170*, 577–592. [[CrossRef](#)] [[PubMed](#)]
23. Li, Z.; Ivanov, A.A.; Su, R.; Gonzalez-Pecchi, V.; Qi, Q.; Liu, S.; Webber, P.; McMillan, E.; Rusnak, L.; Pham, C.; et al. The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nat. Commun.* **2017**, *8*, 14356. [[CrossRef](#)] [[PubMed](#)]
24. Tejera, E.; Cruz-Monteagudo, M.; Burgos, G.; Sanchez, M.E.; Sanchez-Rodriguez, A.; Perez-Castillo, Y.; Borges, F.; Cordeiro, M.; Paz, Y.M.C.; Rebelo, I. Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis. *BMC Med. Genomics* **2017**, *10*, 50. [[CrossRef](#)] [[PubMed](#)]
25. Lopez-Cortes, A.; Paz, Y.M.C.; Cabrera-Andrade, A.; Barigye, S.J.; Munteanu, C.R.; Gonzalez-Diaz, H.; Pazos, A.; Perez-Castillo, Y.; Tejera, E. Gene prioritization, communality analysis, networking and metabolic integrated pathway to better understand breast cancer pathogenesis. *Sci. Rep.* **2018**, *8*, 16679. [[CrossRef](#)]
26. Srivastava, S.; Wang, S.; Tong, Y.A.; Hao, Z.M.; Chang, E.H. Dominant negative effect of a germ-line mutant p53: A step fostering tumorigenesis. *Cancer Res.* **1993**, *53*, 4452–4455.

27. Friend, S.H.; Bernards, R.; Rogelj, S.; Weinberg, R.A.; Rapaport, J.M.; Albert, D.M.; Dryja, T.P. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **1986**, *323*, 643–646. [[CrossRef](#)]
28. Tsuchiya, T.; Sekine, K.; Hinohara, S.; Namiki, T.; Nobori, T.; Kaneko, Y. Analysis of the p16INK4, p14ARF, p15, TP53, and MDM2 genes and their prognostic implications in osteosarcoma and Ewing sarcoma. *Cancer Genet. Cytogenet.* **2000**, *120*, 91–98. [[CrossRef](#)]
29. Gokgoz, N.; Wunder, J.S.; Mousses, S.; Eskandarian, S.; Bell, R.S.; Andrulis, I.L. Comparison of p53 mutations in patients with localized osteosarcoma and metastatic osteosarcoma. *Cancer* **2001**, *92*, 2181–2189. [[CrossRef](#)]
30. Vassilev, L.T.; Vu, B.T.; Graves, B.; Carvajal, D.; Podlaski, F.; Filipovic, Z.; Kong, N.; Kammlott, U.; Lukacs, C.; Klein, C.; et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **2004**, *303*, 844–848. [[CrossRef](#)]
31. Stratton, M.R.; Moss, S.; Warren, W.; Patterson, H.; Clark, J.; Fisher, C.; Fletcher, C.D.; Ball, A.; Thomas, M.; Gusterson, B.A.; et al. Mutation of the p53 gene in human soft tissue sarcomas: Association with abnormalities of the RB1 gene. *Oncogene* **1990**, *5*, 1297–1301. [[PubMed](#)]
32. Moll, U.M.; Petrenko, O. The MDM2-p53 interaction. *Mol. Cancer Res.* **2003**, *1*, 1001–1008. [[PubMed](#)]
33. Tamura, K.; Utsunomiya, J.; Iwama, T.; Furuyama, J.; Takagawa, T.; Takeda, N.; Fukuda, Y.; Matsumoto, T.; Nishigami, T.; Kusuhara, K.; et al. Mechanism of carcinogenesis in familial tumors. *Int. J. Clin. Oncol.* **2004**, *9*, 232–245. [[CrossRef](#)]
34. Sandberg, A.A.; Bridge, J.A. Updates on the cytogenetics and molecular genetics of bone and soft tissue tumors: Osteosarcoma and related tumors. *Cancer Genet. Cytogenet.* **2003**, *145*, 1–30. [[CrossRef](#)]
35. Poos, K.; Smida, J.; Maugg, D.; Eckstein, G.; Baumhoer, D.; Nathrath, M.; Korsching, E. Genomic heterogeneity of osteosarcoma—shift from single candidates to functional modules. *PLoS ONE* **2015**, *10*, e0123082. [[CrossRef](#)] [[PubMed](#)]
36. Sun, L.; Li, J.; Yan, B. Gene expression profiling analysis of osteosarcoma cell lines. *Mol. Med. Rep.* **2015**, *12*, 4266–4272. [[CrossRef](#)]
37. Shi, Z.; Zhou, H.; Pan, B.; Lu, L.; Wei, Z.; Shi, L.; Yao, X.; Kang, Y.; Feng, S. Exploring the key genes and pathways of osteosarcoma with pulmonary metastasis using a gene expression microarray. *Mol. Med. Rep.* **2017**, *16*, 7423–7431. [[CrossRef](#)]
38. Farhan, M.; Wang, H.; Gaur, U.; Little, P.J.; Xu, J.; Zheng, W. FOXO Signaling Pathways as Therapeutic Targets in Cancer. *Int. J. Biol. Sci.* **2017**, *13*, 815–827. [[CrossRef](#)]
39. Shaw, R.J.; Cantley, L.C. Ras, PI(3)K and mTOR signalling controls tumour cell growth. *Nature* **2006**, *441*, 424–430. [[CrossRef](#)]
40. Siqueira, M.F.; Flowers, S.; Bhattacharya, R.; Faibish, D.; Behl, Y.; Kotton, D.N.; Gerstenfeld, L.; Moran, E.; Graves, D.T. FOXO1 modulates osteoblast differentiation. *Bone* **2011**, *48*, 1043–1051. [[CrossRef](#)]
41. Kim, H.N.; Iyer, S.; Ring, R.; Almeida, M. The Role of FoxOs in Bone Health and Disease. *Curr. Top. Dev. Biol.* **2018**, *127*, 149–163. [[CrossRef](#)] [[PubMed](#)]
42. Tan, P.; Guan, H.; Xie, L.; Mi, B.; Fang, Z.; Li, J.; Li, F. FOXO1 inhibits osteoclastogenesis partially by antagonizing MYC. *Sci. Rep.* **2015**, *5*, 16835. [[CrossRef](#)]
43. Coomans de Brachene, A.; Demoulin, J.B. FOXO transcription factors in cancer development and therapy. *Cell. Mol. Life Sci.* **2016**, *73*, 1159–1172. [[CrossRef](#)] [[PubMed](#)]
44. Fu, Z.; Tindall, D.J. FOXOs, cancer and regulation of apoptosis. *Oncogene* **2008**, *27*, 2312–2319. [[CrossRef](#)]
45. Burgering, B.M.; Medema, R.H. Decisions on life and death: FOXO Forkhead transcription factors are in command when PKB/Akt is off duty. *J. Leukoc. Biol.* **2003**, *73*, 689–701. [[CrossRef](#)]
46. Moriishi, T.; Kawai, Y.; Komori, H.; Rokutanda, S.; Eguchi, Y.; Tsujimoto, Y.; Asahina, I.; Komori, T. Bcl2 deficiency activates FoxO through Akt inactivation and accelerates osteoblast differentiation. *PLoS ONE* **2014**, *9*, e86629. [[CrossRef](#)]
47. Guan, H.; Tan, P.; Xie, L.; Mi, B.; Fang, Z.; Li, J.; Yue, J.; Liao, H.; Li, F. FOXO1 inhibits osteosarcoma oncogenesis via Wnt/beta-catenin pathway suppression. *Oncogenesis* **2015**, *4*, e166. [[CrossRef](#)]
48. Tamborero, D.; Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, J.; Kandoth, C.; Reimand, J.; Lawrence, M.S.; Getz, G.; Bader, G.D.; Ding, L.; et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **2013**, *3*, 2650. [[CrossRef](#)]
49. Mendoza, M.C.; Er, E.E.; Blenis, J. The Ras-ERK and PI3K-mTOR pathways: Cross-talk and compensation. *Trends Biochem. Sci.* **2011**, *36*, 320–328. [[CrossRef](#)]

50. Xu, Y.; Li, N.; Xiang, R.; Sun, P. Emerging roles of the p38 MAPK and PI3K/AKT/mTOR pathways in oncogene-induced senescence. *Trends Biochem. Sci.* **2014**, *39*, 268–276. [[CrossRef](#)]
51. De Luca, A.; Maiello, M.R.; D'Alessio, A.; Pergameno, M.; Normanno, N. The RAS/RAF/MEK/ERK and the PI3K/AKT signalling pathways: Role in cancer pathogenesis and implications for therapeutic approaches. *Expert Opin. Ther. Targets* **2012**, *16* (Suppl. 2), S17–S27. [[CrossRef](#)] [[PubMed](#)]
52. Han, G.; Wang, Y.; Bi, W. C-Myc overexpression promotes osteosarcoma cell invasion via activation of MEK-ERK pathway. *Oncol. Res.* **2012**, *20*, 149–156. [[CrossRef](#)] [[PubMed](#)]
53. Samatar, A.A.; Poulidakos, P.I. Targeting RAS-ERK signalling in cancer: Promises and challenges. *Nat. Rev. Drug Discov.* **2014**, *13*, 928–942. [[CrossRef](#)]
54. Daw, N.C.; Chou, A.J.; Jaffe, N.; Rao, B.N.; Billups, C.A.; Rodriguez-Galindo, C.; Meyers, P.A.; Huh, W.W. Recurrent osteosarcoma with a single pulmonary metastasis: A multi-institutional review. *Br. J. Cancer* **2015**, *112*, 278–282. [[CrossRef](#)]
55. Lamora, A.; Mullard, M.; Amiaud, J.; Brion, R.; Heymann, D.; Redini, F.; Verrecchia, F. Anticancer activity of halofuginone in a preclinical model of osteosarcoma: Inhibition of tumor growth and lung metastases. *Oncotarget* **2015**, *6*, 14413–14427. [[CrossRef](#)]
56. Diepenbruck, M.; Christofori, G. Epithelial-mesenchymal transition (EMT) and metastasis: Yes, no, maybe? *Curr. Opin. Cell. Biol.* **2016**, *43*, 7–13. [[CrossRef](#)]
57. Jiang, W.G.; Sanders, A.J.; Katoh, M.; Ungefroren, H.; Gieseler, F.; Prince, M.; Thompson, S.K.; Zollo, M.; Spano, D.; Dhawan, P.; et al. Tissue invasion and metastasis: Molecular, biological and clinical perspectives. *Semin. Cancer Biol.* **2015**, *35*, S244–S275. [[CrossRef](#)]
58. Gross, A.C.; Cam, H.; Phelps, D.A.; Saraf, A.J.; Bid, H.K.; Cam, M.; London, C.A.; Winget, S.A.; Arnold, M.A.; Brandolini, L.; et al. IL-6 and CXCL8 mediate osteosarcoma-lung interactions critical to metastasis. *JCI Insight* **2018**, *3*. [[CrossRef](#)]
59. Liang, C.; Li, F.; Wang, L.; Zhang, Z.K.; Wang, C.; He, B.; Li, J.; Chen, Z.; Shaikh, A.B.; Liu, J.; et al. Tumor cell-targeted delivery of CRISPR/Cas9 by aptamer-functionalized lipopolymer for therapeutic genome editing of VEGFA in osteosarcoma. *Biomaterials* **2017**, *147*, 68–85. [[CrossRef](#)]
60. Weekes, D.; Kashima, T.G.; Zanduetta, C.; Perurena, N.; Thomas, D.P.; Sunter, A.; Vuillier, C.; Bozec, A.; El-Emir, E.; Miletich, I.; et al. Regulation of osteosarcoma cell lung metastasis by the c-Fos/AP-1 target FGFR1. *Oncogene* **2016**, *35*, 2948. [[CrossRef](#)]
61. Kaya, M.; Wada, T.; Akatsuka, T.; Kawaguchi, S.; Nagoya, S.; Shindoh, M.; Higashino, F.; Mezawa, F.; Okada, F.; Ishii, S. Vascular endothelial growth factor expression in untreated osteosarcoma is predictive of pulmonary metastasis and poor prognosis. *Clin. Cancer Res.* **2000**, *6*, 572–577. [[PubMed](#)]
62. Yu, Y.; Luk, F.; Yang, J.L.; Walsh, W.R. Ras/Raf/MEK/ERK pathway is associated with lung metastasis of osteosarcoma in an orthotopic mouse model. *Anticancer Res.* **2011**, *31*, 1147–1152. [[PubMed](#)]
63. Himelstein, B.P.; Asada, N.; Carlton, M.R.; Collins, M.H. Matrix metalloproteinase-9 (MMP-9) expression in childhood osseous osteosarcoma. *Med. Pediatr. Oncol.* **1998**, *31*, 471–474. [[CrossRef](#)]
64. Liu, Y.; Wang, Y.; Teng, Z.; Chen, J.; Li, Y.; Chen, Z.; Li, Z.; Zhang, Z. Matrix metalloproteinase 9 expression and survival of patients with osteosarcoma: A meta-analysis. *Eur. J. Cancer Care Engl.* **2017**, *26*, e12364. [[CrossRef](#)]
65. Guleria, A.; Chandna, S. ATM kinase: Much more than a DNA damage responsive protein. *DNA Repair* **2016**, *39*, 1–20. [[CrossRef](#)]
66. Shiloh, Y.; Ziv, Y. The ATM protein kinase: Regulating the cellular response to genotoxic stress, and more. *Nat. Rev. Mol. Cell Biol.* **2013**, *14*, 197–210. [[CrossRef](#)]
67. Awasthi, P.; Foiani, M.; Kumar, A. ATM and ATR signaling at a glance. *J. Cell Sci.* **2016**, *129*, 1285. [[CrossRef](#)]
68. Ray, A.; Blevins, C.; Wani, G.; Wani, A.A. ATR- and ATM-Mediated DNA Damage Response Is Dependent on Excision Repair Assembly during G1 but Not in S Phase of Cell Cycle. *PLoS ONE* **2016**, *11*, e0159344. [[CrossRef](#)]
69. Blackford, A.N.; Jackson, S.P. ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response. *Mol. Cell* **2017**, *66*, 801–817. [[CrossRef](#)]
70. Kovac, M.; Blattmann, C.; Ribic, S.; Smida, J.; Mueller, N.S.; Engert, F.; Castro-Giner, F.; Weischenfeldt, J.; Kovacova, M.; Krieg, A.; et al. Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. *Nat. Commun.* **2015**, *6*, 8940. [[CrossRef](#)]

71. Nabetani, A.; Ishikawa, F. Alternative lengthening of telomeres pathway: Recombination-mediated telomere maintenance mechanism in human cells. *J. Biochem.* **2011**, *149*, 5–14. [[CrossRef](#)] [[PubMed](#)]
72. di Masi, A.; Cilli, D.; Berardinelli, F.; Talarico, A.; Pallavicini, I.; Pennisi, R.; Leone, S.; Antoccia, A.; Noguera, N.I.; Lo-Coco, F.; et al. PML nuclear body disruption impairs DNA double-strand break sensing and repair in APL. *Cell Death Dis.* **2016**, *7*, e2308. [[CrossRef](#)] [[PubMed](#)]
73. Lallemand-Breitenbach, V.; de The, H. PML nuclear bodies: From architecture to function. *Curr. Opin. Cell Biol.* **2018**, *52*, 154–161. [[CrossRef](#)]
74. Pickett, H.A.; Reddel, R.R. Molecular mechanisms of activity and derepression of alternative lengthening of telomeres. *Nat. Struct. Mol. Biol.* **2015**, *22*, 875–880. [[CrossRef](#)]
75. Voisset, E.; Moravcsik, E.; Stratford, E.W.; Jaye, A.; Palgrave, C.J.; Hills, R.K.; Salomoni, P.; Kogan, S.C.; Solomon, E.; Grimwade, D. Pml nuclear body disruption cooperates in APL pathogenesis and impairs DNA damage repair pathways in mice. *Blood* **2018**, *131*, 636–648. [[CrossRef](#)] [[PubMed](#)]
76. Neumann, A.A.; Watson, C.M.; Noble, J.R.; Pickett, H.A.; Tam, P.P.; Reddel, R.R. Alternative lengthening of telomeres in normal mammalian somatic cells. *Genes Dev.* **2013**, *27*, 18–23. [[CrossRef](#)] [[PubMed](#)]
77. Dilley, R.L.; Verma, P.; Cho, N.W.; Winters, H.D.; Wondisford, A.R.; Greenberg, R.A. Break-induced telomere synthesis underlies alternative telomere maintenance. *Nature* **2016**, *539*, 54–58. [[CrossRef](#)]
78. Kim, J.Y.; Brosnan-Cashman, J.A.; An, S.; Kim, S.J.; Song, K.B.; Kim, M.S.; Kim, M.J.; Hwang, D.W.; Meeker, A.K.; Yu, E.; et al. Alternative Lengthening of Telomeres in Primary Pancreatic Neuroendocrine Tumors Is Associated with Aggressive Clinical Behavior and Poor Survival. *Clin. Cancer Res.* **2017**, *23*, 1598–1606. [[CrossRef](#)]
79. Singhi, A.D.; Liu, T.C.; Roncaioli, J.L.; Cao, D.; Zeh, H.J.; Zureikat, A.H.; Tsung, A.; Marsh, J.W.; Lee, K.K.; Hogg, M.E.; et al. Alternative Lengthening of Telomeres and Loss of DAXX/ATRX Expression Predicts Metastatic Disease and Poor Survival in Patients with Pancreatic Neuroendocrine Tumors. *Clin. Cancer Res.* **2017**, *23*, 600–609. [[CrossRef](#)]
80. Liao, J.Y.; Tsai, J.H.; Jeng, Y.M.; Lee, J.C.; Hsu, H.H.; Yang, C.Y. Leiomyosarcoma with alternative lengthening of telomeres is associated with aggressive histologic features, loss of ATRX expression, and poor clinical outcome. *Am. J. Surg. Pathol.* **2015**, *39*, 236–244. [[CrossRef](#)]
81. Fogli, A.; Demattei, M.V.; Corset, L.; Vaur-Barriere, C.; Chautard, E.; Biau, J.; Kemeny, J.L.; Godfraind, C.; Pereira, B.; Khalil, T.; et al. Detection of the alternative lengthening of telomeres pathway in malignant gliomas for improved molecular diagnosis. *J. Neurooncol.* **2017**, *135*, 381–390. [[CrossRef](#)] [[PubMed](#)]
82. Chong, J.L.; Wenzel, P.L.; Saenz-Robles, M.T.; Nair, V.; Ferrey, A.; Hagan, J.P.; Gomez, Y.M.; Sharma, N.; Chen, H.Z.; Ouseph, M.; et al. E2f1-3 switch from activators in progenitor cells to repressors in differentiating cells. *Nature* **2009**, *462*, 930–934. [[CrossRef](#)] [[PubMed](#)]
83. Hayden, M.S.; Ghosh, S. Signaling to NF-kappaB. *Genes Dev.* **2004**, *18*, 2195–2224. [[CrossRef](#)] [[PubMed](#)]
84. Seoane, J.; Gomis, R.R. TGF-beta Family Signaling in Tumor Suppression and Cancer Progression. *Cold Spring Harb. Perspect. Biol.* **2017**, *9*, a022277. [[CrossRef](#)]
85. Pohl, T.; Gugasyan, R.; Grumont, R.J.; Strasser, A.; Metcalf, D.; Tarlinton, D.; Sha, W.; Baltimore, D.; Gerondakis, S. The combined absence of NF-kappa B1 and c-Rel reveals that overlapping roles for these transcription factors in the B cell lineage are restricted to the activation and function of mature cells. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 4514–4519. [[CrossRef](#)]
86. Liekens, A.M.; De Knijf, J.; Daelemans, W.; Goethals, B.; De Rijk, P.; Del-Favero, J. BioGraph: Unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome. Biol.* **2011**, *12*, R57. [[CrossRef](#)]
87. Wu, X.; Jiang, R.; Zhang, M.Q.; Li, S. Network-based global inference of human disease genes. *Mol. Syst. Biol.* **2008**, *4*, 189. [[CrossRef](#)]
88. Pinero, J.; Queralt-Rosinach, N.; Bravo, A.; Deu-Pons, J.; Bauer-Mehren, A.; Baron, M.; Sanz, F.; Furlong, L.I. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**, *2015*, bav028. [[CrossRef](#)]
89. Fontaine, J.F.; Priller, F.; Barbosa-Silva, A.; Andrade-Navarro, M.A. Genie: Literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.* **2011**, *39*, W455–W461. [[CrossRef](#)]
90. Jourquin, J.; Duncan, D.; Shi, Z.; Zhang, B. GLAD4U: Deriving and prioritizing gene lists from PubMed literature. *BMC Genom.* **2012**, *13* (Suppl. 8), S20. [[CrossRef](#)]

91. Guney, E.; Garcia-Garcia, J.; Oliva, B. GUILDIfy: A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. *Bioinformatics* **2014**, *30*, 1789–1790. [[CrossRef](#)] [[PubMed](#)]
92. Yang, H.; Robinson, P.N.; Wang, K. Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **2015**, *12*, 841–843. [[CrossRef](#)] [[PubMed](#)]
93. Cheng, D.; Knox, C.; Young, N.; Stothard, P.; Damaraju, S.; Wishart, D.S. PolySearch: A web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.* **2008**, *36*, W399–W405. [[CrossRef](#)] [[PubMed](#)]
94. Yue, P.; Melamud, E.; Moulton, J. SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinform.* **2006**, *7*, 166. [[CrossRef](#)] [[PubMed](#)]
95. Lambert, S.A.; Jolma, A.; Campitelli, L.F.; Das, P.K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T.R.; Weirauch, M.T. The Human Transcription Factors. *Cell* **2018**, *175*, 598–599. [[CrossRef](#)]
96. Yevshin, I.; Sharipov, R.; Kolmykov, S.; Kondrakhin, Y.; Kolpakov, F. GTRD: A database on gene transcription regulation-2019 update. *Nucleic Acids Res.* **2019**, *47*, D100–D105. [[CrossRef](#)]
97. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Perturbation-Theory Machine Learning (PTML) Multilabel Model of the ChEMBL Dataset of Preclinical Assays for Antisarcoma Compounds

Alejandro Cabrera-Andrade,^{*,&} Andrés López-Cortés,[&] Cristian R. Munteanu, Alejandro Pazos, Yunierkis Pérez-Castillo, Eduardo Tejera, Sonia Arrasate, and Humbert González-Díaz^{*}



Cite This: *ACS Omega* 2020, 5, 27211–27220



Read Online

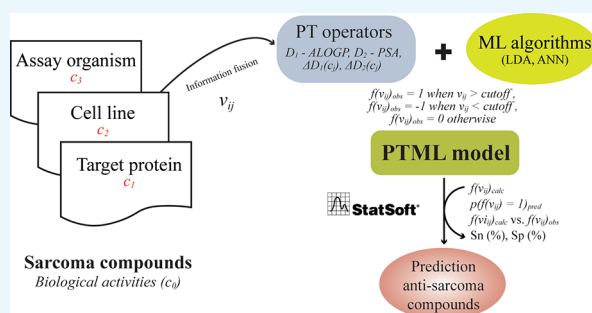
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Sarcomas are a group of malignant neoplasms of connective tissue with a different etiology than carcinomas. The efforts to discover new drugs with antisarcoma activity have generated large datasets of multiple preclinical assays with different experimental conditions. For instance, the ChEMBL database contains outcomes of 37,919 different antisarcoma assays with 34,955 different chemical compounds. Furthermore, the experimental conditions reported in this dataset include 157 types of biological activity parameters, 36 drug targets, 43 cell lines, and 17 assay organisms. Considering this information, we propose combining perturbation theory (PT) principles with machine learning (ML) to develop a PTML model to predict antisarcoma compounds. PTML models use one function of reference that measures the probability of a drug being active under certain conditions (protein, cell line, organism, etc.). In this paper, we used a linear discriminant analysis and neural network to train and compare PT and non-PT models. All the explored models have an accuracy of 89.19–95.25% for training and 89.22–95.46% in validation sets. PTML-based strategies have similar accuracy but generate simplest models. Therefore, they may become a versatile tool for predicting antisarcoma compounds.



INTRODUCTION

Sarcomas are a group of malignant neoplasms of connective tissue. Although their prevalence is much lower than carcinomas, the number of cases is increasing according to the World Health Organization.¹ At the molecular level, their behavior differs from carcinomas, presenting a more varied and complex etiology. This high etiological complexity possibly stems from their mesenchymal origin, which makes it difficult to propose new therapeutic targets for the respective treatment.^{2–6} Representative anticancer compounds tend to have high cytotoxicity and low cellular specificity.⁷ This leads to a decreased efficiency within the treatment and a low remission rate of the disease. However, a description of new molecular markers and the constant performance of drug preclinical assays have generated large amounts of data.^{8–12} This data, if adequately rationalized, may lead in turn to the design of more selective drugs, which takes into account specific drivers based on pathogenic signaling pathways. For instance, the Chemical Database of the European Molecular Biology Laboratory (ChEMBL)^{13,14} contains experimental outcomes for >37,900 different preclinical assays of anti-sarcoma drug candidates. These assays cover a large and structurally heterogeneous series of >34,900 different chemical compounds. Furthermore, the preclinical assays have been carried out on very different experimental conditions. These

experimental conditions include up to 155 different types of biological activity parameters, 36 protein targets, 43 cell lines, and 17 assay organisms. Overall, this forms a large and complex dataset susceptible to analysis so as to extract useful knowledge for drug discovery.

In this context, we can use computational techniques to explore this experimental dataset due to the evident difficulties to analyze it manually. Specifically, cheminformatics methodologies have succeeded in the discovery of new drug candidates effective in the wet-lab.^{15,16} However, many models developed thus far are applied only to carcinomas and/or are focused on homologous series of compounds with one target or a single cell line.^{17–26} In recent years, several studies have focused on applying these methodologies to the study of new types of antisarcoma drugs, mainly on cell lines.^{27–30} However, almost all the models reported have a narrow domain of application because they focus on only one set of conditions, for instance,

Received: July 13, 2020

Accepted: October 6, 2020

Published: October 15, 2020



Table 1. PTML Model Results

series	statistical parameter ^a	predicted statistics (%)	observed set	predicted set	
				$f(v_{ij})_{\text{pred}} = 0$	$f(v_{ij})_{\text{pred}} = 1$
training	Sp	95.63	$f(v_{ij})_{\text{obs}} = 0$	25,647	1172
	Sn	79.64	$f(v_{ij})_{\text{obs}} = 1$	330	1291
	Ac	94.72	total	25,977	2463
validation	Sp	95.79	$f(v_{ij})_{\text{obs}} = 0$	8559	376
	Sn	81.62	$f(v_{ij})_{\text{obs}} = 1$	100	444
	Ac	94.98	total	8659	820

^aSn, sensitivity (%); Sp, specificity (%); Ac, accuracy (%).

Table 2. Variables Used to Fit the PTML Model

condition ^a (c_j)	condition name	symbol	operator formula	operator information
c_0	activity type	$f(v_{ij})_{\text{obs}}$	$=\text{IF}(\text{AND}(v_{ij} > \text{cutoff}(c_0), d(c_0) = 1), 1, \text{IF}(\text{AND}(v_{ij} < \text{cutoff}(c_0), d(c_0) = -1), 1, 0))$	observed classification of the outcome v_{ij} in the assay with conditions c_j
c_0	activity type	$f(v_{ij})_{\text{ref}}$	$n(f(v_{ij})_{\text{obs}} = 1)/n_j$	function of reference if the observed value of probability $p(f(v_{ij}) = 1)_{\text{expt}}$ for the activity v_{ij} of type c_0
$c_j = [c_1, c_2, c_3]$	all conditions (c_j)	$\Delta D_1(c_j)$	$\text{ALOGP}_1 - \langle \text{ALOGP}(c_j) \rangle$	deviation of the molecular descriptors of hydrophobicity/lipophilicity D_1 (ALOGP) and polar surface area D_2 (PSA) from each expected value ($\langle D_1(c_j) \rangle$) or ($\langle D_2(c_j) \rangle$) for the conditions c_j (c_1 = protein target; c_2 = cell line; c_3 = assay organism)
$c_j = [c_1, c_2, c_3]$	all conditions (c_j)	$\Delta D_2(c_j)$	$\text{PSA}_1 - \langle \text{PSA}(c_j) \rangle$	

^aMMA operators with a subset of multiple conditions included in eq 1.

one specific property, target protein, or cell line. Thus, models where multiple conditions of assays are considered at the same time are attractive. Perturbation theory (PT) ideas with machine learning (ML) methods (PT + ML = PTML models) are particularly useful for fitting complex datasets with big data features in drug discovery, proteomics, nanotechnology, *etc.*^{31–41}

PTML models begin with one function of reference that measures the probability of a drug to be active under certain conditions (protein, cell line, organism, *etc.*). Next, PTML models use PT operators (PTOs) to account for the perturbations (deviations) of the input variables of this drug with respect to a population of drugs assayed under the same conditions. ML algorithms are used to establish the relationship between the inputs and the output variable. In cancer research, Speck-Planche *et al.* and other researchers have developed PTML-like models for different types of cancers (with an emphasis on carcinomas) such as bladder, prostate, brain, and breast cancers.^{42–50} In addition, Bediaga *et al.* developed a PTML algorithm for predicting anticancer compounds using data for multiple types of carcinomas at the same time.⁵¹ Speck-Planche *et al.* also recently developed the first PTML-like model for the prediction of antisarcoma compounds using a spectral moment approach.⁵²

In any case, there are no reports of other PTML-like models for antisarcoma compounds. In this study, we carried out a comprehensive compilation, curation, and preprocessing of the ChEMBL dataset for preclinical assays of antisarcoma compounds. After that, we developed the first PTML model able to fit this complex dataset with >37,900 assays and >34,900 compounds. To the best of our knowledge, the study outperforms all previous efforts in terms of simplicity of the model and number of cases, compounds, and cell lines considered.

RESULTS AND DISCUSSION

PTML Antisarcoma Compound Model. The statistical parameters for the PTML model showed a high specificity (Sp) and sensitivity (Sn) for the training series (95.63 and 79.64, respectively). In addition, similar values were obtained for Sp (95.79) and Sn (81.62) in the validation sets. Furthermore, the p-level obtained from the chi-square ($\chi^2 = 16848.08$) was <0.05, indicating that the model is able to perform a statistically significant separation of both classes. It is also interesting to observe the high overall accuracy (Ac) obtained in both sets: over 94% (Table 1). These results suggest that the generated model performs a statistically significant classification of antisarcoma compounds; hence, it can be considered useful for classification models with application in medicinal chemistry. The full list of biological activities (c_0) in the ChEMBL dataset of antisarcoma preclinical experimental assays is shown in Table S1.

The resulting PTML–linear discriminant analysis (LDA) model showed the following formula

$$f(v_{ij})_{\text{calc}} = -11.8545 + 34.8028 \cdot f(v_{ij})_{\text{ref}} + 0.37 \cdot D_1 - 0.0128 \cdot D_2 - 0.3616 \cdot [D_1 - \langle D_1(c_j) \rangle] + 0.0191 \cdot [D_2 - \langle D_2(c_j) \rangle]$$

$$n = 34955, \chi^2 = 16848.08, p < 0.001 \quad (1)$$

The PTML-LDA model was initiated by using as an input the values the function of reference $f(v_{ij})_{\text{ref}}$ for each compound and by adding the effect of perturbations within the system. These perturbation effects refer to the PTOs $\Delta D_k(c_j)$. In eq 1, “i” and “j” are the assay and condition, respectively. Additional coefficients and terms are described in Table 2.

The parameters ALOGP and PSA are widely used in medicinal chemistry because they are related to the lipophilicity of drugs and, consequently, to their capacity to pass through biological membranes or interact with protein

Table 3. Comparison to Other PTML Models of Anticancer Compounds

cancer type ^a	PT ^b	ML ^c	NV ^d	cases ^e	Sn(%) ^f	Sp(%) ^f	ref
sarcoma							
MSS	MMA	LDA	3	37,919	~80	>90	this work
MSS	MA	LDA	>10	3017	>90	>90	52
carcinoma							
bladder	MA	LDA	>10	664	>90	>90	44
bladder		ANN (RBF)	10	664	>95	>95	44
brain	MA	LDA	>10	1236	~90	>90	45
breast	MA	LDA	>10	2272	>85	>90	47
colorectal	MA	LDA	>10	1651	>90	>90	46
colorectal	MA	ANN (RBF)	>10	1651	>90	>90	46
prostate	MA	LDA	>10	1668	>85	>90	49
MCS	MMA	LDA	>10	116,934	>70	~90	51
MCS	MMA	LDA	3	116,934	>70	>90	51
MCS	MMA	ANN	4	116,934	>80	>80	51

^aMSS, multiple sarcoma subtypes; MCS, multiple carcinoma subtypes. ^bPT operators used in PTML models: MMA, multicondition moving average; MA, moving average. ^cML method used for the PTML models: LDA, linear discriminant analysis; ANN, artificial neural networks; RBF, radial basis function; LNN, linear neural networks; E-ANN (RBF), ensemble of artificial neural networks based on the RBF architecture. ^dNV, number of input variables. ^eNumber of preclinical assays. ^fApproximate values for training series.

Table 4. Different Scores Calculated for the Selected Biological Activities (c_0)

activity parameter for $v_{ij}(c_0)$ (unit)	$n_j(c_0)$ ^a	$\langle v_{ij}(c_0) \rangle$ ^b	$d_j(c_0)$ ^c	cutoff (c_0)	$n(f(v_{ij})_{obs} = 1)$ ^d	$p(f(v_{ij})_{obs} = 1/c_0)$ ^e
potency (nM)	31,581	19669.199	-1	100	149	0.005
IC ₅₀ (nM)	1808	228362.82	-1	100	177	0.098
inhibition (%)	690	39.186507	1	50	225	0.326
CC ₅₀ (nM)	450	134445.04	-1	100	4	0.009
activity (%)	404	52.416163	1	50	208	0.515
EC ₅₀ (nM)	379	63578.521	-1	100	44	0.116
TGI (%)	202	43.915842	1	50	102	0.505
T/C	173	26.556832	1	50	28	0.162
IC ₅₀ ($\mu\text{g mL}^{-1}$)	167	64.429402	-1	60	118	0.707
T/C (%)	144	156.92153	1	50	123	0.854
GI ₅₀ (nM)	113	66515.131	-1	100	13	0.115
EC ₅₀ ($\mu\text{g mL}^{-1}$)	90	60.733562	-1	60	57	0.633

^a $n_j(c_0)$, total compounds with experimental values. ^b $\langle v_{ij}(c_0) \rangle$, average calculated of each c_0 biological activity. ^c $d_j(c_0)$, desirability value (1, -1) assigned to each c_0 . ^d $n(f(v_{ij})_{obs} = 1)$, total number of biologically active compounds observed within each c_0 according to the experimental values $v_{ij}(c_0)$ reported for the parameters j . ^e $p(f(v_{ij})_{obs} = 1/c_0)$, probability of a desired biological activity within the conditions c_0 .

hydrophobic pockets.^{53–56} The PTML algorithm has been previously applied to the study of multiple preclinical assays of anticancer drugs. As shown in Table 3, most applications have been directed toward the most prevalent carcinomas among the global population. For instance, Speck-Planche *et al.* reported PTML-like models for bladder,⁴⁴ colorectal,⁴⁶ breast,⁴⁷ prostate⁴⁹ cancers and for multiple carcinoma subtypes.⁵¹ In addition, PTML-like models have been tested in antibrain tumor agents.⁴⁵ Interestingly, Bediaga *et al.* demonstrated the application of a PTML on several types of carcinomas simultaneously and obtained similar Sn and Sp values as we did (>90%).⁵¹ All these PTML-like models are able to account for changes in target proteins, cellular lines, organisms, *etc.* However, they are specific models for carcinomas, not for sarcomas.

It is worth noting that to the best of our knowledge, Speck-Planche *et al.*⁵² seem to be the only researchers to have reported a previous PTML-like model for sarcomas thus far. In their study, the prediction model in external validation resulted in Ac (90.78) and Sp (90.65) values that were lower than what was obtained in our model (Ac = 94.98 and Sp = 95.79). However, our PTML algorithm showed a lower sensitivity in external validation data (81.62%) than the model obtained by

Speck-Planche *et al.* (91.74%). Even when our model had a much lower number of variables and used a stricter cut-off definition for activity class (i.e., IC₅₀ = 0.1 μM instead 1 μM), these aspects alone cannot explain the sensitivity reduction.

The generated PTML-LDA model (eq 1) has important characteristics that allow it to be used within research focused on drug discovery. One of the main advantages of our model is the considerable reduction of input variables for the construction of the algorithm through the inclusion of PTOs. This reduction allowed us to work on datasets with a large amount of information, to define cut-off values, and to calculate the probability of belonging to a class, whether this was a prediction for active compounds (1) or inactive compounds (0). In this way, the Sn or Sp values of the model can be adjusted according to the delimited cut-offs. An ideal prediction model has a reasonable trade-off between Sn and Sp. This means that a high sensitivity is achieved by accepting a relatively low Sp and, conversely, a high Sp is reached by compromising Sn. Sp is synonymous with a true-negative rate, which is related to the false-positive rate,³⁰ so a high specificity in a prediction model for drug discovery implies that it is unlikely to get a positive result in a drug that does not have a desired biological activity. Thus, a positive

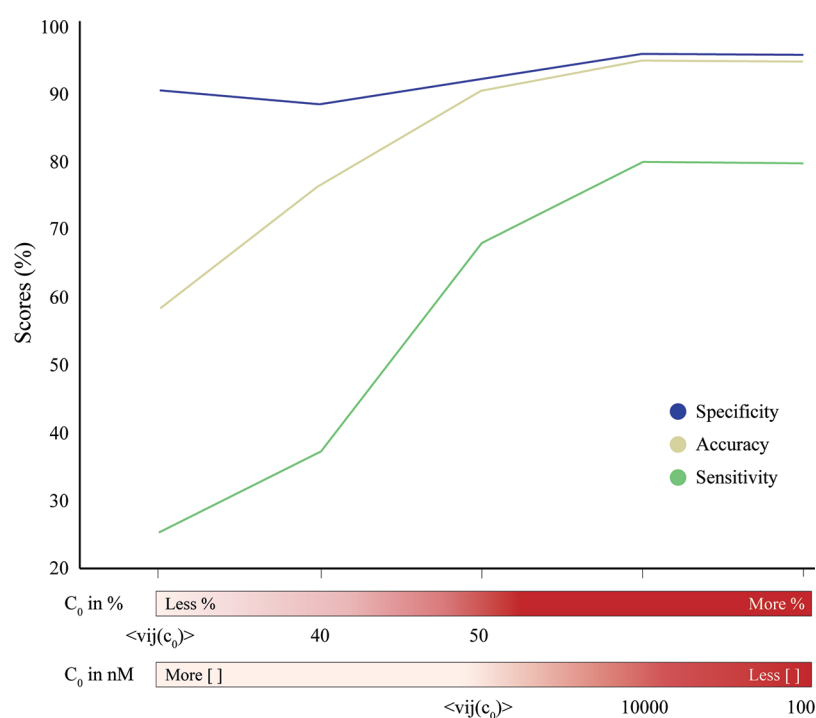


Figure 1. Variation of the specificity, sensitivity, and accuracy values according to the cut-offs implemented. The variation of these scores based on the biological activities c_0 is included in the x -axis. Biological activities c_0 expressed in % (e.g., inhibition, activity, tumor growth inhibition, etc.) and those expressed in nM (e.g., potency, IC_{50} , CI_{50} , etc.) are described. The final model is obtained by applying cut-off values of 50 for c_0 expressed in % and 100 for c_0 expressed in nM.

outcome in a specific model is quite informative in a drug discovery scenario.

On the other hand, a main attribute is the possible combination of several experimental conditions for the prediction of new compounds. In this sense, Speck-Planche *et al.*⁵² used around 3000 interactions derived from 14 cell lines and only considered IC_{50} assays for their model. However, we modeled 37,919 interactions cases comprising 36 protein targets, 43 cell lines, and 17 assay organisms. We also included several different assay types (Table 4). The modeling task we have is more complex not only because of the increment in the chemical diversity but also the wide type of heterogeneity in the interactions (i.e., target types and organisms). The two models cannot be compared in this scenario and our reduction in the ability to detect the true-positive cases (S_n) could be a consequence of this data complexity and also the modeling strategy.

PTML Cut-Off Scanning Study. As mentioned above, the cut-off implemented in the model is a rigorous value that, at the experimental level, is important if one desires to increase effectiveness in the process of discovering antiscarcoma drugs. A restricted value promotes high certainty in the prediction of active compounds for achieving a desired biological action under multiple test conditions.^{57–59} Furthermore, a strict cut-off can decrease the rate of predicted false positives; therefore, if the assay is to be implemented, then it needs a higher sensitivity or higher specificity. This value can be modeled depending on the experimental conditions one wishes to apply. This cut-off value also influences the accuracy within our model. As observed in Figure 1, when using the average $\langle v_{ij}(c_0) \rangle$ calculated for each c_0 , the Ac is not a desirable score. These low statistical values are mainly influenced by the low S_n in the prediction. By increasing the rigor, the model improves

its prediction values for the active compounds (1). When looking at these results, our prediction algorithm not only takes into account several experimental conditions but also restricts the prediction of compounds to those that have true biological activity.

PTML vs ML Model Comparison. Most multitasking or multilabel ML methods are useful for predicting multiple categorical outputs for the same set of input continuous variables.^{60,61} However, our problem was a little different: we had to develop an ML model with only two possible outputs, $f(v_{ij})_{pred} = 1$ or 0, for the same set of input variables. That meant that our model was not multitasking for a single case with a set of input variables containing multiple continuous variables plus multiple categorical input variables. However, we had multiple combinations of input categorical variables or levels for the same set of input continuous variables. Hence, our model was multilabel in the input categorical variables for the same set of input continuous variables. To illustrate this fact, we developed here a comparison of our PTML-LDA model vs classic ML using multiple labeling categorical variables. As seen in Figure 2A, the performance of our PTML-LDA model compared to a classic ML-LDA demonstrates similar values based on S_p , S_n , and Ac. Similarly, when developing neural networks (NN), the results of PTML-NN (Figure 2B) and ML-NN (Figure 2C) are quite similar. One of the advantages of our PTML model is the inclusion of PTOs, which greatly reduces the number of variables to generate the algorithm. Thus, although the statistics of all the models generated are quite similar, the PTML methodology allows for the reduction of variables from 164 variables in classic ML methods to only 5 in the PTML model. All the PTML and non-PTML model results are described in Table S2.

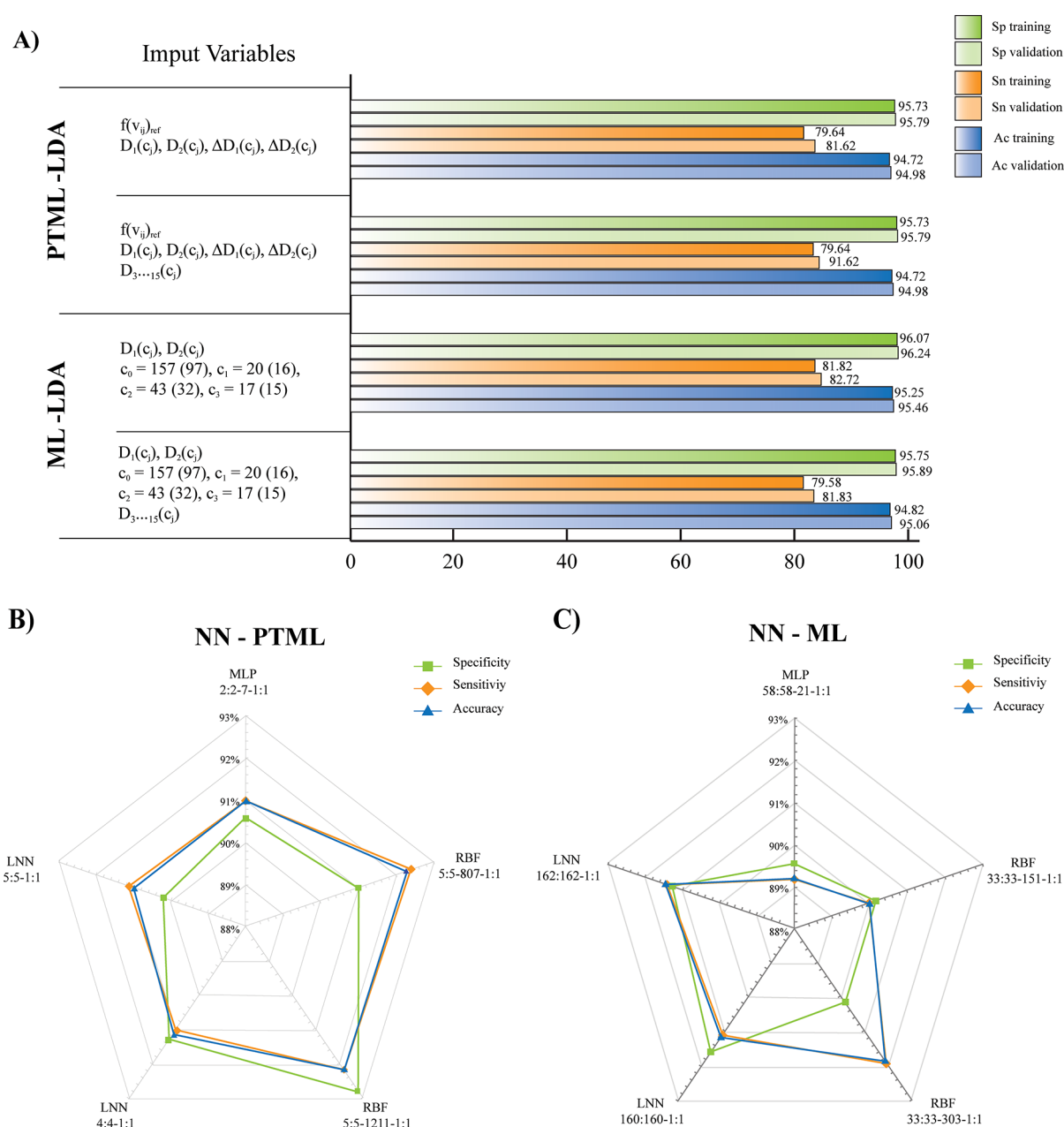


Figure 2. PTML vs ML models. Comparison of sensitivity, specificity, and accuracy of all the generated models. (A) Prediction values of PTML-LDA and ML-LDA models using different types of input variables: $f(v_{ij})_{pred}$ is the function of reference; $D_1(c_i)$ and $D_2(c_j)$ are the ALOGP and PSA descriptors, respectively; $\Delta D_1(c_i)$ and $\Delta D_2(c_j)$ are the deviations of the molecular descriptors of ALOGP and PSA, respectively; $D_3, \dots, D_{15}(c_i)$ are the 12 BCUT molecular descriptors calculated from ChemAxon. Unlike the PTML model, the ML model is calculated with conditions $c_1, c_2,$ and c_3 as a separated set of categorical variables. (B) Prediction values between the neural network-PTML (NN-PTML) and (C) NN-ML models. The NN obtained were multilayer perceptron (MLP), linear neural network (LNN), and radial basis function network (RBF).

PTML vs ML Model with Other Descriptors. Previous studies have considered a wide variety and quantity of molecular descriptors in PTML models. For example, for sarcoma modeling, Speck-Planche *et al.*⁵² used 423 descriptors followed by a feature selection strategy. Similarly, 289 descriptors were used in a PTML model on breast cancer.⁴⁷ We used this approach as a strategy to compare the performance of PTML model vs classic ML techniques including new molecular descriptors (Figure 2A). In this ML study, we included 12 BCUT molecular descriptors (D_k , with $k > 2$) as an input, which were not used in the previous model, and 162 categorical (dummy) variables (C_k). These C_k have

been used to label the multiple conditions of the assays c_j (organisms, proteins, cell lines, *etc.*). One must remember that $D_1 = \text{ALOGP}$ and $D_2 = \text{PSA}$. The new molecular descriptors were D_3, D_4, \dots, D_{14} . The expansion of the variables together with the ML strategies yielded good results but did not outperform what was obtained for the PTML-LDA anti-sarcoma model (as seen in Figure 2A and Table S2) and the number of variables increased to 174 input variables in total. This suggests that by adding different molecular descriptors and probably feature selection strategies, acceptable models for drug discovery can be built. However, our PTML-LDA model based on D_1 and D_2 is a simpler yet effective model.

Table 5. Multiple-Condition Averages for All Sarcoma Assays

	assay condition (c_j) ^a			parameter		
	c_1 = protein (<i>gene</i>)	c_2 = cell line	c_3 = assay organisms ^b	$n_j(c_j)$	$\langle D_1(c_j) \rangle$	$\langle D_2(c_j) \rangle$
O75874 (<i>IDH1</i>)	MD	MD	<i>H. sapiens</i>	31,581	3.778	70.597
MD	MD	MD	<i>M. musculus</i>	1440	2.67	103.712
MD	MD	U2OS	<i>H. sapiens</i>	746	4.421	78.325
MD	MD	HOS	<i>H. sapiens</i>	637	3.603	89.517
MD	MD	MD	<i>H. sapiens</i>	375	3.846	69.876
MD	MD	SAOS-2	<i>H. sapiens</i>	358	4.882	81.659
MD	MD	Sarcoma-180	<i>M. musculus</i>	271	1.108	83.68
MD	MD	MG-63	<i>H. sapiens</i>	241	2.965	111.864
MD	MD	M5076	<i>M. musculus</i>	197	3.033	114.886
MD	MD	HT-1080	<i>H. sapiens</i>	170	2.826	97.731
MD	MD	143B	<i>H. sapiens</i>	131	1.283	141.735
MD	MD	MD	<i>Pseudomonas aeruginosa</i>	130	0.277	142.432
MD	MD	MD	MD	126	1.898	93.448
MD	MD	rhabdomyosarcoma cell	<i>H. sapiens</i>	116	4.036	77.177
MD	MD	CCRF S ⁻¹⁸⁰	<i>M. musculus</i>	109	0.978	140.984
P13053 (<i>Vdr</i>)	MD	MD	<i>Rattus norvegicus</i>	64	5.844	60.476
MD	MD	MES-SA	<i>H. sapiens</i>	64	2.956	89.631
MD	MD	MD	RSV	61	1.277	127.944
MD	MD	6C3HED	<i>M. musculus</i>	60	3.09	97.831
MD	MD	C3H/3T3	MMSV	50	0.327	139.359
P35354 (<i>PTGS2</i>)	MD	MD	<i>H. sapiens</i>	49	3.515	69.152
MD	MD	A204	<i>H. sapiens</i>	44	1.189	106.655
P03359 (<i>pol</i>)	MD	MD	WMSV	44	6.786	204.629
MD	MD	MD	<i>Gallus gallus</i>	43	0.516	106.529
P37231 (<i>PPARG</i>)	MD	MD	<i>H. sapiens</i>	40	5.33	83.835
MD	MD	MD	MMSV	39	0.213	166.782
Q07869 (<i>PPARA</i>)	MD	MD	<i>H. sapiens</i>	37	5.364	81.891
Q13443 (<i>ADAM9</i>)	MD	MD	<i>H. sapiens</i>	35	2.914	91.186
MD	MD	MD	<i>R. norvegicus</i>	34	5.245	64.58
MD	MD	fibroblast	MMSV	33	-1.224	150.956
MD	MD	MD	enterovirus	33	6.348	38.332
MD	MD	MD	human herpesvirus 1	31	6.27	57.306
MD	MD	791T cell line	<i>H. sapiens</i>	28	-1.179	139.194
MD	MD	C3H/3T3	<i>M. musculus</i>	28	1.745	115.047
P08253 (<i>MMP2</i>)	MD	MD	<i>H. sapiens</i>	28	3.31	112.85
MD	MD	MD	human enterovirus 71	28	1.967	124.221
P04637 (<i>TP53</i>), Q00987 (<i>MDM2</i>)	MD	SJSA-1	<i>H. sapiens</i>	27	5.213	49.453
P06401 (<i>PGR</i>)	MD	MD	<i>H. sapiens</i>	26	4.494	32.958
MD	MD	HL-60	<i>H. sapiens</i>	25	3.81	33.754

^aMD, missing data. ^bRSV, Rous sarcoma virus; MLV, murine leukemia virus; MMSV, Moloney murine sarcoma virus; WMSV, Woolly monkey sarcoma virus.

Multiple-Condition Averages in the PTML Antisarcoma Model. In total, we found 83 possible combinations of multiple conditions for all the included sarcoma assays. As shown in Table 5, the $n_j(c_j)$ with the highest number of entries corresponded to tests on human cell lines and on cell lines in *Mus musculus*. The multicondition moving averages (MMAs) used here, $\langle D_1(c_j) \rangle$ and $\langle D_2(c_j) \rangle$, vary significantly along all combinations. However, the anticancer compounds observed for the human osteosarcoma cell lines U2OS, HOS, SAOS-2, MG-63, and 143B and for the fibrosarcoma cell line HT-1080 were in a range of $\langle D_1(c_j) \rangle$ of 1.2–3.7. A similar range was observed in compounds tested in *M. musculus* ($\langle D_1(c_j) \rangle = 1–3$). Interestingly, when comparing these values with the variation of $\langle D_2(c_j) \rangle$, tests on virus lines, such as Moloney murine sarcoma virus and Woolly monkey sarcoma virus, had higher means (between 140 and 205). Since the ALOGP coefficient is a measure widely used in drug discovery to assess

the degree of absorption, distribution in the body, penetration across biological membranes, metabolism, and excretion, this range identified in our results is an important space for the prediction of antisarcoma drugs.^{62,63} Likewise, the range of PSA evidenced in viral line assays may be a better space for this coefficient if it is desired to predict new compounds in these experimental conditions. This may be interesting when defining the validation of a certain antisarcoma compound. Thus, if a compound is significantly predicted in an experimental animal or human cell lines, then it will be possible to propose validations at the preclinical level or in clinical trials, respectively.

How to Use the PTML Model in Practice. The model is capable of scoring the activity of a single compound under different assay conditions. To predict a new compound, first, we have to substitute the expected values of function of reference $f(v_{ij})_{\text{ref}} = p(f(v_{ij}) = 1)_{\text{expt}}$ in the model. As

mentioned, this is the probability of the compound being active for a given biological activity parameter (c_0) (see Table 2). Next, we need to substitute into the equation the values of molecular descriptors $D_1 = \text{ALOGP}$ and $D_2 = \text{PSA}$ of the compound (chemical structure), calculated with the same algorithm used in the ChEMBL dataset. Last, we have to substitute into the equation the average values (expected values) of the molecular descriptors $\langle D_k(c_j) \rangle$ for the specific subset of conditions of the assay c_j we want to predict. In Table S, we show some selected values of these averages with >25 assays reported. It can be noted that the most populated assays in *Homo sapiens* in the dataset were those *in vitro* assays that targeted the protein O75874 (*IDH1*) and that targeted the cell line U2OS. Upon inspecting Table S, we can see that $\langle D_k(c_j) \rangle$ values change for different subsets of conditions c_j . Consequently, when we substitute the different $\langle D_k(c_j) \rangle$ values into the model for the same compound, we can calculate different scores $f(v_{ij})_{\text{calc}}$ of biological activity of the same compound under multiple assay conditions. The full list of the values of $\langle D_k(c_j) \rangle$ appears in Table S3.

CONCLUSIONS

In this research work, we generated a PTML-LDA model constructed with antisarcoma assays obtained from ChEMBL and a heterogeneous set of different cell lines, organisms, and targets. As far as we know, this constitutes the first time that this kind of model was tested for sarcoma comprising 34,955 chemical compounds and 37,919 assays. The PTML-LDA model was compared with classic ML approaches like the neural network and also with non-PT consideration. The rate of true positives and true negatives is similar when comparing PTML-LDA to other prediction models. PTML-LDA reduces the amount of input variables (ALOGP and PSA) needed, thus increasing the simplicity and interpretability of the model.

METHODS

ChEMBL Data Curation and Preprocessing. In total, we downloaded >370,000 outcomes for preclinical assays of antisarcoma drug candidates from the ChEMBL database. The keywords (fields) used for the search were as follows: Sarcoma (Assay) and also keywords for more relevant cell osteosarcoma lines MG-63, U2O2, HOS, SAOS-2, and 143B. After that, we carried out a data fusion of the datasets obtained into one single raw dataset. The working dataset was curated by eliminating all duplicated entries. We also eliminated all cases with missing values of biological activity (v_{ij}) and/or molecular descriptors. The molecular descriptors used were the same as those precalculated by the ChEMBL database where $D_1 = \text{logP}$ and $D_2 = \text{PSA}$.^{13,14} The final dataset obtained after curation contained 37,919 cases comprising 36 protein targets, 43 cell lines, and 17 assay organisms (Table S1). For comparison and exploration with other models, we additionally computed 12 BCUT molecular descriptors⁶⁴ with ChemAxon (<http://www.chemaxon.com>). The classical unweighted Burden descriptors as well as those weighted by charge and hydrogen bond properties were calculated. The lowest and the three highest eigenvalues were used for descriptor calculation.

To train the model, we split this dataset into two data subsets: training and validation series. We performed a random, stratified, and representative selection of training/validation cases. To accomplish this task, we sorted the cases by n_j (from highest to lowest) as well as by assay conditions:

biological activity, protein accession, cell line, and assay organism (alphabetically from A to Z). After this, we selected every fourth case (1 out of 4) to form a training subset (75% of cases) and validation subset (25% of cases). The result of each experimental assay is the value obtained from the quantification of each biological activity and named v_{ij} (“i” and “j” represent the assay and conditions, respectively). Each biological activity depends on the conditions c_j ($c_0, c_1, c_2, \dots, c_n$) used in each assay. Thus, the conditions taken into account in the data preprocessing were $c_0 = \text{biological activity}$, $c_1 = \text{protein accession}$, $c_2 = \text{cell line}$, and $c_3 = \text{assay organism}$. From v_{ij} , each experimental assay was discretized based on the desirability $d(c_0)$. This variable was defined as 1 when the result of the desired biological activity depended on an increased value of v_{ij} and -1 when the desired biological activity depended on a lower value of v_{ij} . Thus, the discretized value $f(v_{ij})_{\text{obs}}$ was calculated as follows: $f(v_{ij})_{\text{obs}} = 1$ when $v_{ij} > \text{cut-off}$ and $d(c_0) = 1$. The function $f(v_{ij})_{\text{obs}} = 1$ when $v_{ij} < \text{cut-off}$ and $d(c_0) = -1$; otherwise, $f(v_{ij})_{\text{obs}} = 0$. The value $f(v_{ij})_{\text{obs}} = 1$ refers to a strong effect of the compound over the target. Since $d(c_0)$ has a direct relationship with $f(v_{ij})_{\text{obs}}$, we applied a rational cut-off for each c_0 , which will be discussed later. Briefly, the cut-off for properties related to drug concentrations and described in nM (potency, IC_{50} , CC_{50} , EC_{50} , GI_{50} , etc.) was set at 100. For properties described in % (inhibition, activity, TGI, among others), the cut-off was set at 50. Last, to calculate the probability of these expected values, we evaluated the relationship between the total number of the observed $n(f(v_{ij}) = 1)_{\text{obs}}$ within the level of biological activity desired for the condition c_j and the total number of compounds n_j that were described in that same condition. In this sense, we have that $p(f(v_{ij})_{\text{obs}} = 1)_{\text{expt}} = n(f(v_{ij}) = 1)_{\text{obs}}/c_0$.

PTML Linear Model. The multicondition moving averages (MMAs) are PTOs similar to Box–Jenkins moving average operators. However, MMAs are PTOs accounting for perturbations (changes) in multiple conditions c_j at the same time, while MA quantifies changes in only one condition. By using linear discriminant analysis (LDA),⁶⁵ we obtained a PTML-LDA equation as follows

$$f(v_{ij})_{\text{calc}} = a_0 + a_1 \cdot f(v_{ij})_{\text{ref}} + \sum_{k=1}^{k_{\text{max}}} a_{k_j} \cdot D_k + \sum_{k=1, j=0}^{k_{\text{max}}, j_{\text{max}}} a_{k_j} \cdot \Delta D_k(c_j)$$

The model generates an output score $f(v_{ij})_{\text{calc}}$ that refers to a score function for a biological activity v_{ij} under the assay conditions c_j . The LDA algorithm includes the Mahalanobis' distance metric,⁶⁵ which makes it possible to infer predictive values through a probability calculation $p(f(v_{ij}) = 1)_{\text{pred}}$. For the variable selection, we detected specific perturbations within the conditions c_j that will be adjusted to anticancer properties through a forward-stepwise strategy.⁶⁵ Such conditions as $c_1 = \text{protein accession}$, $c_2 = \text{cell line}$, and $c_3 = \text{assay organism}$ were significant, so we took them into consideration in our model. Through $p(f(v_{ij}) = 1)_{\text{pred}}$, we predicted the activity of each compound by applying the function $f(v_{ij})_{\text{pred}} = 1$ when $p(f(v_{ij}) = 1)_{\text{pred}} > 0.5$ or $f(v_{ij})_{\text{pred}} = 0$.

For comparison, we also used a strategy that is not based on perturbation theory. In this sense, besides the molecular descriptors, we added conditions c_1 , c_2 , and c_3 as a separate set of categorical variables. A total of 237 variables were needed to represent all conditions. Filtering using the variance of each

variable leads to a total of 162 variables, including ALOGP and PSA.

The evaluation of the discriminant model was calculated from Wilks' lambda (Λ) as follows

$$\Lambda = \left[\frac{1}{1 + \lambda} \right]$$

where Λ is chi-square distributed for $df = (k - 1)$, k is equal to the number of parameters estimated, and $\lambda = \left[\frac{\sum (z_j - \bar{z})^2}{\sum (z_{ij} - \bar{z}_i)^2} \right]$.

For ML, besides LDA, we also used neural networks (NN) with different architectures. STATISTICA software was used in both cases. The final networks obtained were multilayer perceptron (MLP), linear neural network (LNN), and radial basis function network (RBF). All these ML strategies were applied with perturbation and nonperturbation theory. The predicted 1 or 0 values were used to determine the specificity or true-negative rate (Sp), sensitivity or true-positive rate (Sn), and accuracy (Ac) when compared to the observed values. Thus, when $f(v_{ij})_{pre} = f(v_{ij})_{obs}$, the cases were determined to be correct.⁶⁵

The metrics to evaluate the performance of all the prediction models were Ac, Sn, and Sp using the following formulae

$$Ac = \frac{\text{number of correctly classified compounds}}{\text{total number of compounds}}$$

$$Sn = \frac{\text{number of correctly classified active compounds}}{\text{total number of active compounds}}$$

$$Sp = \frac{\text{number of correctly classified inactive compounds}}{\text{total number of inactive compounds}}$$

■ ASSOCIATED CONTENT

■ Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsomega.0c03356>.

ChEMBL dataset of antisarcoma preclinical experimental assays for the PTML model; results of the analyzed models for sarcoma biological activities; all the multiple-condition averages for all sarcoma assays (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Alejandro Cabrera-Andrade – *Grupo de Bio-Quimioinformática and Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de Las Américas, Quito 170125, Ecuador; RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, A Coruña 15071, Spain; orcid.org/0000-0001-9702-6618; Email: raul.cabrera@udla.edu.ec*

Humbert González-Díaz – *Department of Organic Chemistry II and Basque Center for Biophysics, University of Basque Country UPV/EHU, Leioa 48940, Biscay, Spain; Ikerbasque, Basque Foundation for Science, Bilbao 48011, Biscay, Spain; orcid.org/0000-0002-9392-2797; Email: humberto.gonzalezdiaz@ehu.es*

Authors

Andrés López-Cortés – *RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, A Coruña 15071, Spain;*

Centro de Investigación Genética y Genómica, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Quito 170129, Ecuador

Cristian R. Munteanu – *RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, A Coruña 15071, Spain; Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), A Coruña 15006, Spain; Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Campus de Elviña s/n, A Coruña 15071, Spain; orcid.org/0000-0002-5628-2268*

Alejandro Pazos – *RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, A Coruña 15071, Spain; Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), A Coruña 15006, Spain*

Yunierkis Pérez-Castillo – *Grupo de Bio-Quimioinformática and Escuela de Ciencias Físicas y Matemáticas, Universidad de Las Américas, Quito 170125, Ecuador*

Eduardo Tejera – *Grupo de Bio-Quimioinformática and Facultad de Ingeniería y Ciencias Aplicadas, Universidad de Las Américas, Quito 170125, Ecuador*

Sonia Arrasate – *Department of Organic Chemistry II and Basque Center for Biophysics, University of Basque Country UPV/EHU, Leioa 48940, Biscay, Spain*

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acsomega.0c03356>

Author Contributions

*A.C.-A. and A.L.-C. contributed equally to the study.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge research grants from Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P), and Basque government (IT1045-16). The authors also acknowledge the support of Ikerbasque, Basque Foundation for Science. This work was supported by Universidad de Las Américas and the Collaborative Project in Genomic Data Integration (CICLOGEN) PI17/01826 funded by the Carlos III Health Institute from the Spanish National Plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER)—“A way to build Europe”. This project was also supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia ED431D 2017/16 and “Drug Discovery Galician Network” ref. ED431G/01 and the “Galician Network for Colorectal Cancer Research” (ref. ED431D 2017/23) and finally by the Spanish Ministry of Economy and Competitiveness for its support through the funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union. Additional support was offered by the Consolidation and Structuring of Competitive Research Units—Competitive Reference Groups (ED431C 2018/49), funded by the Ministry of Education, University and Vocational Training of the Xunta de Galicia endowed with EU FEDER funds.

REFERENCES

- (1) Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R. L.; Torre, L. A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424.
- (2) Hui, J. Y. C. Epidemiology and Etiology of Sarcomas. *Surg. Clin. North Am.* **2016**, *96*, 901–914.
- (3) Sidaway, P. Sarcoma: Genetic determinants of sarcoma risk revealed. *Nat. Rev. Clin. Oncol.* **2016**, *13*, 590.
- (4) Thomas, D. M.; Ballinger, M. L. Etiologic, environmental and inherited risk factors in sarcomas. *J. Surg. Oncol.* **2015**, *111*, 490–495.
- (5) HaDuong, J. H.; Martin, A. A.; Skapek, S. X.; Mascarenhas, L. Sarcomas. *Pediatr. Clin. North Am.* **2015**, *62*, 179–200.
- (6) Yang, J.; Ren, Z.; Du, X.; Hao, M.; Zhou, W. The role of mesenchymal stem/progenitor cells in sarcoma: update and dispute. *Stem Cell Investig.* **2014**, *1*, 18.
- (7) Double, J.; Barrass, N.; Barnard, N. D.; Navaratnam, V. Toxicity testing in the development of anticancer drugs. *Lancet. Oncol.* **2002**, *3*, 438–442.
- (8) Yap, T. A.; Sandhu, S. K.; Workman, P.; de Bono, J. S. Envisioning the future of early anticancer drug development. *Nat. Rev. Cancer* **2010**, *10*, 514–523.
- (9) Williams, R. J.; Walker, I.; Takle, A. K. Collaborative approaches to anticancer drug discovery and development: a Cancer Research UK perspective. *Drug Discovery Today* **2012**, *17*, 185–187.
- (10) Heinemann, F.; Huber, T.; Meisel, C.; Bundschus, M.; Leser, U. Reflection of successful anticancer drug development processes in the literature. *Drug Discovery Today* **2016**, *21*, 1740–1744.
- (11) Sun, J.; Wei, Q.; Zhou, Y.; Wang, J.; Liu, Q.; Xu, H. A systematic analysis of FDA-approved anticancer drugs. *BMC Syst. Biol.* **2017**, *11*, 87.
- (12) Carvalho-Silva, D.; Pierleoni, A.; Pignatelli, M.; Ong, C.; Fumis, L.; Karamanis, N.; Carmona, M.; Faulconbridge, A.; Hercules, A.; McAuley, E.; Miranda, A.; Peat, G.; Spitzer, M.; Barrett, J.; Hulcoop, D. G.; Papa, E.; Koscielny, G.; Dunham, I. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* **2019**, *47*, D1056–D1065.
- (13) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.
- (14) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.
- (15) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538–1546.
- (16) Ali, M.; Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys. Rev.* **2019**, *11*, 31–39.
- (17) Wang, J.; Yun, D.; Yao, J.; Fu, W.; Huang, F.; Chen, L.; Wei, T.; Yu, C.; Xu, H.; Zhou, X.; Huang, Y.; Wu, J.; Qiu, P.; Li, W. Design, synthesis and QSAR study of novel isatin analogues inspired Michael acceptor as potential anticancer compounds. *Eur. J. Med. Chem.* **2018**, *144*, 493–503.
- (18) Pogorzelska, A.; Sławiński, J.; Żołnowska, B.; Szafranski, K.; Kawiak, A.; Chojnacki, J.; Ulenberg, S.; Zielińska, J.; Bączek, T. Novel 2-(2-alkylthiobenzenesulfonyl)-3-(phenylprop-2-ynylideneamino)-guanidine derivatives as potent anticancer agents - Synthesis, molecular structure, QSAR studies and metabolic stability. *Eur. J. Med. Chem.* **2017**, *138*, 357–370.
- (19) Sławiński, J.; Szafranski, K.; Pogorzelska, A.; Żołnowska, B.; Kawiak, A.; Macur, K.; Belka, M.; Bączek, T. Novel 2-benzylthio-5-(1,3,4-oxadiazol-2-yl)benzenesulfonamides with anticancer activity: Synthesis, QSAR study, and metabolic stability. *Eur. J. Med. Chem.* **2017**, *132*, 236–248.
- (20) Singh, H.; Kumar, R.; Singh, S.; Chaudhary, K.; Gautam, A.; Raghava, G. P. S. Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. *BMC Cancer* **2016**, *16*, 77.
- (21) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. SMILES-based QSAR approaches for carcinogenicity and anticancer activity: comparison of correlation weights for identical SMILES attributes. *Anti-Cancer Agents Med. Chem.* **2011**, *11*, 974–982.
- (22) González-Díaz, H.; Bonet, I.; Terán, C.; De Clercq, E.; Bello, R.; García, M. M.; Santana, L.; Uriarte, E. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* **2007**, *42*, 580–585.
- (23) González-Díaz, H.; Viña, D.; Santana, L.; de Clercq, E.; Uriarte, E. Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg. Med. Chem.* **2006**, *14*, 1095–1107.
- (24) González-Díaz, H.; Gia, O.; Uriarte, E.; Hernández, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. Markovian chemicals “in silico” design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: discovery of anticancer compounds. *J. Mol. Model.* **2003**, *9*, 395–407.
- (25) Jung, M.; Kim, H.; Kim, M. Chemical genomics strategy for the discovery of new anticancer agents. *Curr. Med. Chem.* **2003**, *10*, 757–762.
- (26) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189–199.
- (27) Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; Wilson, C. J.; Lehár, J.; Kryukov, G. V.; Sonkin, D.; Reddy, A.; Liu, M.; Murray, L.; Berger, M. F.; Monahan, J. E.; Morais, P.; Meltzer, J.; Korejwa, A.; Jané-Valbuena, J.; Mapa, F. A.; Thibault, J.; Bric-Furlong, E.; Raman, P.; Shipway, A.; Engels, I. H.; Cheng, J.; Yu, G. K.; Yu, J.; Aspesi, P.; de Silva, M.; Jagtap, K.; Jones, M. D.; Wang, L.; Hatton, C.; Palessandolo, E.; Gupta, S.; Mahan, S.; Sougnez, C.; Onofrio, R. C.; Liefeld, T.; MacConaill, L.; Winckler, W.; Reich, M.; Li, N.; Mesirov, J. P.; Gabriel, S. B.; Getz, G.; Ardlie, K.; Chan, V.; Myer, V. E.; Weber, B. L.; Porter, J.; Warmuth, M.; Finan, P.; Harris, J. L.; Meyerson, M.; Golub, T. R.; Morrissey, M. P.; Sellers, W. R.; Schlegel, R.; Garraway, L. A. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–607.
- (28) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. CORAL: classification model for predictions of anti-sarcoma activity. *Curr. Top. Med. Chem.* **2012**, *12*, 2741–2744.
- (29) Vos, H. I.; Coenen, M. J. H.; Guchelaar, H.-J.; Maroeska, D.; te Loo, D. M. The role of pharmacogenetics in the treatment of osteosarcoma. *Drug Discovery Today* **2016**, *21*, 1775–1786.
- (30) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463–477.
- (31) Blázquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck-Heck cascade reactions. *RSC Adv.* **2016**, *6*, 38602–38610.
- (32) M Casañola-Martin, G.; Le-Thi-Thu, H.; Pérez-Giménez, F.; Marrero-Ponce, Y.; Merino-Sanjuán, M.; Abad, C.; González-Díaz, H. Multi-output Model with Box-Jenkins Operators of Quadratic Indices for Prediction of Malaria and Cancer Inhibitors Targeting Ubiquitin-Proteasome Pathway (UPP) Proteins. *Curr. Protein Pept. Sci.* **2016**, *17*, 220–227.

- (33) Romero-Durán, F. J.; Alonso, N.; Yañez, M.; Caamaño, O.; García-Mera, X.; González-Díaz, H. Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology* **2016**, *103*, 270–278.
- (34) Kleandrova, V. V.; Luan, F.; González-Díaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S. Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. *Environ. Sci. Technol.* **2014**, *48*, 14686–14694.
- (35) Luan, F.; Kleandrova, V. V.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. D. S. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale* **2014**, *6*, 10623–10630.
- (36) Alonso, N.; Caamaño, O.; Romero-Durán, F. J.; Luan, F.; Cordeiro, M. N. D. S.; Yañez, M.; González-Díaz, H.; García-Mera, X. Model for High-Throughput Screening of Multitarget Drugs in Chemical Neurosciences: Synthesis, Assay, and Theoretic Study of Rasagiline Carbamates. *ACS Chem. Neurosci.* **2013**, *4*, 1393–1403.
- (37) González-Díaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Curr. Top. Med. Chem.* **2013**, *13*, 1713–1741.
- (38) Kleandrova, V. V.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS Comb. Sci.* **2016**, *18*, 490–498.
- (39) Speck-Planche, A.; Cordeiro, M. N. D. S. Simultaneous virtual prediction of anti-*Escherichia coli* activities and ADMET profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS Comb. Sci.* **2014**, *16*, 78–84.
- (40) Speck-Planche, A.; Cordeiro, M. N. D. S. Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based in Silico Approach for the Design of Virtual Anti-Hepatitis C Leads. *ACS Comb. Sci.* **2017**, *19*, 501–512.
- (41) Speck-Planche, A.; Cordeiro, M. N. D. S. Computer-aided discovery in antimicrobial research: In silico model for virtual screening of potent and safe anti-pseudomonas agents. *Comb. Chem. High Throughput Screening* **2015**, *18*, 305–314.
- (42) Speck-Planche, A.; Cordeiro, M. N. D. S. Erratum to: Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Diversity* **2017**, *21*, 525.
- (43) Speck-Planche, A.; Cordeiro, M. N. D. S. Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Diversity* **2017**, *21*, 511–523.
- (44) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. *Anti-Cancer Agents Med. Chem.* **2013**, *13*, 791–800.
- (45) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anti-Cancer Agents Med. Chem.* **2012**, *12*, 678–685.
- (46) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorg. Med. Chem.* **2012**, *20*, 4848–4855.
- (47) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *Eur. J. Pharm. Sci.* **2012**, *47*, 273–279.
- (48) Cordeiro, M. N.; Speck-Planche, A. Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. *Curr. Top. Med. Chem.* **2012**, *12*, 2703–2704.
- (49) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg. Med. Chem.* **2011**, *19*, 6239–6244.
- (50) Wei, D.-Q.; Selvaraj, G.; Kaushik, A. C. Computational Perspective on the Current State of the Methods and New Challenges in Cancer Drug Discovery. *Curr. Pharm. Des.* **2018**, *24*, 3725–3726.
- (51) Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb. Sci.* **2018**, *20*, 621–632.
- (52) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Fragment-based QSAR model toward the selection of versatile anti-sarcoma leads. *Eur. J. Med. Chem.* **2011**, *46*, 5910–5916.
- (53) Chittchang, M.; Gleeson, M. P.; Ploypradith, P.; Ruchirawat, S. Assessing the drug-likeness of lamellarins, a marine-derived natural product class with diverse oncological activities. *Eur. J. Med. Chem.* **2010**, *45*, 2165–2172.
- (54) Hansch, C.; Verma, R. P. A QSAR study for the cytotoxic activities of taxoids against macrophage (MPhi)-like cells. *Eur. J. Med. Chem.* **2009**, *44*, 274–279.
- (55) Roy, K.; Pratim Roy, P. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. *Eur. J. Med. Chem.* **2009**, *44*, 2913–2922.
- (56) Sarkar, A.; Anderson, K. C.; Kellogg, G. E. Computational analysis of structure-based interactions and ligand properties can predict efflux effects on antibiotics. *Eur. J. Med. Chem.* **2012**, *52*, 98–110.
- (57) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **2014**, *66*, 334–395.
- (58) Leeson, P. Drug discovery: Chemical beauty contest. *Nature* **2012**, *481*, 455–456.
- (59) Arnott, J. A.; Planey, S. L. The influence of lipophilicity in drug discovery and design. *Expert Opin. Drug Discovery* **2012**, *7*, 863–875.
- (60) Yuan, H.; Paskov, I.; Paskov, H.; González, A. J.; Leslie, C. S. Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* **2016**, *6*, 31619.
- (61) Nikolova, O.; Moser, R.; Kemp, C.; Gönen, M.; Margolin, A. A. Modeling gene-wise dependencies improves the identification of drug response biomarkers in cancer studies. *Bioinformatics* **2017**, *33*, 1362–1369.
- (62) Waring, M. J. Lipophilicity in drug discovery. *Expert Opin. Drug Discovery* **2010**, *5*, 235–248.
- (63) Giaginis, C.; Tsopelas, F.; Tsantili-Kakoulidou, A. The Impact of Lipophilicity in Drug Discovery: Rapid Measurements by Means of Reversed-Phase HPLC. *Methods Mol. Biol.* **1824**, 1824, 217–228.
- (64) Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.
- (65) Hill, T.; Lewicki, P. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining. In *Discriminant Function Analysis*; 1st ed.; StatSoft, Inc.: 2006; 155–164.



Perspective

A Multi-Objective Approach for Anti-Osteosarcoma Cancer Agents Discovery through Drug Repurposing

Alejandro Cabrera-Andrade ^{1,2,3,*} , Andrés López-Cortés ^{3,4,5} ,
Gabriela Jaramillo-Koupermann ⁶, Humberto González-Díaz ^{7,8} , Alejandro Pazos ^{3,9} ,
Cristian R. Munteanu ^{3,9} , Yunierkis Pérez-Castillo ^{1,10} and Eduardo Tejera ^{1,11,*}

¹ Grupo de Bio-Quimioinformática, Universidad de Las Américas, Quito 170125, Ecuador; yunierkis.perez@udla.edu.ec

² Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de Las Américas, Quito 170125, Ecuador

³ Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, CITIC, Campus Elviña s/n, 15071 A Coruña, Spain; aalc84@gmail.com (A.L.-C.); apazos@udc.es (A.P.); c.munteanu@udc.es (C.R.M.)

⁴ Centro de Investigación Genética y Genómica, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Quito 170129, Ecuador

⁵ Latin American Network for Implementation and Validation of Clinical Pharmacogenomics Guidelines (RELIVAF-CYTED), 28029 Madrid, Spain

⁶ Laboratorio de Biología Molecular, Subproceso de Anatomía Patológica, Hospital de Especialidades Eugenio Espejo, Quito 170403, Ecuador; gaby_jaramillok@yahoo.com

⁷ Department of Organic and Inorganic Chemistry, and Basque Center for Biophysics CSIC-UPV/EHU, University of the Basque Country UPV/EHU, 48940 Leioa, Spain; humberto.gonzalezdiaz@ehu.es

⁸ IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

⁹ Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), 15006 A Coruña, Spain

¹⁰ Escuela de Ciencias Físicas y Matemáticas, Universidad de Las Américas, Quito 170125, Ecuador

¹¹ Facultad de Ingeniería y Ciencias Agropecuarias, Universidad de Las Américas, Quito 170125, Ecuador

* Correspondence: raul.cabrera@udla.edu.ec (A.C.-A.); eduardo.tejera@udla.edu.ec (E.T.)

Received: 2 October 2020; Accepted: 12 November 2020; Published: 22 November 2020



Abstract: Osteosarcoma is the most common type of primary malignant bone tumor. Although nowadays 5-year survival rates can reach up to 60–70%, acute complications and late effects of osteosarcoma therapy are two of the limiting factors in treatments. We developed a multi-objective algorithm for the repurposing of new anti-osteosarcoma drugs, based on the modeling of molecules with described activity for HOS, MG63, SAOS2, and U2OS cell lines in the ChEMBL database. Several predictive models were obtained for each cell line and those with accuracy greater than 0.8 were integrated into a desirability function for the final multi-objective model. An exhaustive exploration of model combinations was carried out to obtain the best multi-objective model in virtual screening. For the top 1% of the screened list, the final model showed a BEDROC = 0.562, EF = 27.6, and AUC = 0.653. The repositioning was performed on 2218 molecules described in DrugBank. Within the top-ranked drugs, we found: temsirolimus, paclitaxel, sirolimus, everolimus, and cabazitaxel, which are antineoplastic drugs described in clinical trials for cancer in general. Interestingly, we found several broad-spectrum antibiotics and antiretroviral agents. This powerful model predicts several drugs that should be studied in depth to find new chemotherapy regimens and to propose new strategies for osteosarcoma treatment.

Keywords: osteosarcoma; machine learning; multi-objective model; virtual screening; drug repositioning

1. Introduction

Osteosarcoma (OS) is the most common primary bone tumor in children, adolescents and young adults, representing approximately 3.5% of all childhood cancers and 56% of malignant bone tumors in children. Its incidence rate ranges between 1 and 5 cases per million people and it is usually diagnosed in patients who are 10 to 19 years old. OS follows a bimodal distribution, with an initial peak in late adolescence and young adulthood and a second peak in old age [1].

The management of patients diagnosed with OS has not changed in recent decades. Current systemic OS first-line therapy includes cycles of cisplatin, doxorubicin, and high-dose methotrexate (MAP). Second-line therapy can integrate some tyrosine kinase inhibitors such as sorafenib and everolimus, plus antineoplastic agents like etoposide, topotecan and cyclophosphamide [2]. Neoadjuvant chemotherapy is generally administered for a period of 10 weeks, followed by the surgical resection of the compromised tumor area and radiotherapy. If 90% or more of the tumor area shows necrosis, additional cycles of postoperative therapy are applied to reject micrometastasis [3,4].

The prognosis of this disease is highly variable, possibly due to its high rate of tumor mutations, which leads to widespread dysregulation in cell signaling pathways and genomic instability [5]. Patients with localized disease show a 5-year survival rate of 65 to 70%, while for those who develop metastases the rate drops to 19–30% [6]. These metastatic events involve the lung parenchyma in 75% of the cases and distant skeletal sites [7,8], hindering treatment efficacy [9]. In this scenario, current therapy shows little response sensitivity and the survival rate decreases considerably.

Despite current chemotherapy regimens being the most effective strategy for OS treatment, patients' sensitivity to these agents regarding the toxic side-effects and antitumor effects varies considerably [10,11]. Several clinical trials have developed experimental designs to improve survival rates by testing dose intensification, and also adding or combining various chemotherapeutic agents [12]. There is a dose effect on treatment response, but several studies have shown that high-dose chemotherapy may not increase survival rates any more than less toxic moderate doses [13]. Due to the lack of tumor specificity or metastasis events or the complex etiology of these bone tumors, the anti-OS compounds currently used have a narrow therapeutic index and no increase in survival rates have been achieved in the last three decades [14], thus the therapeutic strategies need to be optimized.

The development and validation of novel therapeutic compounds is a time-consuming and labor-intensive process. Drug repositioning, which explores potential novel uses for known molecules based on prediction algorithms has become an effective and innovative approach [15–17]. One approach is based on multi-objective computational models where the repositioning process is addressed from a set of potentially desirable solutions. To do this, some computational techniques have been applied and these include Quantitative Structure-Activity Relationship (QSAR) and Ligand-Based Virtual Screening, which aid the identification of hit structures [18–21]. These QSAR studies are used to perform virtual drug screening that has been integrated into the drug discovery pipeline and could save both time and money, especially in the early phase of drug discovery [22,23].

In this sense, the application of these models is of high interest to researchers specializing in cancer. Several studies have focused on the description of new therapeutic agents, especially for treating carcinomas [24–32]. However, very few have concentrated on tumors of mesenchymal origin [33,34]. Thus, we developed a multi-objective model for the prediction of drugs with potential biological activity towards OS, one of the most prevalent cancers in pediatric populations where current chemotherapy treatments have not varied in the last decades.

2. Results

2.1. Datasets and Molecular Descriptors

The ChEMBL database reports a total of 1250 compounds with biological activity for HOS, MG63, U2OS, and SAOS2 cell lines. Of these, 1036 shows complete information on their biological activities evaluated by IC₅₀, GI₅₀, and C₅₀ assays (Table S1). Before constructing the prediction algorithms for

each cell line, we inspected all those compounds reported in the DrugBank and separated them from this list for later use in virtual screening (VS).

Of the 1036 compounds, 28 drugs are reported in DrugBank for the HOS cell line, 30 for MG63, 31 for SAOS2, and 32 for U2OS. In this way, we removed these 121 drugs from the 1036, and the prediction models were built on the remaining 915 compounds. Thus, the calculation of the molecular descriptors (MDs) was performed on 277 compounds described for HOS, 124 for MG63, 173 for U2OS and 341 for SAOS2 (Figure 1A) and we obtained 500 ISIDA variables for each cell line (Table S2).

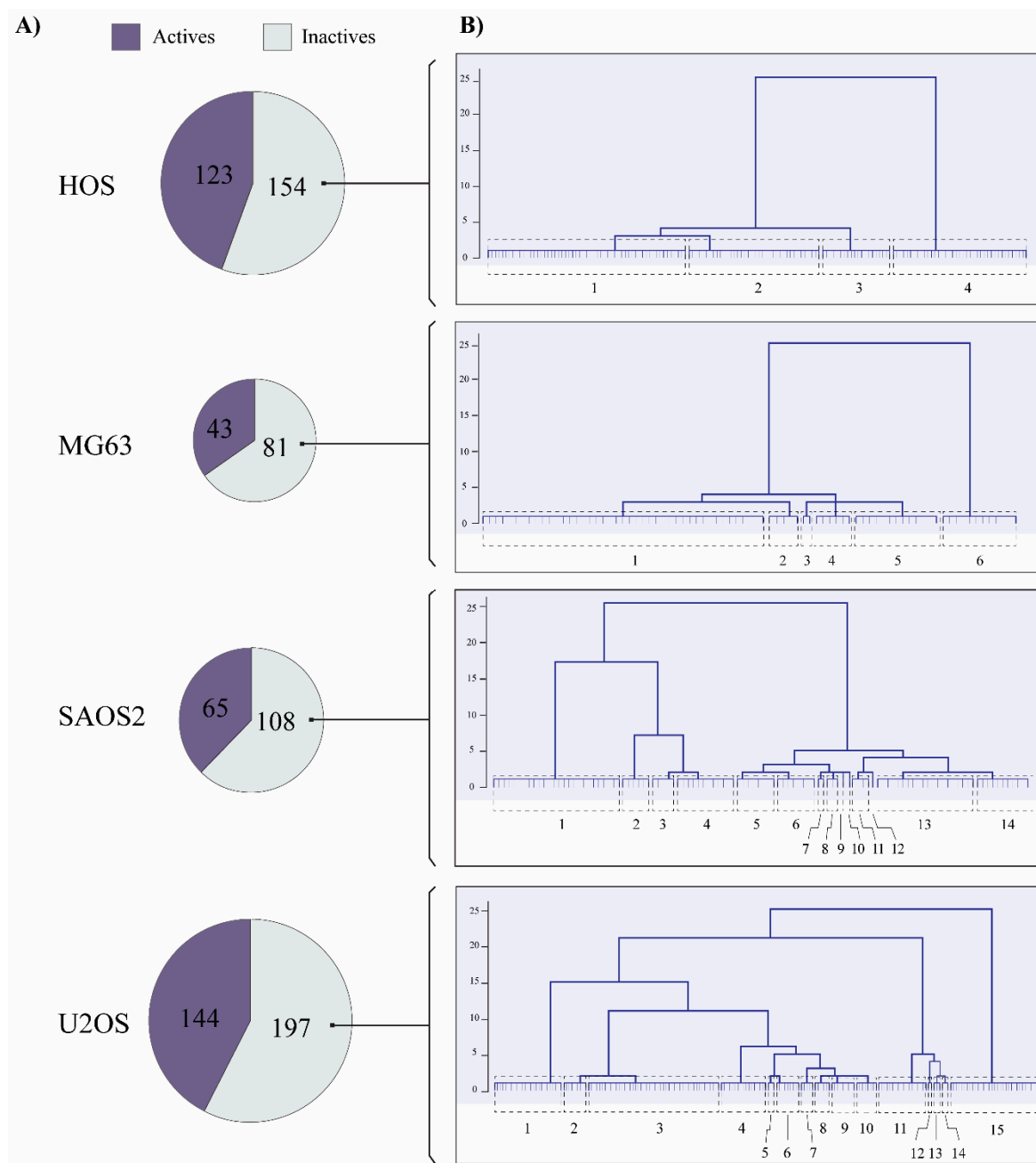


Figure 1. The chemical diversity of inactive compounds in OS cell lines. (A) Compounds with biological activity reported in ChEMBL for OS cell lines. (B) Dendrograms calculated for inactive compounds in the HOS, MG63, SAOS2 and U2OS cell lines.

Since inactive compounds are described in greater numbers than active ones in all cell lines, we evaluated the chemical diversity in the inactive series by applying a hierarchical clustering.

Thus, we calculated the degree of similarity in the inactive compounds to balance the data through stratified random sampling instead of a solely random partitioning. As a result, hierarchical representations were generated in which the clusters at each level of the hierarchy were created by merging clusters at the next level down [35]. In our case, we chose a strict cut-off to show all the possible groups and to make sure we had a wide representation of the chemical space within the data in the inactive series. From this, we identified four clusters in the HOS cell line, six in MG63, 14 in SAOS2 and 15 in U2OS (Figure 1B).

Therefore, we separated 24 inactive compounds for all cell lines (24 in each) reported in the DrugBank, as mentioned above, and then 7, 14, 19 and 29 compounds from each list. The balanced datasets resulted in 246 molecules for HOS, 86 for MG63, 130 for SAOS2, and 288 for U2OS, with a ratio of 1:1 between active and inactive compounds.

2.2. Construction of Models

The prediction algorithms used were: support vector machine (SVM), random forest (RF), neural networks (NN), decision tree (DTREE), k-Nearest Neighbors (KNN), and a scalable end-to-end tree boosting system (XGBoost) [36].

As seen in Figure 2, each of the six trained models demonstrated different performance metrics on the external data. HOS had similar achievements in the six learning techniques, but only the SVM, RF and XGBoost models showed optimal accuracy (AC) for subsequent assembly (0.836, 0.828 and 0.833 respectively). For MG63, the best strategies were SVM (0.882) and KNN (0.833). Prediction values when using RF, NN, XGBoost DTREE were less than 80%, and the true positive rate was lower than 0.7.

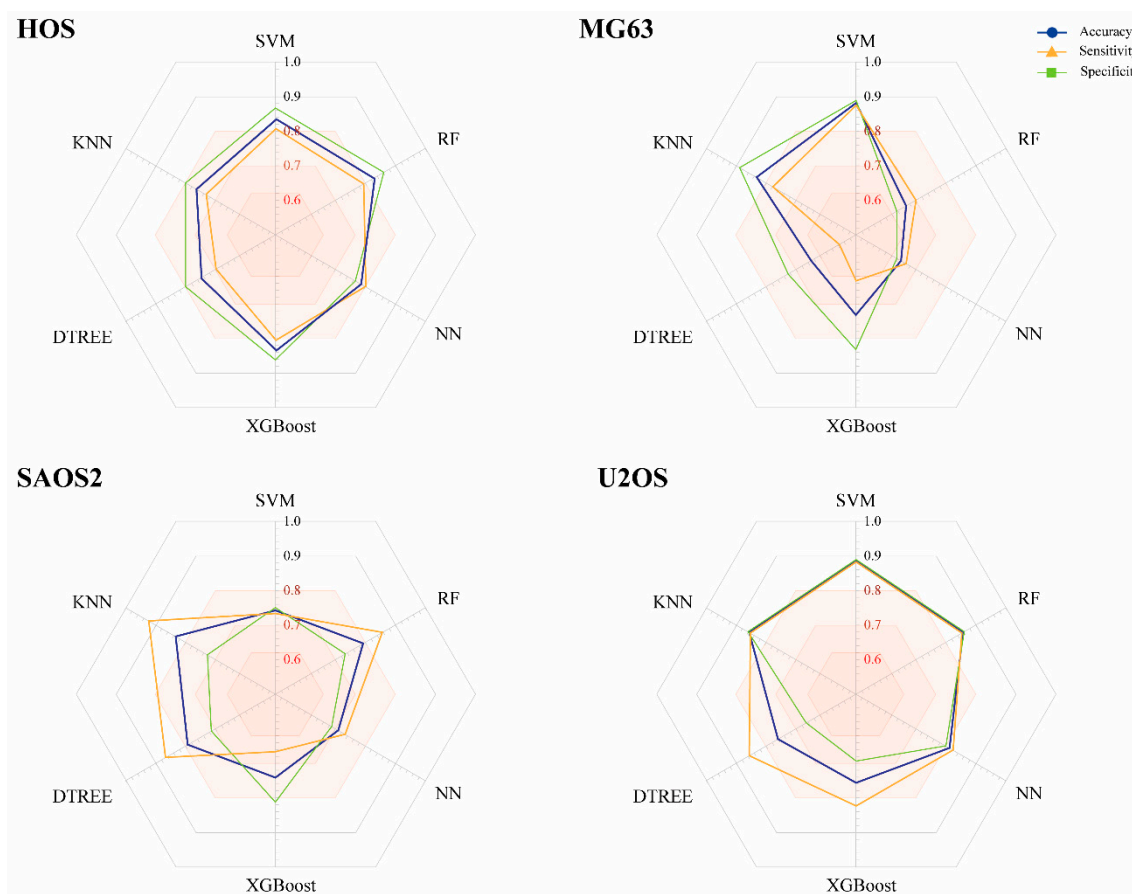


Figure 2. Performance of machine learning models constructed from compounds with biological activity for the HOS, MG63, SAOS2 and U2OS cell lines. Accuracy, sensitivity and specificity values correspond to the external data.

SAOS2 modeling resulted in only one algorithm with AC greater than 0.8, namely KNN (0.833). Lastly, the prediction models for U2OS with ACs > 0.8 were SVM, RF, NN, and KNN (0.886, 0.857, 0.812, and 0.857 respectively), all of whose SN and SP rates were higher than 0.8. XGBoost and DTREE were not taken into account for the assembly.

It is interesting to note that the feature selection by genetic algorithm (GA) reduced from 500 to 19–101, depending on each dataset. This strategy allowed us to generate the models described above and obtain the desired performance measurements for the final model (Table S3).

2.3. Multi-Objective Model Assessment and Virtual Screening

The AC, SN and specificity (SP) evaluated the performance of a model based on its training and data described as external, but these metrics do not always describe a desirable recovery rate at the time of performing screening for drug repositioning [37]. Therefore, the multi-objective model was evaluated in a virtual screening setting, where we mainly took into account the Area Under the Accumulation Curve (AUC), the Boltzmann-Enhanced Discrimination of ROC (BEDROC) and recovery efficiency (EF) at 1% of the screened list. The VS was developed on a dataset of 772 compounds. Of these, 14 corresponded to drugs used for previously described OS treatment, 653 were decoy molecules calculated from these 14 drugs, and 105 compounds were described as inactive and removed during the data balancing process.

We used all base-models with an AC greater than 0.8 in previously described external validation, and tested all possible combinations. Based on the VS results, the best multi-objective model was made up of the desirability values of the algorithms HOS-SVM, HOS-RF, MG63-SVM, SAOS2-KNN, U2OS-NN, and U2OS-KNN.

As seen in Figure 3A, our strategy generated a prediction method with better early recognition rates than the individual models. BEDROC is a metric that assigns more weight to early ranked molecules than late ranked molecules, therefore the initial enrichment was weighted. This enrichment was higher in our multi-objective model (BEDROC = 0.562) when calculated with an $\alpha = 160.9$. This means that our algorithm turned out to be the best strategy for recognizing “active” anti-sarcoma molecules in 1% of the list of therapeutic drugs for OS. Likewise, the EF value was higher in our algorithm when analyzing the recovery rate at 0.01. The EF values calculated for a recovery efficiency of 1% in the base models HOS-RF, SAOS2-KNN, and HOS-SVM were 20.68, 13.8, and 20.68, respectively, while in our method they reached 27.57. This indicates that with our protocol, it is possible to retrieve almost 27 times more the number of multi-targeted compounds in the first 1% of the ranked list than what is expected from a uniform distribution of the active ones in the virtual screening database.

When analyzing the active retrieved fractions of all the models (Figure 3B), one notices that all protocols have similar recovery rates within 30% of the data screened. However, a closer inspection of the screened data shows that the multi-objective has the highest recovery rate of anti-OS compounds at 1% (or less) of the data screened (Figure 3C). This algorithm recognized four compounds within the first six positions. These retrieve rates suggest that a strategy made up of several methods is capable of predicting those molecules described as active within 1% of a screened list. In a drug repositioning scenario, this is important since the compounds ranked in the first positions have a high probability of presenting biological activity in vitro. Using our prediction algorithm, a prediction rate of 59.9% would be expected in 1% of screening for drugs with anti-OS activity.

2.4. Analysis of Repurposed Drugs

The screening weighted four principal drug classes in the first 1% of the screened list: anti-infectives for systemic use (antimycobacterials, macrolides, protease inhibitors and tetracyclines; which represent 55%); antineoplastic/immunomodulating agents (immunosuppressants, protein kinase inhibitors and taxanes; 32%); dermatological/immunosuppressant (agents for dermatitis, excluding corticosteroids; 4%); and antiparasitics (an antinematodal agent and a broad-spectrum endectocide; 9%). The first two

groups represent more than 85% of all repositioned drugs (Figure 4A). All the desirability values for each repositioned drug are detailed in Table S4.

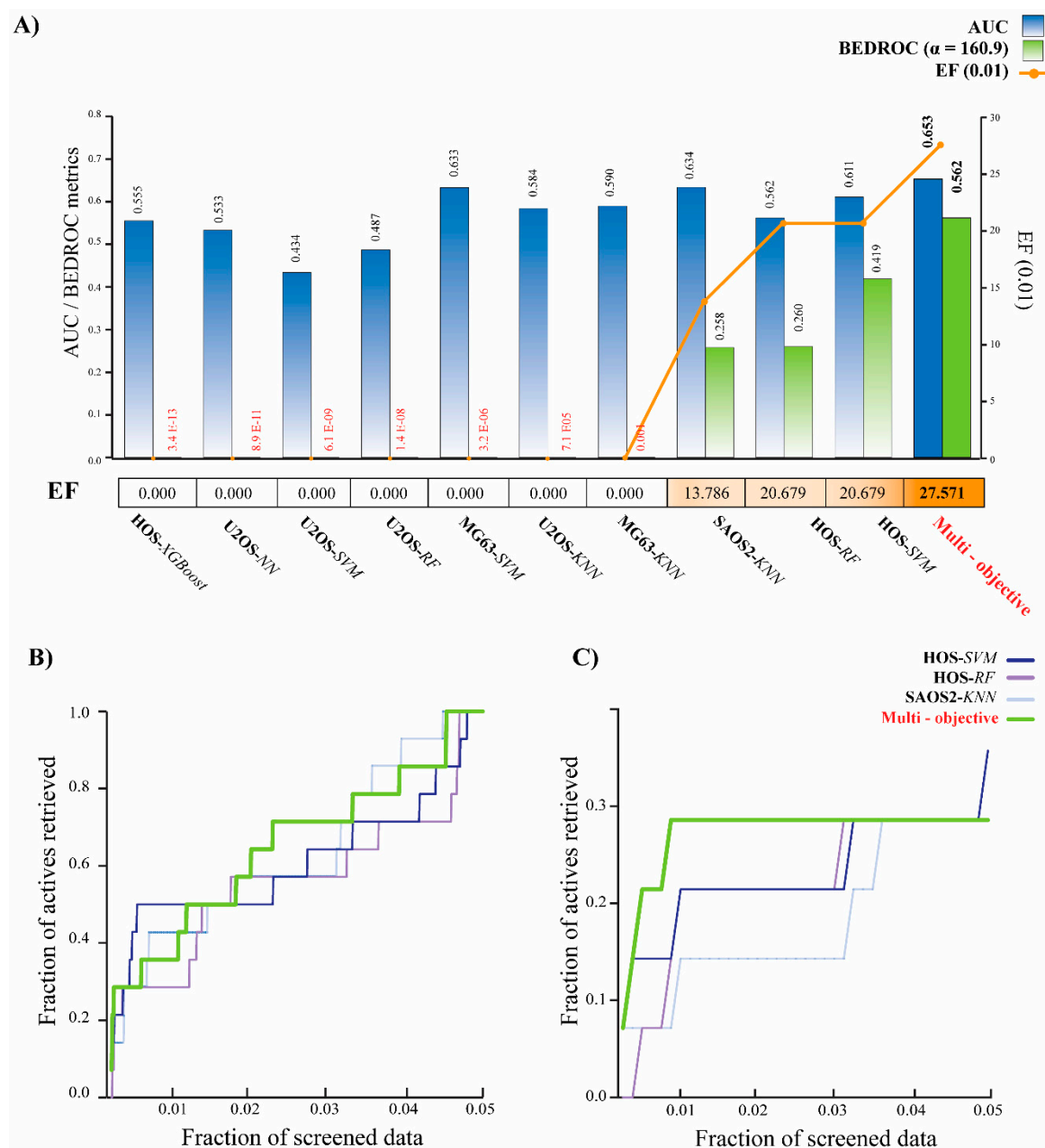


Figure 3. Results of the performance of base models and multi-objective models for the VS. (A) Comparison of AUC values (black bars) and BEDROC with $\alpha = 160.9$ of base models and the multi-objective algorithm. (B) Accumulative curves for the four top-performing VS protocols. The comparison includes the best 3 simple models and the multi-objective algorithm. Results are presented for the whole screening, and (C) for the top 5% of screened data.

The action mechanism of antineoplastic and immunomodulating agent mainly inhibits the mTOR pathway and microtubules polymerization. Among these, temsirolimus, paclitaxel, sirolimus (rapamicine), everolimus, cabazitaxel and docetaxel are ranked at the top. On the other hand, broad-spectrum antibacterials described as drugs that bind to the 30S/50s subunit of bacterial ribosome, HIV-1 protease inhibitors and antimycobacterials, which inhibit DNA-dependent RNA bacterial polymerase, were weighted in the screening. We also found two molecules used for the treatment

of HIV, described as inhibitors of HIV-1 protease (tipranavir and fosamprenavir) (Figure 4B). It is interesting to note that several repositioned drugs have been found within clinical trials for cancer patients. Out of the antineoplastic and immunomodulating agents, only cabazitaxel has not yet been studied in trials related to bone sarcomas. Moreover, broad-spectrum antibacterial compounds such as clarithromycin, erythromycin, doxycycline and tetracycline are top-ranked drugs that are registered in clinical trials for carcinomas.

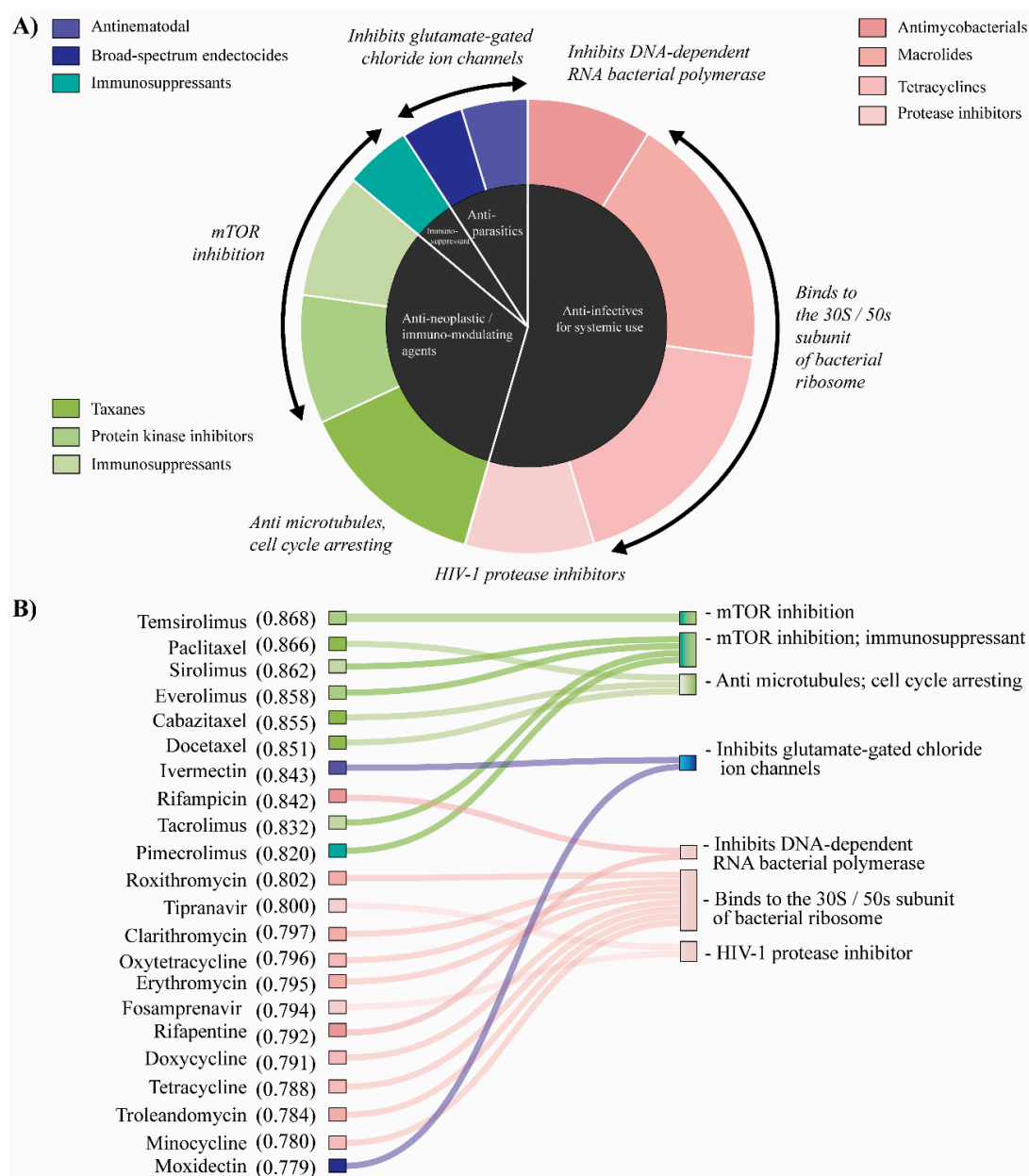


Figure 4. Repositioned drugs for OS treatment. (A) The central pie chart (black) shows the distribution of the 4 main drug classes repositioned at the first 1% of the screened list, whereas the outer pie chart shows the groups they represent. Each color represents a specific group obtained from the Anatomical Therapeutic Chemical (ATC) classification system. The action mechanisms of the screened drugs are also included in italics. (B) Correlation between the top-ranked drugs using the multi-objective model and its action mechanism. Listed are the first 22 positions (1%) of the 2218 DrugBank compounds screened. Drugs and their desirability values obtained using the prediction algorithm are described in the left column, while their action mechanism is in the right column. The colors represent the drug groups described in the graph above.

3. Discussion

Several reports have shown that multi-objective models have a better prediction rate during screening time since they approach the problem with a particular perspective from a set of potentially desirable solutions [18,21,38,39]. In our case, each of these possible desirable solutions was made up of each algorithm constructed from the described compounds with activity for the OS cell lines HOS, MG63, SAOS2 and U2OS. One of the major outcomes of this study is the improvement of the AUC and BEDROC values obtained in the VS, especially the EF of the multi-objective model when comparing with the performance of the base models (Figure 3A). This suggests that our algorithm improves the recognition rate of molecules described as therapeutic for OS treatment, especially within 1% of the data screened (Figure 3B). Specifically, the EF obtained indicates that it is possible to retrieve in the first 1% of the ranked list almost 27 times more multi-targeted compounds than what is expected from a uniform distribution of the active ones in the virtual screening database, something that is not obtained from the algorithms generated by each cell line.

Drug repositioning is an effective strategy for finding novel drug-disease relationships for existing molecules. The development of these strategies has gained considerable interest in recent years compared with de novo drug discovery pipeline, which demands more research time and experimental hours in the case of new drug development, and requires a greater financial investment. On the other hand, the use of already proven drugs is a highly efficient, low-cost and low-risk strategy since screening is carried out on molecules that have passed all clinical safety tests at Phase I, Phase II, and Phase III [40,41]. We used our multi-objective approach in order to propose new agents with chemotherapeutic activity for osteosarcoma treatment. Given the high recovery rate of active compounds obtained in our model (EF 0.01 = 27,571), we considered the first 22 highest-ranking compounds belonging to 1% of the 2218 approved FDA drugs reported in the DrugBank.

Of these 22 drugs, 13 (59.1%) are enrolled in clinical trials for cancer patients (reviewed at <https://clinicaltrials.gov/>) (Table S5): temsirolimus, paclitaxel, sirolimus/rapamycin, everolimus, cabazitaxel, docetaxel, rifampicin, tacrolimus, clarithromycin, erythromycin, doxycycline, tetracycline and minocycline. Interestingly, only five of these drugs are included in trials of patients with OS: temsirolimus, paclitaxel, sirolimus/rapamycin, everolimus, and docetaxel. The remaining 10 drugs (ivermectin, pimecrolimus, roxithromycin, tipranavir, oxytetracycline, fosamprenavir, rifapentine, troleandomycin and moxidectin) are not registered in any clinical trial for cancer patients, however, their action mechanisms are similar to various chemotherapeutic agents used in oncology practice.

Cancer cells are characterized by unregulated proliferation, which leads to cellular undifferentiation and disruption on the function of tissues. Cell proliferation can be caused by a checkpoint failure in cell cycle or a disruption in the cell death pathway. In this sense, any agent that affects the metabolism of cancer cells by reducing or inhibiting cell proliferation and promotes apoptosis is a potential target for cancer treatment [42]. Several agents used as first-line treatment for OS, such as methotrexate, doxorubicin, etoposide, cisplatin and ifosfamide, induce a disruption in these cellular functions, either by interrupting nucleotide synthesis, by DNA synthesis by inhibiting topoisomerase II, or by binding to double-strand DNA to promote apoptosis. On the other hand, several second-line drugs act on mTOR, a pathway considered pathogenic within the development and progression of OS [43–45], and on the formation of microtubules, inhibiting the progression from the G1 to the S phase of the cell cycle. In the top six positions of our screening, we found chemotherapeutic drugs described as therapeutic agents for various types of cancer (temsirolimus, paclitaxel, sirolimus, everolimus, cabazitaxel and docetaxel). Indeed, these compounds belong to one of the four principal drug classes found in our repositioning called antineoplastic and immune-modulatory agents (Figure 4A). Their action mechanism resembles those previously described as second-line drugs, which mainly inhibit mTOR and interfere with the microtubule depolymerization (Figure 4B). Interestingly, cabazitaxel is the only one of these six top-ranked compounds that is not reported in clinical trials of OS patients. This molecule is a semi-synthetic derivative of a natural taxoid that considerably increases overall survival versus mitoxantrone after prior docetaxel treatment in patients

with metastatic castration-resistant prostate cancer [46–48]. Cabazitaxel induces cell cycle arresting by interacting with the microtubule depolymerization by what is defined as a microtubule destabilizing agent. These types of agents show high antineoplastic activity and have been reported in previous studies into drug repositioning [49]. Although they commonly used in pediatric oncology [50], the microtubule-stabilizing taxanes are not often used to treat childhood cancers due to limited activity, even if safety is observed [51]. In this sense, cabazitaxel can be an important therapeutic agent for the treatment of OS, especially in patients who can progress onto it after docetaxel.

It is interesting to note that 54.5% of the total predicted compounds (12 out of 22) are classified as anti-infectives for systemic use. More specifically, taking into account the Anatomical Therapeutic Chemical (ATC) classification system, our protocol weighted several macrolides (roxithromycin, clarithromycin, erythromycin and troleandomycin), tetracyclines (oxytetracycline, doxycycline, tetracycline and minocycline), protease inhibitors (tipranavir and fosamprenavir) and antimycobacterial (rifampicin and rifapentine) as possible anti-OS agents (Figure 4B). On the one hand, prior studies into cancer therapy have noted the importance of macrolide and tetracycline compounds in cancer treatment [52,53]. Some authors have suggested that these groups of compounds inhibit the action of matrix metalloproteinases (MMPs) in order to reduce the degree of tumor invasion and metastases [54]. Others have observed that these drugs act on mitochondrial biogenesis [55,56], disrupting this process and thus increasing the effectiveness of chemotherapy or radiotherapy on tumor cells. On the other hand, the therapeutic action of HIV-protease inhibitors for the treatment of cancer has been reported. Although these molecules are not expected to cross-react with human peptides, preclinical data suggest that their antitumor activity may be linked in part to the inhibition of endopeptidases, such as metalloproteases and proteasomes [57]. Of our repositioned drugs, clarithromycin, erythromycin and doxycycline are currently under study as possible therapeutic agents for leukemia, colorectal, prostate and lung cancer, among others [58–61], and are involved in clinical trials of cancer patients (Table S5). Based on our findings, these agents could demonstrate antitumor activity in bone tumors.

These results may be promising for future preclinical and clinical studies. The lack of therapeutic options for OS should be the basis of searches for new agents as potential treatments. The discovery of molecular targets in OS will be part of the development of new molecules that could give these patients more options.

4. Materials and Methods

4.1. Preprocessing Datasets and Molecular Descriptors

Prediction models were developed from compounds described in the Chemical database (Version 25) of the European Molecular Biology Laboratory (ChEMBL) [62,63] with biological activity against the OS cell lines HOS (ChEMBL614736), MG63 (ChEMBL614347), SAOS (ChEMBL614894) and U2OS (ChEMBL615023). We considered all standard values evaluated by IC₅₀ (half-maximal inhibitory concentration), GI₅₀ (percentage of cell growth inhibition at a fixed concentration), and EC₅₀ (a concentration that inhibited half the cell culture growth), and from these scores, we defined a class for each compound.

Compounds with standard values > 10 µM were classified as inactive (0), and those with values < 10 µM as active (1). In those drugs where two or more assays are reported, and the standard values classify these compounds in different classes, the final criteria were assigned by most of the set of classes obtained. If more than 75% of the tests obtained classify a compound in the same class, this drug was included in the study, otherwise it was rejected. On the other hand, compounds that did not show information about their biological activity, inconclusive data about their activity, or incomplete information regarding ChEMBL ID or canonical SMILES were removed from the analysis.

We used the ChemAxon's JChem for Excel (18.8.0.253) [64] software to code the chemical structures in SMILES format. This information was converted to SD files (SDFs) and the structure of each compound was standardized using ChemAxon's Standardizer [65]. Explicit hydrogen atoms

were removed. Then we normalized specific chemotypes, such as nitro to one unique representation, the rings aromatization, the curation of tautomeric forms, the stripping of salts and small fragments. Furthermore, all duplicate structures were identified using the EdiSDF tool within the ISIDA/QSPR package and subsequently withdrawn from the list [66].

Two-dimensional molecular descriptors were computed with ISIDA Fragmentor 2017 [67,68]. The types of descriptors calculated were: Sequences of atoms and bonds; Atom-centered fragments based on sequences of atoms and bonds; Atom-centered fragments based on sequences of atoms and bonds of fixed length; and Triplets. For these calculations, the minimum and maximum length of fragments as sequences were set to 2 and 8, respectively. Molecular descriptors were calculated separately for each dataset.

The computed descriptors for each dataset were first filtered to remove those present in less than 1% of the compounds. Next, the Minimal Redundancy Maximal Relevance (mRMR) algorithm [69] was employed to keep the top 500 features in each dataset. For mRMR, the Mutual Information Quotient (MIQ) score was used as a features-ranking metric. This subset of 500 selected molecular descriptors was employed for QSAR modeling.

4.2. Machine Learning Models and Quality Evaluation

In each cell line under study, the amount of active and inactive compounds varied considerably. This unbalance in the dataset is not desirable for modeling. In order to balance the classes, the following procedure was executed in all cell lines. (1) Using the previously computed molecular descriptors, we carried out a hierarchical clustering. We applied the interval measure, the Euclidean distance and Ward's method for clustering, both in active and inactive compounds for each cell line. The IBM SPSS Statistics software v.25 (IBM Corp., Armonk, NY, USA) was employed to generate dendrograms and define all the clusters within the data. (2) Once the number of clusters had been defined, we continued with a random stratify extraction of the same amount of compound in both classes. This procedure had previously been used by other authors in order to obtain a balanced data representative of the chemical diversity space [70]. The training set consisted of 75% of randomly chosen compounds from the balanced dataset and the remaining percentage was utilized as external data. The external dataset was used to evaluate the model performance metrics.

To obtain each model, we applied genetic algorithms as feature selection by considering an initial population of 50 chromosomes and 30 generations. For validation of the fitness function in GA, we performed a cross-validation strategy using the average balanced classification rate (BRC) across 100 random splits (bootstrap sampling). This means that in each generation, 100 models were evaluated and the average AC was extracted. The models used together with genetic algorithm were: the support vector machine, random forest, neural networks, decision tree, k-Nearest Neighbors, and a scalable end-to-end tree boosting system [36]. The SVM kernel was fixed to RBF. For performance metrics of models, we calculated the total accuracy (AC), sensitivity (SN), specificity (SP) and the balanced classification rate (BCR) as follows:

$$AC = \frac{\text{Number of correctly classified compounds}}{\text{Total number of compounds}} \quad (1)$$

$$SN = \frac{\text{Number of correctly classified active compounds}}{\text{Total number active compounds}} \quad (2)$$

$$SP = \frac{\text{Number of correctly classified inactive compounds}}{\text{Total number inactive compounds}} \quad (3)$$

$$BCR = \frac{SN + SP}{2} \times (1 - |SN - SP|) \quad (4)$$

4.3. Multi-Objective Model Assembly and Virtual Screening

The construction of the multi-objective model was performed by computing the global desirability as:

$$D_1 = (d(y_1)d(y_2), \dots, d(y_k))^{1/k} \quad (5)$$

where y_k corresponds with the desirability scores of each cell line ($k = 1, \dots, 4$). For each of the cell lines, several possible models are available. The resulting prediction of each model for a given compound resulted in a score linked to the class membership- either a prediction score for an active class, and/or score for an inactive class. In all cases, we established the score for which the compound was active against the cell line. However, the calculated class membership for some machine learning algorithm occurred in the range of positives and negatives, where active compounds showed positive values and vice versa. For all the cases, the geometrical mean of all scores of the compound to be active in a particular cell line (or to be 0–1, normalized transformation) was used as a desirability score for each model (y_k). Since there are several possible combinations, we performed an exhaustive exploration to obtain the best possible model. Hence, we explored the combination of all possible models in the computation of each $d(y_k)$ and consequently D_1 in order to obtain the best performance in early recognition metrics for virtual screening.

For VS, we developed a dataset with those antitumor compounds used in the current management of osteosarcoma, not included in either the training or external sets for any cell line and with compounds validated in clinical studies for OS, published on the US government's Clinical Trials website (<https://clinicaltrials.gov/>). As first- and second-line therapy drugs, we included [4,12,71,72]: doxorubicin (ChEMBL53463), methotrexate (ChEMBL34259), ifosfamide (ChEMBL1024), etoposide (ChEMBL44657), sorafenib (ChEMBL1336), cyclophosphamide (ChEMBL88), docetaxel (ChEMBL92), gemcitabine (ChEMBL888), dactinomycin (ChEMBL1554) and vincristine (ChEMBL90555). Additionally, we incorporated as validated drugs in clinical trials: temsirolimus (ChEMBL1201182) [73,74], ridaforolimus (ChEMBL2103839) [75,76], sirolimus (ChEMBL413) [77] and pazopanib (ChEMBL477772) [78,79].

As inactive compounds for screening, we considered those molecules withdrawn in the data balancing process (previously described, thus not employed for model training and selection), and common ChEMBL compounds for the four cell lines that showed no biological activity (standard values > 10 μ M). Additionally, we generated Decoy molecules based on the selected active compounds by employing the DUD-E server 5 [80]. We incorporated around 50 inactive molecules for each active compound, which is the proportion used in the DUD-E database that is widely employed to validate virtual screening workflows [80].

The performance of our models within this VS scenario was evaluated by computing the AUAC, BEDROC and EF [81,82]:

$$AUC = 1 - \frac{1}{n} \sum_{i=1}^n x_i \quad (6)$$

$$EF = \frac{\sum_{i=1}^n \delta_i}{\chi^n}, \text{ where } \delta_i = \begin{cases} 1 & r_i \leq \chi^N \\ 0 & r_i > \chi^N \end{cases} \quad (7)$$

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}}, \text{ with } RIE_{min} = \frac{1 - e^{-\alpha R_a}}{R_a(1 - e^{-\alpha})}, RIE_{max} = \frac{1 - e^{-\alpha R_a}}{R_a(1 - e^{-\alpha})} \text{ and} \quad (8)$$

$$RIE = \frac{\frac{1}{n} \sum_{i=1}^n e^{-\alpha x_i}}{\frac{1}{N} \left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1} \right)}$$

In the above equations, n represents the number of active compounds, N the total number of compounds in the dataset, x_i the relative ranking of active compound i in the ranked list, χ the fraction of data for which EF will be computed, R_a the rate of active compounds in the dataset (n/N), and α is the α parameter which ensures that active compounds ranked at the beginning of the ordered list result in higher weights than those at the tail. The α parameter is computed using the following equation:

$$\theta(1 - e^{-\alpha}) - 1 + e^{-\alpha z} = 0 \quad (9)$$

where z represents the fraction of the ranked list at which enrichment is important and θ is the expected contribution of the enrichment at this $z\%$ fraction to the overall enrichment.

5. Conclusions

In conclusion, this study presents a multi-objective prediction algorithm developed from compounds described with biological activity for the osteosarcoma cell lines HOS, MG63, SAOS2 and U2OS. The performance of this multi-objective model considerably improves the recognition rate in a virtual screening scenario, developed on drugs used as first- and second-line treatment for OS. Specifically, a high level of performance was observed for the recognition of molecules with biological activity within 1%. Using this ML algorithm on 2218 compounds described in the DrugBank, we found several antineoplastic agents currently being studied in clinical trials for the treatment of OS. Interestingly, Cabazitaxel is a compound with chemotherapeutic activity that is being studied in several clinical trials for different types of carcinomas and not in sarcomas, therefore it can be taken into account for clinical validations in patients with OS. Furthermore, several broad-spectrum antibiotics, for instance clarithromycin, erythromycin and doxycycline, were top-ranked drugs in our screening. These compounds have already been studied in various types of carcinomas, so they comprise an interesting group of drugs for developing therapeutic validation studies in bone cancers. One of the main limitations was the lack of experimental validation of the drugs proposed for repositioning because this was an initial study. Although it is true that several compounds have already been studied in clinical trials, the validation process of their biological activities on bone tumor cells is an indispensable step in order to proceed with a validation strategy in patients, hence this will be the next procedure in further studies.

Supplementary Materials: The following are available online at <http://www.mdpi.com/1424-8247/13/11/409/s1>. Table S1. ChEMBL dataset of compounds with biological activity for HOS, MG63, U2OS and SAOS2 cell lines. Table S2. ISIDA Molecular descriptors calculated for each cell line's dataset. Table S3. Performance on machine learning models for HOS, MG63, SAOS2 and U2OS cell lines. Table S4. Ranking and desirability values of drugs repositioned by the multi-objective model. Table S5. Clinical trials reported for the 22 top-ranked compounds repositioned by the multi-objective model.

Author Contributions: E.T., A.C.-A. and Y.P.-C. conceived the project and wrote the manuscript. E.T. and Y.P.-C. designed the algorithms. A.C.-A. and A.L.-C. implemented the algorithm and performed the data analysis. G.J.-K. made substantial contributions to the discussion section. C.R.M., H.G.-D. and A.P. helped with study design and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad de Las Américas, Quito, Ecuador, grant number ENFRCA.18.01, by Ministry of Competitiveness and Economy (CTQ2016-74881-P), Ministry of Science and Innovation (PID2019-104148GB-I00), and Basque Government (IT1045-16)-2016–2021.

Acknowledgments: This work was supported by Universidad de Las Américas (Quito, Ecuador) and the Competitive Reference Groups (Ref. ED431C 2018/49) in Galicia, Spain.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ottaviani, G.; Jaffe, N. The epidemiology of osteosarcoma. *Cancer Treat. Res.* **2009**, *152*, 3–13.
2. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology (NCC Guidelines)—Bone Cancer. Available online: <https://www.nccn.org/> (accessed on 4 March 2020).
3. Misaghi, A.; Goldin, A.; Awad, M.; Kulidjian, A.A. Osteosarcoma: A comprehensive review. *SICOT J.* **2018**, *4*, 12. [[CrossRef](#)]
4. Biermann, J.S.; Chow, W.; Reed, D.R.; Lucas, D.; Adkins, D.R.; Agulnik, M.; Benjamin, R.S.; Brigman, B.; Budd, G.T.; Curry, W.T.; et al. NCCN Guidelines Insights: Bone Cancer, Version 2.2017. *J. Natl. Compr. Cancer Netw.* **2017**, *15*, 155–167. [[CrossRef](#)]

5. de Azevedo, J.W.V.; de Medeiros Fernandes, T.A.A.; Fernandes, J.V., Jr.; de Azevedo, J.C.V.; Lanza, D.C.F.; Bezerra, C.M.; Andrade, V.S.; de Araujo, J.M.G.; Fernandes, J.V. Biology and pathogenesis of human osteosarcoma. *Oncol. Lett.* **2020**, *19*, 1099–1116.
6. Xin, S.; Wei, G. Prognostic factors in osteosarcoma: A study level meta-analysis and systematic review of current practice. *J. Bone Oncol.* **2020**, *21*, 100281. [[CrossRef](#)]
7. Marko, T.A.; Diessner, B.J.; Spector, L.G. Prevalence of Metastasis at Diagnosis of Osteosarcoma: An International Comparison. *Pediatr. Blood Cancer* **2016**, *63*, 1006–1011. [[CrossRef](#)]
8. Duchman, K.R.; Gao, Y.; Miller, B.J. Prognostic factors for survival in patients with high-grade osteosarcoma using the Surveillance, Epidemiology, and End Results (SEER) Program database. *Cancer Epidemiol.* **2015**, *39*, 593–599. [[CrossRef](#)]
9. Song, K.; Song, J.; Lin, K.; Chen, F.; Ma, X.; Jiang, J.; Li, F. Survival analysis of patients with metastatic osteosarcoma: A Surveillance, Epidemiology, and End Results population-based study. *Int. Orthop.* **2019**, *43*, 1983–1991. [[CrossRef](#)]
10. Taran, S.J.; Taran, R.; Malipatil, N.B. Pediatric Osteosarcoma: An Updated Review. *Indian J. Med. Paediatr. Oncol.* **2017**, *38*, 33–43. [[CrossRef](#)]
11. Vos, H.I.; Coenen, M.J.; Guchelaar, H.J.; Te Loo, D.M. The role of pharmacogenetics in the treatment of osteosarcoma. *Drug Discov. Today* **2016**, *21*, 1775–1786. [[CrossRef](#)]
12. Durfee, R.A.; Mohammed, M.; Luu, H.H. Review of Osteosarcoma and Current Management. *Rheumatol. Ther.* **2016**, *3*, 221–243. [[CrossRef](#)]
13. Omer, N.; Le Deley, M.C.; Piperno-Neumann, S.; Marec-Berard, P.; Italiano, A.; Corradini, N.; Bellera, C.; Brugieres, L.; Gaspar, N. Phase-II trials in osteosarcoma recurrences: A systematic review of past experience. *Eur. J. Cancer* **2017**, *75*, 98–108. [[CrossRef](#)]
14. Harrison, D.J.; Geller, D.S.; Gill, J.D.; Lewis, V.O.; Gorlick, R. Current and future therapeutic approaches for osteosarcoma. *Expert Rev. Anticancer Ther.* **2018**, *18*, 39–50. [[CrossRef](#)]
15. Lo, Y.C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546. [[CrossRef](#)]
16. Brown, A.S.; Patel, C.J. A review of validation strategies for computational drug repositioning. *Brief. Bioinform.* **2018**, *19*, 174–177. [[CrossRef](#)]
17. Pushpakom, S.; Iorio, F.; Eyers, P.A.; Escott, K.J.; Hopper, S.; Wells, A.; Doig, A.; Williams, T.; Latimer, J.; McNamee, C.; et al. Drug repurposing: Progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **2019**, *18*, 41–58. [[CrossRef](#)]
18. Ma, X.H.; Shi, Z.; Tan, C.; Jiang, Y.; Go, M.L.; Low, B.C.; Chen, Y.Z. In-silico approaches to multi-target drug discovery: Computer aided multi-target drug design, multi-target virtual screening. *Pharm. Res.* **2010**, *27*, 739–749. [[CrossRef](#)]
19. Huang, G.; Li, J.; Wang, P.; Li, W. A Review of Computational Drug Repositioning Approaches. *Comb. Chem. High Throughput Screen* **2017**, *20*, 831–838. [[CrossRef](#)]
20. Park, K. A review of computational drug repurposing. *Transl. Clin. Pharmacol.* **2019**, *27*, 59–63. [[CrossRef](#)]
21. Cruz-Monteagudo, M.; Schurer, S.; Tejera, E.; Perez-Castillo, Y.; Medina-Franco, J.L.; Sanchez-Rodriguez, A.; Borges, F. Systemic QSAR and phenotypic virtual screening: Chasing butterflies in drug discovery. *Drug Discov. Today* **2017**, *22*, 994–1007. [[CrossRef](#)]
22. Murphy, R.F. An active role for machine learning in drug development. *Nat. Chem. Biol.* **2011**, *7*, 327–330. [[CrossRef](#)]
23. Ramsay, R.R.; Popovic-Nikolic, M.R.; Nikolic, K.; Uliassi, E.; Bolognesi, M.L. A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* **2018**, *7*, 3. [[CrossRef](#)]
24. Nagamalla, L.; Kumar, J.V.S. In silico screening of FDA approved drugs on AXL kinase and validation for breast cancer cell line. *J. Biomol. Struct. Dyn.* **2020**. [[CrossRef](#)]
25. Kumar, R.; Chaudhary, K.; Singla, D.; Gautam, A.; Raghava, G.P. Designing of promiscuous inhibitors against pancreatic cancer cell lines. *Sci. Rep.* **2014**, *4*, 4668. [[CrossRef](#)]
26. Issa, N.T.; Stathias, V.; Schurer, S.; Dakshanamurthy, S. Machine and deep learning approaches for cancer drug repurposing. *Semin. Cancer Biol.* **2020**. [[CrossRef](#)]
27. Koudijs, K.K.M.; Terwisscha van Scheltinga, A.G.T.; Bohringer, S.; Schimmel, K.J.M.; Guchelaar, H.J. Personalised drug repositioning for Clear Cell Renal Cell Carcinoma using gene expression. *Sci. Rep.* **2018**, *8*, 5250. [[CrossRef](#)]

28. Wei, G.G.; Gao, L.; Tang, Z.Y.; Lin, P.; Liang, L.B.; Zeng, J.J.; Chen, G.; Zhang, L.C. Drug repositioning in head and neck squamous cell carcinoma: An integrated pathway analysis based on connectivity map and differential gene expression. *Pathol. Res. Pract.* **2019**, *215*, 152378. [[CrossRef](#)]
29. Singh, H.; Kumar, R.; Singh, S.; Chaudhary, K.; Gautam, A.; Raghava, G.P. Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. *BMC Cancer* **2016**, *16*, 77. [[CrossRef](#)]
30. Speck-Planche, A.; Cordeiro, M. Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.* **2017**, *21*, 511–523. [[CrossRef](#)]
31. Bediaga, H.; Arrasate, S.; Gonzalez-Diaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb. Sci* **2018**, *20*, 621–632. [[CrossRef](#)]
32. Lopez-Cortes, A.; Paz, Y.M.C.; Guerrero, S.; Cabrera-Andrade, A.; Barigye, S.J.; Munteanu, C.R.; Gonzalez-Diaz, H.; Pazos, A.; Perez-Castillo, Y.; Tejera, E. OncoOmics approaches to reveal essential genes in breast cancer: A panoramic view from pathogenesis to precision medicine. *Sci. Rep.* **2020**, *10*, 5285. [[CrossRef](#)]
33. Li, X.; Yan, M.L.; Yu, Q. Identification of candidate drugs for the treatment of metastatic osteosarcoma through a subpathway analysis method. *Oncol. Lett.* **2017**, *13*, 4378–4384. [[CrossRef](#)]
34. Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N. Fragment-based QSAR model toward the selection of versatile anti-sarcoma leads. *Eur. J. Med. Chem.* **2011**, *46*, 5910–5916. [[CrossRef](#)]
35. Chalise, P.; Koestler, D.C.; Bimali, M.; Yu, Q.; Fridley, B.L. Integrative clustering methods for high-dimensional molecular data. *Transl. Cancer Res.* **2014**, *3*, 202–216.
36. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
37. Braga, R.C.; Alves, V.M.; Silva, A.C.; Nascimento, M.N.; Silva, F.C.; Liao, L.M.; Andrade, C.H. Virtual screening strategies in medicinal chemistry: The state of the art and current challenges. *Curr. Top. Med. Chem.* **2014**, *14*, 1899–1912. [[CrossRef](#)]
38. Bolognesi, M.L.; Cavalli, A. Multitarget Drug Discovery and Polypharmacology. *Chem. Med. Chem.* **2016**, *11*, 1190–1192. [[CrossRef](#)]
39. Perez-Castillo, Y.; Sanchez-Rodriguez, A.; Tejera, E.; Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M.; Le-Thi-Thu, H.; Pham-The, H. A desirability-based multi objective approach for the virtual screening discovery of broad-spectrum anti-gastric cancer agents. *PLoS ONE* **2018**, *13*, e0192176. [[CrossRef](#)]
40. Xue, H.; Li, J.; Xie, H.; Wang, Y. Review of Drug Repositioning Approaches and Resources. *Int. J. Biol Sci* **2018**, *14*, 1232–1244. [[CrossRef](#)]
41. Langedijk, J.; Mantel-Teeuwisse, A.K.; Slijkerman, D.S.; Schutjens, M.H. Drug repositioning and repurposing: Terminology and definitions in literature. *Drug Discov. Today* **2015**, *20*, 1027–1034. [[CrossRef](#)]
42. Parvathaneni, V.; Kulkarni, N.S.; Muth, A.; Gupta, V. Drug repurposing: A promising tool to accelerate the drug discovery process. *Drug Discov. Today* **2019**, *24*, 2076–2085. [[CrossRef](#)]
43. Ding, L.; Congwei, L.; Bei, Q.; Tao, Y.; Ruiguang, W.; Heze, Y.; Bo, D.; Zhihong, L. mTOR: An attractive therapeutic target for osteosarcoma? *Oncotarget* **2016**, *7*, 50805–50813. [[CrossRef](#)]
44. Bishop, M.W.; Janeway, K.A. Emerging concepts for PI3K/mTOR inhibition as a potential treatment for osteosarcoma. *F1000Research* **2016**, *5*, 1590. [[CrossRef](#)]
45. Cabrera-Andrade, A.; Lopez-Cortes, A.; Jaramillo-Koupermann, G.; Paz, Y.M.C.; Perez-Castillo, Y.; Munteanu, C.R.; Gonzalez-Diaz, H.; Pazos, A.; Tejera, E. Gene Prioritization through Consensus Strategy, Enrichment Methodologies Analysis, and Networking for Osteosarcoma Pathogenesis. *Int. J. Mol. Sci* **2020**, *21*, 1053. [[CrossRef](#)]
46. de Bono, J.S.; Oudard, S.; Ozguroglu, M.; Hansen, S.; Machiels, J.P.; Kocak, I.; Gravis, G.; Bodrogi, I.; Mackenzie, M.J.; Shen, L.; et al. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: A randomised open-label trial. *Lancet* **2010**, *376*, 1147–1154. [[CrossRef](#)]
47. de Wit, R.; de Bono, J.; Sternberg, C.N.; Fizazi, K.; Tombal, B.; Wulfing, C.; Kramer, G.; Eymard, J.C.; Bamias, A.; Carles, J.; et al. Cabazitaxel versus Abiraterone or Enzalutamide in Metastatic Prostate Cancer. *N. Engl. J. Med.* **2019**, *381*, 2506–2518. [[CrossRef](#)]

48. Oudard, S.; Fizazi, K.; Sengelov, L.; Daugaard, G.; Saad, F.; Hansen, S.; Hjalmer-Eriksson, M.; Jassem, J.; Thiery-Vuillemin, A.; Caffo, O.; et al. Cabazitaxel Versus Docetaxel As First-Line Therapy for Patients With Metastatic Castration-Resistant Prostate Cancer: A Randomized Phase III Trial-FIRSTANA. *J. Clin. Oncol.* **2017**, *35*, 3189–3197. [CrossRef]
49. Lo, Y.C.; Senese, S.; France, B.; Gholkar, A.A.; Damoiseaux, R.; Torres, J.Z. Computational Cell Cycle Profiling of Cancer Cells for Prioritizing FDA-Approved Drugs with Repurposing Potential. *Sci. Rep.* **2017**, *7*, 11261. [CrossRef]
50. Reynolds, C.P.; Kang, M.H.; Maris, J.M.; Kolb, E.A.; Gorlick, R.; Wu, J.; Kurmasheva, R.T.; Houghton, P.J.; Smith, M.A. Initial testing (stage 1) of the anti-microtubule agents cabazitaxel and docetaxel, by the pediatric preclinical testing program. *Pediatr. Blood Cancer* **2015**, *62*, 1897–1905. [CrossRef]
51. Amoroso, L.; Castel, V.; Bisogno, G.; Casanova, M.; Marquez-Vega, C.; Chisholm, J.C.; Doz, F.; Moreno, L.; Ruggiero, A.; Gerber, N.U.; et al. Phase II results from a phase I/II study to assess the safety and efficacy of weekly nab-paclitaxel in paediatric patients with recurrent or refractory solid tumours: A collaboration with the European Innovative Therapies for Children with Cancer Network. *Eur. J. Cancer* **2020**, *135*, 89–97. [CrossRef]
52. Hussain, A.; Dar, M.S.; Bano, N.; Hossain, M.M.; Basit, R.; Bhat, A.Q.; Aga, M.A.; Ali, S.; Hassan, Q.P.; Dar, M.J. Identification of dinactin, a macrolide antibiotic, as a natural product-based small molecule targeting Wnt/beta-catenin signaling pathway in cancer cells. *Cancer Chemother. Pharmacol.* **2019**, *84*, 551–559. [CrossRef]
53. Gupta, A.; Okesli-Armlovich, A.; Morgens, D.; Bassik, M.C.; Khosla, C. A genome-wide analysis of targets of macrolide antibiotics in mammalian cells. *J. Biol. Chem.* **2020**, *295*, 2057–2067. [CrossRef]
54. Bahrami, F.; Morris, D.L.; Pourgholami, M.H. Tetracyclines: Drugs with huge therapeutic potential. *Mini Rev. Med. Chem* **2012**, *12*, 44–52. [CrossRef]
55. Fiorillo, M.; Toth, F.; Sotgia, F.; Lisanti, M.P. Doxycycline, Azithromycin and Vitamin C (DAV): A potent combination therapy for targeting mitochondria and eradicating cancer stem cells (CSCs). *Aging* **2019**, *11*, 2202–2216. [CrossRef]
56. Lamb, R.; Ozsvari, B.; Lisanti, C.L.; Tanowitz, H.B.; Howell, A.; Martinez-Outschoorn, U.E.; Sotgia, F.; Lisanti, M.P. Antibiotics that target mitochondria effectively eradicate cancer stem cells, across multiple tumor types: Treating cancer like an infectious disease. *Oncotarget* **2015**, *6*, 4569–4584. [CrossRef]
57. Maksimovic-Ivanic, D.; Fagone, P.; McCubrey, J.; Bendtzen, K.; Mijatovic, S.; Nicoletti, F. HIV-protease inhibitors for the treatment of cancer: Repositioning HIV protease inhibitors while developing more potent NO-hybridized derivatives? *Int. J. Cancer* **2017**, *140*, 1713–1726. [CrossRef]
58. Petroni, G.; Stefanini, M.; Pillozzi, S.; Crociani, O.; Becchetti, A.; Arcangeli, A. Data describing the effects of the Macrolide Antibiotic Clarithromycin on preclinical mouse models of Colorectal Cancer. *Data Brief.* **2019**, *26*, 104406. [CrossRef]
59. Van Nuffel, A.M.; Sukhatme, V.; Pantziarka, P.; Meheus, L.; Sukhatme, V.P.; Bouche, G. Repurposing Drugs in Oncology (ReDO)-clarithromycin as an anti-cancer agent. *Ecancermedicalscience* **2015**, *9*, 513. [CrossRef]
60. de Jong, J.; Hellemans, P.; De Wilde, S.; Patricia, D.; Masterson, T.; Manikhas, G.; Myasnikov, A.; Osmanov, D.; Cordoba, R.; Panizo, C.; et al. A drug-drug interaction study of ibrutinib with moderate/strong CYP3A inhibitors in patients with B-cell malignancies. *Leuk. Lymphoma* **2018**, *59*, 2888–2895. [CrossRef]
61. Markowska, A.; Kaysiewicz, J.; Markowska, J.; Huczynski, A. Doxycycline, salinomycin, monensin and ivermectin repositioned as cancer drugs. *Bioorg. Med. Chem. Lett.* **2019**, *29*, 1549–1554. [CrossRef]
62. Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; De Veij, M.; Félix, E.; Magarinos, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940. [CrossRef]
63. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrian-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. [CrossRef]
64. Chem Axon J. Chem for Office. Available online: <https://chemaxon.com> (accessed on 12 March 2020).
65. Chem Axon Chemaxon Standardizer. Available online: <http://www.chemaxon.com> (accessed on 24 March 2020).
66. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Hoonakker, F.; Tetko, I.V.; Marcou, G. ISIDA-Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput. Aided Drug Des.* **2008**, *4*, 191. [CrossRef]


67. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010**, *29*, 855–868. [[CrossRef](#)]
68. Varnek, A.; Fourches, D.; Hoonakker, F.; Solovev, V.P. Substructural fragments: An universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided Mol. Des.* **2005**, *19*, 693–703. [[CrossRef](#)]
69. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [[CrossRef](#)]
70. Potter, T.; Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **1998**, *41*, 478–488. [[CrossRef](#)]
71. Serra, M.; Hattinger, C.M. The pharmacogenomics of osteosarcoma. *Pharmacogenom. J.* **2017**, *17*, 11–20. [[CrossRef](#)]
72. Hattinger, C.M.; Vella, S.; Tavanti, E.; Fanelli, M.; Picci, P.; Serra, M. Pharmacogenomics of second-line drugs used for treatment of unresponsive or relapsed osteosarcoma patients. *Pharmacogenomics* **2016**, *17*, 2097–2114. [[CrossRef](#)]
73. Schwartz, G.K.; Tap, W.D.; Qin, L.X.; Livingston, M.B.; Undevia, S.D.; Chmielowski, B.; Agulnik, M.; Schuetze, S.M.; Reed, D.R.; Okuno, S.H.; et al. Cixutumumab and temsirolimus for patients with bone and soft-tissue sarcoma: A multicentre, open-label, phase 2 trial. *Lancet Oncol.* **2013**, *14*, 371–382. [[CrossRef](#)]
74. Trucco, M.M.; Meyer, C.F.; Thornton, K.A.; Shah, P.; Chen, A.R.; Wilky, B.A.; Carrera-Haro, M.A.; Boyer, L.C.; Ferreira, M.F.; Shafique, U.; et al. A phase II study of temsirolimus and liposomal doxorubicin for patients with recurrent and refractory bone and soft tissue sarcomas. *Clin. Sarcoma Res.* **2018**, *8*, 21. [[CrossRef](#)]
75. Demetri, G.D.; Chawla, S.P.; Ray-Coquard, I.; Le Cesne, A.; Staddon, A.P.; Milhem, M.M.; Penel, N.; Riedel, R.F.; Bui-Nguyen, B.; Cranmer, L.D.; et al. Results of an international randomized phase III trial of the mammalian target of rapamycin inhibitor ridaforolimus versus placebo to control metastatic sarcomas in patients after benefit from prior chemotherapy. *J. Clin. Oncol.* **2013**, *31*, 2485–2492. [[CrossRef](#)]
76. Chawla, S.P.; Staddon, A.P.; Baker, L.H.; Schuetze, S.M.; Tolcher, A.W.; D’Amato, G.Z.; Blay, J.Y.; Mita, M.M.; Sankhala, K.K.; Berk, L.; et al. Phase II study of the mammalian target of rapamycin inhibitor ridaforolimus in patients with advanced bone and soft tissue sarcomas. *J. Clin. Oncol.* **2012**, *30*, 78–84. [[CrossRef](#)]
77. Qayed, M.; Cash, T.; Tighiouart, M.; MacDonald, T.J.; Goldsmith, K.C.; Tanos, R.; Kean, L.; Watkins, B.; Suessmuth, Y.; Wetmore, C.; et al. A phase I study of sirolimus in combination with metronomic therapy (CHOAnome) in children with recurrent or refractory solid and brain tumors. *Pediatr. Blood Cancer* **2020**, *67*, e28134. [[CrossRef](#)]
78. van der Graaf, W.T.; Blay, J.Y.; Chawla, S.P.; Kim, D.W.; Bui-Nguyen, B.; Casali, P.G.; Schoffski, P.; Aglietta, M.; Staddon, A.P.; Beppu, Y.; et al. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): A randomised, double-blind, placebo-controlled phase 3 trial. *Lancet* **2012**, *379*, 1879–1886. [[CrossRef](#)]
79. Longhi, A.; Paioli, A.; Palmerini, E.; Cesari, M.; Abate, M.E.; Setola, E.; Spinnato, P.; Donati, D.; Hompland, I.; Boye, K. Pazopanib in relapsed osteosarcoma patients: Report on 15 cases. *Acta Oncol.* **2019**, *58*, 124–128. [[CrossRef](#)]
80. Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594. [[CrossRef](#)]
81. Truchon, J.F.; Bayly, C.I. Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508. [[CrossRef](#)]
82. Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *J. Comput. Aided Mol. Des.* **2008**, *22*, 213–228. [[CrossRef](#)]

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

SCIENTIFIC REPORTS



OPEN

Gene prioritization, communality analysis, networking and metabolic integrated pathway to better understand breast cancer pathogenesis

Andrés López-Cortés^{1,2}, César Paz-y-Miño¹, Alejandro Cabrera-Andrade^{3,4}, Stephen J. Barigye⁵, Cristian R. Munteanu^{2,6}, Humberto González-Díaz^{7,8}, Alejandro Pazos^{2,6}, Yunierkis Pérez-Castillo^{4,9} & Eduardo Tejera^{4,10}

Consensus strategy was proved to be highly efficient in the recognition of gene-disease association. Therefore, the main objective of this study was to apply theoretical approaches to explore genes and communities directly involved in breast cancer (BC) pathogenesis. We evaluated the consensus between 8 prioritization strategies for the early recognition of pathogenic genes. A communality analysis in the protein-protein interaction (PPI) network of previously selected genes was enriched with gene ontology, metabolic pathways, as well as oncogenomics validation with the OncoPPI and DRIVE projects. The consensus genes were rationally filtered to 1842 genes. The communality analysis showed an enrichment of 14 communities specially connected with ERBB, PI3K-AKT, mTOR, FOXO, p53, HIF-1, VEGF, MAPK and prolactin signaling pathways. Genes with highest ranking were TP53, ESRI, BRCA2, BRCA1 and ERBB2. Genes with highest connectivity degree were TP53, AKT1, SRC, CREBBP and EP300. The connectivity degree allowed to establish a significant correlation between the OncoPPI network and our BC integrated network conformed by 51 genes and 62 PPI. In addition, CCND1, RAD51, CDC42, YAP1 and RPA1 were functional genes with significant sensitivity score in BC cell lines. In conclusion, the consensus strategy identifies both well-known pathogenic genes and prioritized genes that need to be further explored.

BC is a complex and heterogeneous disease. This pathology represents a significant health problem and is characterized by an intricate interplay between different biological aspects such as environmental determinants, signaling pathway alterations, metabolic abnormalities, hormone disruption, gene expression deregulation, DNA genomics alterations and ethnicity^{1,2}.

The heterogeneity of BC can be observed at molecular, histological and functional levels, all of which have clinical implications³. The 95% of mammary tumors are adenocarcinomas. The *in situ* carcinoma is classified into ductal carcinoma *in situ* and lobular carcinoma *in situ*⁴. On the other hand, the malignant cells of the infiltrating

¹Centro de Investigación Genética y Genómica, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Mariscal Sucre Avenue, 170129, Quito, Ecuador. ²RNASA-IMEDIR, Computer Sciences Faculty, University of Coruna, 15071, Coruna, Spain. ³Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de las Américas, Avenue de los Granados, 170125, Quito, Ecuador. ⁴Grupo de Bio-Quimioinformática, Universidad de las Américas, Avenue de los Granados, 170125, Quito, Ecuador. ⁵Department of Chemistry, McGill University, 801 Sherbrooke Street West, Montreal, QC, H3A 0B8, Canada. ⁶INIBIC, Institute of Biomedical Research, CHUAC, UDC, 15006, Coruna, Spain. ⁷Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Biscay, Spain. ⁸IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Biscay, Spain. ⁹Escuela de Ciencias Físicas y Matemáticas, Universidad de las Américas, Avenue de los Granados, 170125, Quito, Ecuador. ¹⁰Facultad de Ingeniería y Ciencias Agropecuarias, Universidad de las Américas, Avenue de los Granados, 170125, Quito, Ecuador. Correspondence and requests for materials should be addressed to A.L.-C. (email: aalc84@gmail.com) or E.T. (email: eduardo.tejera@udla.edu.ec)

ductal carcinoma are classified as lobular, tubular, medullary, papillary and metaplastic⁵. However, the histopathologic classification coupled with the molecular subtyping of the estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and the PAM50 mRNA-based assay generate five different intrinsic molecular subtypes: luminal A (ER+ and/or PR+, HER2-, low Ki67), luminal B (ER+ and/or PR+, HER+ or HER- with high Ki67), basal-like (ER-, PR-, HER2-, cytokeratin 5/6+, and/or HER1+), HER2-enriched (ER-, PR-, HER2+) and normal-like^{3,6-9}.

The major BC hallmarks are related to cell proliferation, differentiation and cell apoptosis processes that are associated to the deregulation of the cell cycle and the impairment of DNA repair processes¹⁰. However, the underlying molecular interactions of these processes are to-date not well understood and the corresponding network of the mechanistic interplay and physical interactions between individual genes, proteins and metabolites are unexplored due to the fact that most pathways are complex connected to regulate particular cellular processes¹¹. For this reason, BC genes need to be understood as being part of a complex network¹². In general, genes involved in the BC progression represent a broad class of proteins such as transcription factors, chromatin remodelers, growth factors, growth factor receptors, signal transducers and DNA repair genes¹³. The individual key players of BC progression are classified as oncogenes, tumor suppressor genes and genomic stability genes¹⁴. These genes are playing a key role in the regulation of cell cycle, cell proliferation and cell differentiation¹⁵.

Despite what is known up to date, we still have not a complete, integrative understanding about the association between BC driver genes, networks and metabolic pathways. Hence, the consensus strategy (CS) had proved to be an efficient way to explore gene-disease association^{15,16}. Therefore, we will include several prioritization strategies that will be integrated using a CS in order to rank the genes in the gene-disease association. The consensus result will be integrated in network analysis and metabolic pathway analysis in order to identify relevant pathogenic genes and pathogenic pathways related to BC. The aim of this study is to apply several theoretical approaches to explore BC, specially those genes directly involved in the pathogenesis through a multi-objective design.

Methods

Selection of pathogenic genes for validation. The methodology used below is similar to that previously described by Tejera *et al.*¹⁷. The validation strategy for prioritization on pathogenic genes was performed from the identification of specific genes involved in the BC pathogenesis. Through a search in Scopus and PubMed databases, a gene was considered as pathogenic if: (1) the silencing or induced overexpression of the proposed gene in organism models generate a clinical phenotype like BC (Group G1), and (2) at least one polymorphism was associated with BC in meta-analysis studies (Group G2)^{17,18}.

The full gene list of G1 (n = 59) and G2 (n = 101) can be found in Tables S1 and S2, respectively. While the 145 unique genes combining G1 + G2 and its corresponding Entrez Gene ID identifier can be found in Table S3.

Prioritization algorithms and Consensus strategy. The prioritization methods were selected according to two criteria: (1) full available platform in web service, and (2) requiring only the disease name for gene prioritization. The eight bioinformatics tools that met these criteria were Glad4U¹⁹, DisgeNet²⁰, Génie²¹, SNPs3D²², Guildify²³, CIPHER²⁴, Phenolyzer²⁵ and Polysearch²⁶. These prioritization algorithms present several characteristics that have been previously evaluated by several authors^{15,27}. The previously selected prioritization tools were well integrated in the CS¹⁷. Each gene “i” in the ranked list provided by each method “j” was normalized ($GeneN_{i,j}$) which means, the normalized score of the gene “i” in the method “j” in order to integrate all methods for the Consensus approach. For the final score per gene we considered the average normalized score as well as the number of methods that predict the gene “n_i” using:

$$Gene_i = \sqrt{\left(\frac{n_i - 1}{12 - 1}\right) \left(\frac{1}{j} \sum_j GeneN_{i,j}\right)} \quad (1)$$

The equation (1) corresponds with the geometrical mean between the average score of each gene obtained in each method and the normalized score according to the number of methods which predict the gene-disease association¹⁷. The geometrical mean, using the square root, is applied because it is more sensitive to extreme values than the arithmetic mean. Therefore, genes are ordered according to the $Gene_i$ values. This sorting will produce a ranking that further normalized leading to the final score of each gene ($ConsenScore_i$). The final list has 19,989 prioritized genes. To reduce this list we used the already predefined pathogenic genes (G1 and G2) and the following equation (2):

$$I_i = \frac{TP_i}{FP_i + 1} ConsenScore_i \quad (2)$$

where TP and FP were the true positive and false positive values (up to the ranking value of the $Gene_i$), respectively. The maximal value of I_i is the maximal compromise between the TP and FP rate compensated with the ranking index of each gene.

Enrichment analysis. Pathway enrichment analysis and gene ontology (GO) were performed using David Bioinformatics Resource^{28,29}. Revigo was used to simplify the high number of genes and GO terms, maintaining it with highest specificity^{30,31}. In addition, RSpider was used to obtain integrated information from the Kyoto Encyclopedia of Genes and Genomes (KEGG)^{32,33}. RSpider will produce statistical analysis of the enrichment and a network representation integrating the information in both databases. This tool connects into non-interrupted sub-network component as many input genes as possible using minimal number of missing genes³².

Protein-protein interaction network analysis. The protein-protein interaction (PPI) network with a highest confidence cutoff of 0.9 and zero node addition was created using the String Database³⁴. The confidence score is the approximate probability that a predicted link exists between two enzymes in the same metabolic map. The String Database takes into account known and predicted interactions³⁴. The centrality indexes calculation and network visualization was analyzed through the Cytoscape software³⁵. The communality network analysis (CNA) was performed by clique percolation method using the CFinder software³⁶. The CNA provides a better topology description of the network overlapping modules that correspond with relevant biological information and including the location of highly connected sub-graphs (k-cliques)¹⁷. The different k-cliques present different number of communities and genes per community. The selection of the k-clique value will define our further analysis. The higher the k-clique value is, the lower the number of communities that integrate it and vice versa. In our network, both extremes (too small or too high k-clique values) generate imbalance in the gene distribution present in each community. In order to minimize this bias, we used “S” index detailed in equation (3)¹⁷, where N_g^k and N_c^k represent the number of genes in each community and the number of communities for a defined k-clique cutoff value:

$$S^k = \frac{|\text{mean}(N_g^k) - \text{median}(N_g^k)|}{N_c^k} \quad (3)$$

In order to provide a weight of the pathways integrating also network information we used the *PathScore_m* defined as¹⁷: if *ConsenScore_i^k* is the *ConsenScore_i* of the gene “i” in the community “k” then: (1) Each community “k” was weighted as: $W_k = \sum \text{ConsenScore}_i^k / N_k$, where N_k is the number of communities. (2) Each pathway “m” was weighted as: $\text{PathRankScore}_m = \sum W_k^m / N_k^m$, where W_k^m is the weight (W_k) of each community connected with the pathway “m” and N_k^m is the number of communities connected with the pathway “m”. (3) A second weight was given to the pathway “m” (*PathGeneScore_m*) considering all the genes involved in the pathway as: $\text{PathGeneScore}_m = \sqrt{(\text{ConsenScore}_i^m)^{n_m} / N_m}$, where “ N_m ” is the total number of genes in the pathway “m” while “ n_m ” is the number of those genes which are also found in the PPI network. *ConsenScore_i^m* is the average of the *ConsenScore_i* of all genes present in the pathway “m”. (4) The final score associated with the pathway “m” (*PathScore_m*) is calculated as the geometrical mean between *PathGeneScore_m* and the normalized *PathRankScore_m*.

K-mean analysis. Once the k-clique cutoff is defined, there are several communities that need also to be rationally reduced. We proposed a K-mean clustering analysis using the following variables: PathScore, average degree and average consensus ranking of the genes in that community. The cluster analysis will lead us to group communities with similar patterns according to predefined variables.

Oncogenomics validation with the OncoPPI BC network and the DRIVE project. OncoPPI reports the generation of a cancer-focused PPI network, and identification of more than 260 high-confidence cancer-associated PPI according to Li *et al.*, and Ivanov *et al.*^{37,38}. In addition, the OncoPPI BC network is confirmed by 94 genes and 170 PPI experimentally analyzed in BC cell lines^{37,38}. The correlation of the degree centrality by means of Spearman p-value test between the OncoPPI BC network and our String PPI network, and between the OncoPPI BC network and our BC integrated network allows validation of all the high-confidence breast cancer-focused PPI analyzed in cell lines and proposed in our study.

On the other hand, the DRIVE project (deep RNAi interrogation of visibility effects in cancer) is the larger-scale gene knockdown experiment to discover functional gene requirements across diverse sets of cancer³⁹. According to McDonald *et al.*, DRIVE constructed deep coverage shRNA lentiviral libraries targeting 7,838 human genes (e.g. druggable enzymes) with a median of 20 shRNAs per gene and used to screen 398 cancer cell lines, including 24–25 BC cell lines, in order to analyze cell viability³⁹. shRNA activity was aggregated to gene-level activity by Redundant siRNA Activity method (RSA). According to König *et al.*, RSA method uses all shRNA reagents against a given gene to calculate a statistical significance that knockdown of gene X leads to loss of viability⁴⁰. Genes with RSA value (sensitivity score) ≤ -3 for >50% of cancer cell lines were deemed essential, genes with RSA ≤ -3 for 1–49% of cancer cell lines were deemed active and genes with RSA ≤ -3 for 0% of cancer cell lines were deemed inert. Regarding our study, we analyzed the sensitivity score of the Consensus genes, the most relevant communities, pathogenic genes, the BC integrated network and the OncoPPI BC network in all cancer cell lines and BC cell lines.

Results

Consensus prioritization. The analyses of pathogenic genes in all bioinformatics tools are presented in Table 1. However, not all methods are able to identify the 145 proposed BC pathogenic genes.

CS is the method with highest identification of pathogenic genes in G1 and G2 datasets at the lower 1% of the data (199 of 19,989 genes). CS identified the 49.2% of G1 set in the initial 1% and almost 80% of G1 and G2 genes in the 5% of the final gene list (29 and 116 genes, respectively) followed by Phenolyzer method²⁵. The identification of the pathogenic genes is important but it is also relevant a low rank for those genes. Therefore, we also included the average rank of the detected genes as presented in Table 2.

The rank of the detected genes using CS is actually not superior to Guildify²³, and it is actually very close to Phenolyzer²⁵. However, considering both criteria recovering and ranking, CS is superior recovering in the first 1% more genes (10% more than Phenolyzer) in the average 50 top genes. Similarly, in the initial 10% of the data (1998 genes) Consensus recovers almost 20% more genes than Phenolyzer and 50% more than Guildify in the average 280 initial genes.

Methods	1%			5%			10%			20%			50%		
	G1	G2	G1 + G2	G1	G2	G1 + G2	G1	G2	G1 + G2	G1	G2	G1 + G2	G1	G2	G1 + G2
GLAD4U	6.8	4.5	3.2	15.3	15.3	12.3	20.3	22.5	19.4	32.2	34.2	30.3	45.8	47.7	43.9
Disgenet	0.0	0.0	0.0	1.7	1.8	1.3	8.5	4.5	3.2	10.2	9.0	6.5	15.3	12.6	9.7
Genie	3.4	1.8	1.3	5.1	2.7	2.6	6.8	4.5	4.5	47.5	27.9	31.0	67.8	55.0	56.1
SNP3D	11.9	8.1	5.8	22.0	26.1	20.6	35.6	37.8	32.9	44.1	54.1	47.7	59.3	65.8	60.6
Guildify	18.6	16.2	14.8	18.6	23.4	20.0	23.7	28.8	25.2	44.1	36.9	38.1	76.3	69.4	70.3
Cipher	3.4	2.7	1.9	5.1	7.2	5.8	13.6	14.4	12.3	20.3	16.2	15.5	25.4	21.6	20.0
Phenolyzer	47.5	29.7	31.6	79.7	55.0	60.6	86.4	71.2	74.2	88.1	85.6	85.2	94.9	98.2	96.8
Polyssearch	0.0	0.0	0.0	1.7	0.9	0.6	1.7	0.9	0.6	3.4	1.8	1.3	5.1	4.5	3.2
Consensus	49.2	42.3	40.6	76.3	84.7	80.0	83.1	98.2	92.3	93.2	100.0	97.4	96.6	100.0	98.7

Table 1. Identification (in %) of pathogenic genes in each approach.

Methods	1%			5%			10%			20%			50%		
	G1	G2	G1 + G2	G1	G2	G1 + G2	G1	G2	G1 + G2	G1	G2	G1 + G2	G1	G2	G1 + G2
GLAD4U	4.2	2.7	1.9	20.3	10.3	8.1	30.6	18.6	14.4	64.5	26.9	27.4	123.6	71.0	53.0
Disgenet	0.0	0.0	0.0	2.5	1.4	0.6	5.1	2.7	1.9	6.3	5.0	3.5	12.4	7.2	5.4
Genie	11.9	6.3	4.5	27.6	8.7	10.3	50.8	51.5	36.1	273.6	146.2	107.4	389.5	247.9	174.0
SNP3D	6.9	4.6	3.3	24.1	17.9	13.6	60.7	34.4	26.7	104.4	63.2	48.5	214.9	108.8	84.6
Guildify	97.8	39.5	31.4	97.8	120.1	78.6	424.7	226.9	169.0	1576.3	551.4	508.5	3531.5	1863.9	1370.9
Cipher	2.5	2.7	1.9	20.8	18.8	14.4	89.4	45.7	33.2	133.2	51.6	43.4	204.7	116.7	81.2
Phenolyzer	95.3	45.7	36.0	355.8	191.4	147.2	441.9	323.7	221.4	461.2	399.3	264.1	532.0	444.7	298.7
Polyssearch	0.0	0.0	0.0	1.7	0.9	0.6	1.7	0.9	0.6	4.2	2.3	1.6	6.3	5.2	3.7
Consensus	91.7	66.9	46.2	372.5	271.3	189.5	510.5	400.2	277.0	989.5	430.4	356.2	1392.2	430.4	413.5

Table 2. Average ranking of identified pathogenic genes in each method.

The number of prioritized genes is really elevated (19,989) and consequently a rational cutoff needs to be applied. The maximal value of I_i is 0.787148315 and corresponds with a ranking value of 1842. Therefore, our final reduced list for BC comprises the first 1842 genes (Fig. 1a). The entire gene list as well as their scores and ranking can be found in Table S4. In the 1842 genes there are 91.5% of predefined pathogenic genes.

Enrichment analysis of breast cancer related genes and protein-protein interaction network. The enrichment analysis of GO terms related to biological processes (BP) and metabolic pathways was carried on in the 1842 genes. The GO enrichment results into more than 300 terms with a false discovery rate (FDR) < 0.01. In order to simplify this list we used Revigo to calculate the GO term frequencies³⁰.

Tables S5 and S6 present a full list of BP in BC genes. We only consider terms with a frequency < 0.05%. The BP that present low frequency are more specific and therefore they give a greater biological meaning⁴¹. Several BP such as ERBB2 signaling pathway, DNA synthesis involved in DNA repair, phosphatidylinositol-3-phosphate biosynthetic process, cellular response to epidermal growth factor stimulus and positive regulation of tyrosine phosphorylation of STAT3 protein are directly associated with the BC pathogenesis⁴²⁻⁴⁴.

The enrichment analysis of the KEGG pathways generated significant association (FDR) between BC and the PI3K-AKT, FOXO, ERBB, RAS, prolactin and MAPK signaling pathways⁴⁵⁻⁵¹. The BP and enriched pathways are consistent between them and also with scientific knowledge about BC (Table S7).

To better understand BC behavior, in addition to the association between BP and enrichment pathways, it was important to supplement information through a network analysis. With the indicated cutoff of 0.9, the final interaction network had 1484 nodes, corresponding with the 80.6% of the initial Consensus genes ($n = 1842$). The best-ranked k-clique was 9 ($S_k = 0.126$) with 49 communities (Fig. 1b and Table S8).

Of the 1484 network nodes, only 496 were part of one of the 49 communities (k-clique 9). The network with 1484 genes presented 124 of the 145 predefined pathogenic genes (86%). The sub-network of 496 genes comprises 63 of 145 (43%) predefined pathogenic genes. In this reduction there is an enrichment of the pathogenic genes considering that hypergeometric probability test (HPT) provides a $p < 0.01$. This means that the number of pathogenic genes in this group is higher than what would be expected at random. On the other hand, the average degree of the pathogenic genes was 37.4 which was statistically significant higher than non pathogenic genes (18.1) at $p < 0.05$. This result indicates that the average degree of the genes in the network could be associated to BC.

The metabolic pathways obtained by previous enrichment analysis is weighted considering the consensus score of the genes involved as well as their participation in the interaction network. The results presented in Table 3 (Table S9) shown that some metabolic pathways are present in several communities while others are poorly represented. Among the most relevant signaling pathways with highest *PathScore* for BC were ERBB, prolactin, mTOR, p53, FOXO, HIF-1, MAPK, PI3K-AKT and VEGF signaling pathways.

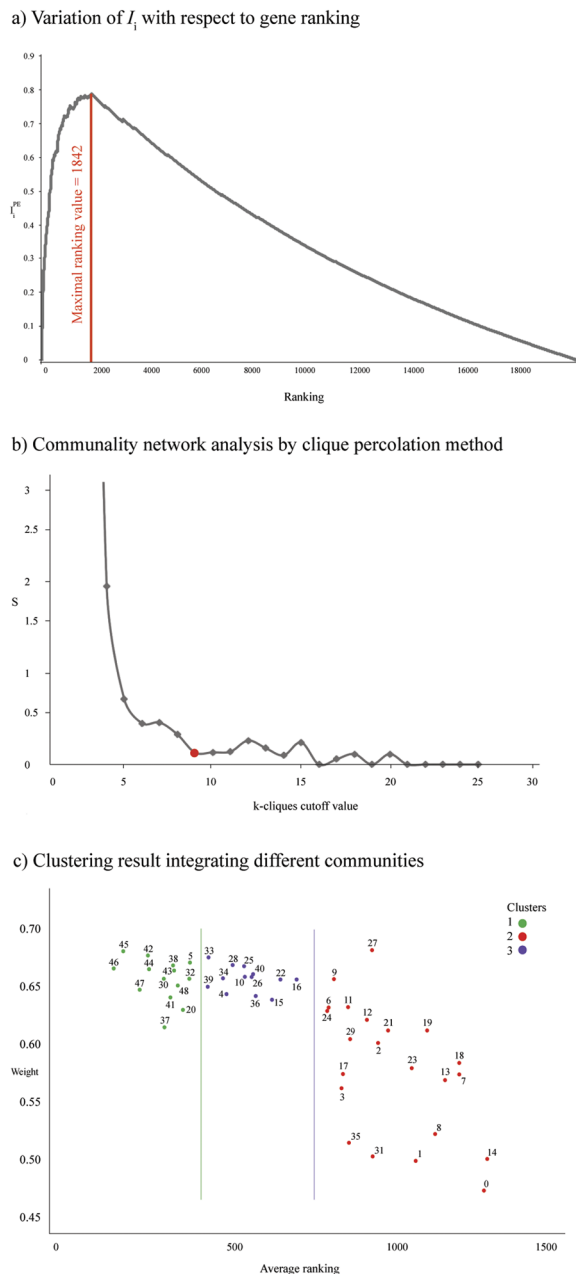


Figure 1. (a) Variation of I_i with respect to genes ranking. The maximal value of I_i is 0.787148315 and corresponds with a ranking value of 1842 genes. (b) Community network analysis by clique percolation method. Values of S^k with respect to each k -clique cutoff value. (c) Clustering result (3 clusters) integrating different communities. Green circles represent cluster 1, blue circles represent cluster 2, and purple circles represent cluster 3. X-axis represents the average ranking of communities and Y-axis represents weight of pathogenic genes.

In order to reduce the 49 communities, which is a relative high number, we considered a K-mean cluster analysis using Euclidian distance with the following variables: average node degree in each community, *ConsenScore*_{*i*} of each gene in the community, and the average *PathScore* in each community. The 14 most relevant communities of cluster 1 were: 46 (0.664), 45 (0.677), 47 (0.646), 42 (0.674), 44 (0.663), 30 (0.655), 37 (0.616), 41 (0.640), 43 (0.662), 38 (0.666), 48 (0.649), 32 (0.655), 5 (0.668) and 20 (0.630). These communities could comprise the most relevant BC genes and pathways (Fig. 1c).

Table 4 details genes that make up the main communities and the HPT p-values (Table S10). HPT evaluates the relevance of the pathogenic genes in the communities. The top 20 genes with highest connectivity degree were TP53, AKT1, SRC, CREBBP, EP300, JUN, CTNBN1, RAC1, PIK3CA, EGFR, MAPK8, MAPK1, STAT3, ESR1, MAPK14, CCND1, GRB2, CDK2, FOS and CDKN1A. In addition, 19 of these 20 genes were found in the 14 most relevant communities. The sub-network of genes comprised in the 14 communities is presented in Figs 2, S1(a) and S1(b).

Pathways	PathRank	N Community	PathGene	PathScore	Community
ERBB signaling pathway	0.815143	14	0.715853953	0.763886926	4 25 26 33 34 36 38 40 42 43 44 46 47 48
Prolactin signaling pathway	0.795867	15	0.72857406	0.761477386	4 6 11 33 34 36 38 39 40 42 43 44 46 47 48
mTOR signaling pathway	0.815500	4	0.687676019	0.748865671	4 36 42 44
p53 signaling pathway	0.735875	8	0.735254081	0.735564475	4 9 10 12 16 30 32 42
FOXO signaling pathway	0.787647	17	0.683991499	0.733991752	4 5 6 11 12 22 34 36 38 39 42 43 44 45 46 47 48
HIF-1 signaling pathway	0.796182	11	0.673983105	0.7325388	2 4 5 22 34 36 38 41 42 45 46
VEGF signaling pathway	0.799750	16	0.663653015	0.728530369	4 6 11 25 26 33 34 36 38 42 43 44 45 46 47 48
Homologous recombination	0.689800	5	0.744804648	0.716774892	9 24 27 30 32
Thyroid hormone signaling pathway	0.801071	14	0.626992865	0.708707323	4 5 10 20 28 33 34 35 36 37 43 44 46 47
Adherens junction	0.794533	15	0.630206366	0.70761569	4 5 11 25 26 28 33 36 38 40 43 44 46 47 48
Adipocytokine signaling pathway	0.831000	6	0.596127825	0.703833945	4 5 10 42 46 48
TNF signaling pathway	0.790667	12	0.621398946	0.700941819	4 6 11 16 36 39 41 42 45 46 47 48
Neurotrophin signaling pathway	0.794800	15	0.61762929	0.700636681	4 6 11 25 34 36 38 39 40 43 44 45 46 47 48
B cell receptor signaling pathway	0.839583	12	0.583361014	0.699842972	4 33 34 36 38 39 42 44 45 46 47 48
Fc epsilon RI signaling pathway	0.785500	14	0.623089264	0.699597468	4 6 11 25 33 34 36 38 40 43 44 46 47 48
Cell cycle	0.705455	11	0.681447933	0.693347346	4 5 9 10 12 13 22 29 30 32 47
Insulin resistance	0.854000	4	0.560416943	0.691806381	4 5 42 46
PI3K-AKT signaling pathway	0.802462	13	0.584009347	0.68457654	4 22 26 33 34 35 36 38 42 44 45 46 47
Focal adhesion	0.800353	17	0.576200699	0.679090513	4 11 22 25 26 33 34 36 38 40 42 43 44 45 46 47 48
AMPK signaling pathway	0.817000	4	0.562233667	0.677749885	4 10 42 44
NOD-like receptor signaling pathway	0.786500	10	0.580649858	0.675781853	4 6 11 36 39 41 43 46 47 48
Sphingolipid signaling pathway	0.782615	13	0.576929156	0.671947642	4 6 11 33 34 35 36 43 44 45 46 47 48
T cell receptor signaling pathway	0.776857	14	0.577623933	0.669874076	4 6 11 25 26 34 36 38 39 40 44 46 47 48
JAK-STAT signaling pathway	0.830000	6	0.523496172	0.659167523	4 10 34 42 44 46
RAS signaling pathway	0.780833	18	0.548420257	0.654388889	4 8 11 22 25 26 33 34 36 38 40 42 43 44 45 46 47 48
Mismatch repair	0.720200	5	0.582186126	0.647526407	9 15 24 30 32
Estrogen signaling pathway	0.731111	18	0.559789644	0.639740908	1 3 4 6 14 20 31 34 35 36 38 39 40 41 44 45 46 47
MAPK signaling pathway	0.777053	19	0.514896219	0.63253574	4 6 8 11 20 22 25 26 34 36 38 39 42 43 44 45 46 47 48
RAP1 signaling pathway	0.736048	21	0.539811636	0.630338853	1 4 6 11 14 22 25 26 31 33 34 35 36 38 42 43 44 45 46 47 48

Table 3. Pathway enrichment analysis (k-clique 9) and their associated weights.

Breast cancer integrated network. Figure S2 shows the BC integrated network conformed by 334 genes and proposed by this study: genes from the most relevant communities ($n = 84$), pathogenic genes (G1 + G2) ($n = 115$), PAM50 genes ($n = 26$), the ERBB signaling pathway ($n = 54$), the FOXO signaling pathway ($n = 27$), the HIF-1 signaling pathway ($n = 40$), the MAPK signaling pathway ($n = 68$), the mTOR signaling pathway ($n = 31$), the p53 signaling pathway ($n = 40$), the PI3K-AKT signaling pathway ($n = 114$) and the VEGF signaling pathway ($n = 31$).

Additionally, Fig. 3 shows a circular chord diagram of the BC integrated network to better understand the PPI in BC. Genes of the most relevant communities were most associated with MAPK, PI3K-AKT and HIF-1 signaling pathways. Pathogenic genes were most associated with PI3K-AKT, MAPK and FOXO signaling pathways. PAM50 genes were most associated with PI3K-AKT, ERBB and HIF-1 signaling pathways. The ERBB and FOXO signaling pathways were most associated with PI3K-AKT and MAPK signaling pathways. The prolactin, mTOR, p53, HIF-1 and MAPK signaling pathways were most associated with PI3K-AKT and FOXO signaling pathways. The VEGF signaling pathway was most associated with ERBB and MAPK signaling pathways. Finally, the PI3K-AKT signaling pathway was most associated with MAPK and FOXO signaling pathways (Table S11).

PAM50 subtypes. Regarding the intrinsic molecular subtypes obtained from the PAM50 mRNA-based assay^{3,6-9,52-54}, the CS identified 31 of 50 (62%) PAM50 genes. Focused heatmap of classification by nearest centroids selected genes for each subtype: luminal A ($n = 7$), normal-like ($n = 6$), luminal B ($n = 6$), HER2-enriched ($n = 7$), and basal-like ($n = 5$). The average ranking between luminal A (637.1) with normal-like (624.8), luminal B (106.2) with HER2-enriched (98), and basal-like (738.6) was correlated with the heatmap dendrogram of the centroid models of subtype of Parker *et al.*³.

The PPI network created using String Database allowed identifying 26 of 50 (52%) PAM50 genes. The expression patterns of PAM50 are detailed in Table S12³. Additionally, the PPI between PAM50 and genes of the most relevant communities, pathogenic genes, and the most relevant KEGG signaling pathways in BC are detailed in Table S12.

Oncogenomics validation with the OncoPPI BC network. Of the 1484 genes that make up the String Database³⁴, 77 genes (5.2%) were part of the OncoPPI BC network^{37,38}. The degree centrality allowed to establish a significant correlation (Spearman $p < 0.001$; $r^2 = 0.273$) between the OncoPPI BC network and genes of this

network present in our String Database. On the other hand, of the 334 genes that make up the BC integrated network, 51 genes (15%) were part of the OncoPPi BC network. The degree centrality allowed to establish a significant correlation (Spearman $p < 0.05$; $r^2 = 0.237$) between the OncoPPi BC network and genes of this network present in our BC integrated network (Table S13).

Figure 4 shows the correlation of PPI between the OncoPPi BC network and our BC integrated network. This sub-network is conformed by 20 genes of the most relevant communities, 3 PAM50 genes, 4 pathogenic genes (G1 + G2), 7 genes of the PI3K-AKT signaling pathway, 1 gene of the ERBB signaling pathway, 2 genes of the FOXO signaling pathway, 1 gene of the HIF-1 signaling pathway and 13 multiple signaling pathway genes. Finally, this sub-network has 62 breast cancer-associated PPI according to the OncoPPi network (Table S14).

Oncogenomics validation with DRIVE. Regarding our results, DRIVE detected 70.6% (1300/1842) of the Consensus genes, of which 3.08% (40 genes) was essential (sensitivity score ≤ -3) in all cancer cell lines ($n = 398$) and 4.15% (54 genes) presented sensitivity score ≤ -3 in $>50\%$ of BC cell lines ($n = 24-25$)³⁹. DRIVE detected 82% (273/334) of genes that make up the BC integrated network, of which 2.93% (8 genes) was essential in all cancer cell lines and 5.50% (15 genes) presented sensitivity score ≤ -3 in $>50\%$ of BC cell lines. Regarding genes that make up the most relevant communities, DRIVE detected 94% (79/84), of which 3.80% (3 genes) was essential in all cancer cell lines and 6.33% (5 genes) presented sensitivity score ≤ -3 in $>50\%$ of BC cell lines, observing an enrichment in the detection in contrast with the Consensus genes. Similarly, DRIVE detected 81% (76/94) of genes that make up the OncoPPi BC network, of which 3.95% (3 genes) was essential in all cancer cell lines and 6.58% (5 genes) presented sensitivity score ≤ -3 in $>50\%$ of BC cell lines. DRIVE detected 76% (110/145) of pathogenic genes G1 + G2, of which 2.73% (3 genes) was essential in all cancer cell lines and 4.55% (5 genes) presented sensitivity score ≤ -3 in $>50\%$ of BC cell lines (Fig. 5a,b). Finally, we proposed a normalized gene list according to the Consensus genes and the sensitivity score ≤ -3 in all cancer cell lines (Table S15) and BC cell lines (Table S16).

Additionally, Fig. 5c shows a Venn diagram of 54 genes with significant sensitivity score (≤ -3) in $>50\%$ of BC cell lines. Of which, CCND1, CDC42, YAP1, RPA1 and RAD51 integrated the most relevant communities, CCND1, CDC42, ITGAV, TFDP1 and TRRAP integrated the OncoPPi BC network, CCND1, CDC42, RPA1, RAD51, CDK1, SMC2, XRCC6, ITGAV, PLK1, MCL1, BCL2L1, ITGB5, RBX1, PPP2RIA and CRKL integrated the BC integrated network, and finally, all 54 genes were part of the Consensus genes. On the other hand, the Venn diagram of the essential genes in all cancer cell lines is shown in Fig. S3.

Integrated metabolic network and compounds. The reference global network from the 1842 genes was mapped obtaining three significant models ($p < 0.005$) using RSpider³². Model 1 has 662 initial genes, model 2 has 724 initial genes and model 3 has 746 initial genes. The p-value indicates the probability for a random gene/protein list to have a maximal connected component of the same or larger size. This p-value is computed by Monte Carlo simulation as described by Antonov *et al.*³².

The expanded integrated metabolic network (model 3) (Fig. S4) allows the entrance of 299 (957 in total) genes in order to bring connections between initial genes. However, it incorporates 66 compounds that also acts as connectors. These compounds obtained from the integrated metabolic network are fully detailed in Table S17.

Discussion

The CS improves the detection and prioritization of pathogenic genes. In our study, 19,989 genes were analyzed and after prioritization analysis we obtained a top ranking of 1842 genes where the top 10 genes with highest ranking were TP53, ESR1, BRCA2, BRCA1, ERBB2, CHECK2, CCND1, AR, RAD51 and ATM; and where 137 of 145 (94.5%) predefined pathogenic genes associated with BC were identified. CS is the method with highest identification of pathogenic genes in G1 and G2 datasets. Regarding both datasets, CS identified the 40.6% of G1 + G2 sets in the 1% and the 92.3% of G1 + G2 sets in the 10% of the final gene list compared to the second best method (Phenolyzer) that identifies the 31.6% of G1 + G2 sets in the 1% and the 74.2% of G1 + G2 sets in the 10% of the final gene list. Previous studies by Tejera *et al.* and Cruz-Monteaquedo *et al.*, have shown that CS in prioritization improves the detection of genes related with specific pathologies such as Parkinson's and preeclampsia^{17,55}. The importance of combining different prioritization strategies can remove noisy information and increase the relevance of gene-disease association¹⁷. Therefore, this study proves for the first time that CS improves the early enrichment ability of genes related with BC pathogenesis.

The BP from the Consensus genes allowed obtaining already expected information associated with BC. The most relevant BP with major biological meaning were: ERBB2 signaling pathway, whose overexpression can increase tyrosine kinase activities triggering down-stream pathways⁵⁶. DNA synthesis involved in DNA repair, in which DNA lesions have been found to be repairable by proteins either under clinical trials for current drug targets, namely BRCA1 and PARP-1^{42,57}. Phosphatidylinositol-3-phosphate plays a key regulatory function in cell survival, proliferation, migration, angiogenesis and apoptosis⁵⁸. The epidermal growth factor cellular stimulus generates the overexpression of EGFR triggering poor clinical outcomes in BC. Finally, the major signaling pathways activated by EGFR receptors are mediated by PI3K, RAS/MAPK and JNK resulting in a plethora of biological functions^{44,59}.

It is hard to establish a pathway ranking according to their implications in BC without further enrichment analysis. It is the main reason to combine the analysis of the PPI network. The String Database network with 1484 nodes already comprises the 85.5% of predefined pathogenic genes. The sub-network containing only genes belonging to some communities have the 43% of predefined pathogenic genes. On the other hand, the average degree of the pathogenic genes (37.4) was statistically significant higher than non-pathogenic genes (18.1). That is, the connectivity degree could be associated with the pathogenicity in this network.

Communities	Genes	Average ConsenScore _i	Average Rank	Average Degree	N pathogenic	Pathogenic genes/genes	HPT* (p-value)
46	CREBBP MAPK14 AKT1 SRC ESR1 JUN RAC3 CCND1 NFKB1 RELA	0.939	147.4	138	4	0.400	0.007783988
45	AKT1 MMP9 BCL2 VEGFA JUN TP53 TGFB1 IL6 FGF2 MMP2	0.924	181.8	181.8	7	0.700	3.25867E-06
47	MAPK14 CTNNB1 MAPK8 RAC1 SRC ABL1 MAPK1 JUN RAC3 STAT3 TP53 CCND1 FOS	0.899	240.62	45.62	3	0.231	0.098109212
42	AKT1 VEGFA JUN LEP TGFB1 IGF1 IL6 INS SERPINE1	0.887	269.89	101.3	6	0.667	2.72754E-05
44	CDH2 CTNNB1 AKT1 RAC1 SRC CDC42 CDH1 PIK3CA CCND1	0.885	275	141.11	4	0.444	0.00500697
30	RPA1 RPA3 CDK4 RAD51C ATM ATR DMC1 NBN MRE11 RBBP8 H2AFX RAD51	0.862	328.83	42.67	5	0.417	0.002288344
37	CREBBP PPARA MED1 NCOA1 CARM1 NCOA6 YAP1 CTGF WWTR1 NCOA2	0.862	330.1	60.6	0	0.000	N/A
41	MMP9 VEGFA JUN STAT3 CXCL8 IL6 TIMP1 MMP2 IL1B	0.853	352	80.2	5	0.556	0.000452371
43	CDH2 MAPK14 CTNNB1 MAPK8 RAC1 SRC CDC42 ABL1 CCND1	0.849	365.56	124.67	2	0.222	0.182829173
38	PIK3CA EGF EGFR GRB2 ERBB2 ERBB3 ERBB4 CBL PLCG1	0.848	362.33	89.3	3	0.333	0.037259742
48	MAPK14 MAPK8 RAC1 SRC ABL1 MAPK1 LCK STAT3 FYN	0.841	379.33	127.11	1	0.111	0.562833095
32	CDK2 RPA1 RPA3 CDK4 ATM DMC1 MLH1 MRE11 BLM TOP3A H2AFX RAD51	0.824	421.25	48.75	2	0.250	0.080438401
5	CREBBP SRA1 CITED2 PPARGC1A EP300 PPARA MED1 NR1P1 NCOA1	0.8	423.2	76.8	0.0	0.000	N/A
20	CREBBP JUN TP53 ATF2 KAT2B SMARCB1 IRF1 NR3C1 SMARCE1 HMGB1 ARID1A	0.8	398.7	85.4	1.0	0.091	0.636520998

Table 4. Genes present in the most relevant communities in k-clique 9. *HPT: Hypergeometric probability test.

TP53, AKT1, SRC, CREBBP, EP300, JUN, CTNNB1, RAC1, PIK3CA, EGFR, MAPK8, MAPK1, STAT3, ESR1, MAPK14, CCND1, GRB2, CDK2, FOS and CDKN1A are those genes with highest connectivity degree. The 95% of these genes (19/20) are present in at least one of the 14 most relevant communities. The minimal average ranking, the highest average degree and the Euclidean distance for the identification of clusters using K-mean allowed to determine that the cluster 1 conformed by the 14 communities (46, 45, 47, 42, 44, 30, 37, 41, 43, 38, 48, 32, 5 and 20) are more related with BC.

The CNA determined 84 genes present in the most relevant communities, of which, 12 were BC driver genes according to The Cancer Genome Atlas (TCGA) and the IntOGen web platform⁶⁰. In addition, 35 were tier 1 in the Cancer Gene Census⁶¹, and 19 of these were cancer hallmarks according to COSMIC^{62,63}, and Hanahan and Weinberg (Table S18)^{10,64}. Oncogenes were ERBB2, CCND1, EGFR, PIK3CA, ERBB3, CDK4, MAPK1, ABL1, LCK and RAC1; tumor suppressor genes were ATM, CDH1, EP300, ATR and BLM; and genes with both features were TP53, ESR1, ERBB4 and CREBBP.

On the other hand, the top 10 statistically significantly mutated genes identified by MutSigCVv.1.4 across the BC samples (n = 1087) in the Pan-Cancer Atlas were PIK3CA (34.7%), TP53 (34.7%), CDH1 (13.3%), GATA3 (12.8%), MAP3K1 (9.1%), PTEN (6.1%), RUNX1 (4.8%), NF1 (4.6%), MAP2K4 (4.4%) and ARID1A (4.3%)^{65,66}. The CS identified the 80% and the CNA analyzed the 40% of these genes.

Regarding the pathway enrichment analysis (k-clique 9) using David Bioinformatics Resource²⁸, the most significant BC signaling pathways for the most relevant communities were ERBB, prolactin, mTOR, p53, FOXO, HIF-1, VEGF, PI3K-AKT and MAPK signaling pathways.

The ERBB signaling pathway members form cell-surface receptors with extracellular domains yielding ligand-binding specificity⁶⁷. Downstream signaling from these receptors proceeds via tyrosine phosphorylation mediating signal transduction events that control cell proliferation, migration and survival. However, aberrant ERBB activation in BC can increase transcriptional expression⁴⁴. Genes of the most relevant communities that make up this pathway were MAPK1, MAPK8, ABL1, SRC, AKT1, PIK3CA, EGFR, ERBB3, EGF, ERBB2, CBL, GRB2, PLCG1, ERBB4 and JUN.

The prolactin signaling pathway and its downstream JAK2/STAT5 pathway are involved in the mammary gland development⁶⁸. Furthermore, prolactin and its receptor were found to play a permissive role in oncogene-induced mammary tumors⁶⁹. Genes of the most relevant communities that make up this signaling pathway were MAPK1, FOS, NFKB1, ESR1, RELA, MAPK8, MAPK14, SRC, CCND1, AKT1, INS, STAT3, PIK3CA, GRB2 and IRF1.

The PI3K-AKT-mTOR pathway plays a significant role in proliferation and cell survival in BC⁷⁰. The PI3K heterodimer (p85 and p110) phosphorylates phosphatidylinositol 4,5 biphosphate to phosphatidylinositol 3,4,4-triphosphate, which in turn leads to the phosphorylation of AKT, which has impact on cancer cell cycling, survival and growth⁴⁵. In addition, mTOR is associated with cell metabolism and cancer cell growth^{32,45}. Regarding antitumor efficacy, Woo *et al.*, suggests that both AKT and mTOR inhibitors have greater antitumor activity in BC⁷¹. Genes of the most relevant communities that make up the mTOR signaling pathway were MAPK1, AKT1, INS, IGF1, PIK3CA and GRB2; and that make up the PI3K-AKT signaling pathway were MAPK1, NFKB1,

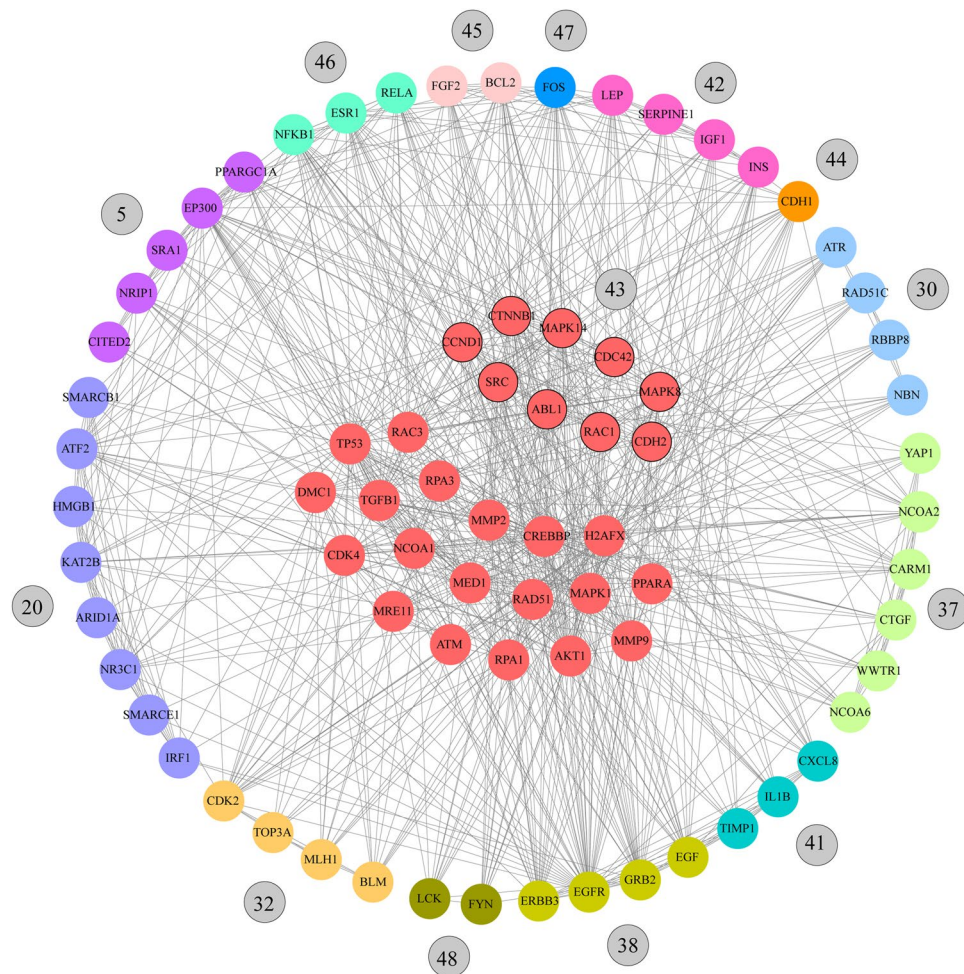


Figure 2. Community network analysis for k-clique 9. Red nodes represent genes that are part of several communities. The other colors correspond with the most relevant communities obtained.

RELA, FGF2, BCL2, RAC1, CCND1, AKT1, IGF1, INS, IL6, VEGFA, PIK3CA, GRB2, EGFR, EGF, CDK2, CDK4, TP53 and ATF2.

The p53 tumor suppressor holds distinction as the most frequently mutated gene in human cancer⁷². Acting as a transcription factor, p53 plays a critical role in growth-inhibition, angiogenesis, apoptosis and cell migration⁷³. Genes of the most relevant communities that make up this pathway were CCND1, IGF1, SERPINE1, CDK2, CDK4, ATM, ATR and TP53.

FOXO transcription factors play a critical role in pathological processes in BC. Those transcription factors regulate phosphorylation, acetylation and ubiquitination⁷⁴. Genes of the most relevant communities that make up this pathway were CREBBP, EP300, MAPK1, MAPK8, MAPK14, CCND1, TGFB1, AKT1, IGF2, INS, STAT3, IL6, PIK3CA, EGFR, EGF, GRB2, CDK2 and ATM.

Hypoxic conditions increase levels of HIF-1 signaling pathway in BC, inducing the expression of genes involved in angiogenesis, resistance to oxidative stress, cell proliferation, apoptosis and metastasis⁷⁵. Genes of the most relevant communities that make up this pathway were CREBBP, EP300, MAPK1, NFKB1, RELA, BCL2, AKT1, SERPINE1, IGF1, INS, STAT3, VEGFA, IL6, TIMP1, PIK3CA, PLCG1, EGFR, EGF and ERBB2.

The VEGF signaling pathway not only contributes to angiogenesis and vascular permeability but also contributes in BC tumorigenesis⁷⁶. Genes of the most relevant communities that make up this pathway were MAPK1, RAC3, MAPK14, RAC1, SRC, CDC42, AKT1, VEGFA, PIK3CA and PLCCG1.

MAPK signaling pathway is involved in cell growth, proliferation, differentiation, migration, and apoptosis^{77–79}. Genes of the most relevant communities that make up this pathway were MAPK1, FOS, RAC3, NFKB1, RELA, FGF2, MAPK8, MAPK14, RAC1, CDC42, TGFB1, AKT1, IGF1, INS, VEGFA, EGFR, EGF, GRB2, TP53, JUN and ATF2.

According to Li *et al.* and Ivanov *et al.*^{37,38}, the integration of cancer genes into networks offers opportunities to reveal PPI with therapeutic significance. The PPI mediates the regulation of oncogenic signals that are essential to cellular proliferation and survival, and thus represent potential targets for drug discovery. However, only a small portion of the PPI landscape has been described³⁷. The OncoPPI BC network was conformed by 94 genes and 170 PPI experimentally analyzed in BC cell lines^{37,38}. We carried out the validation of our String Database and our BC integrated network by comparing the degree centrality of both networks with the OncoPPI BC network^{37,38}.

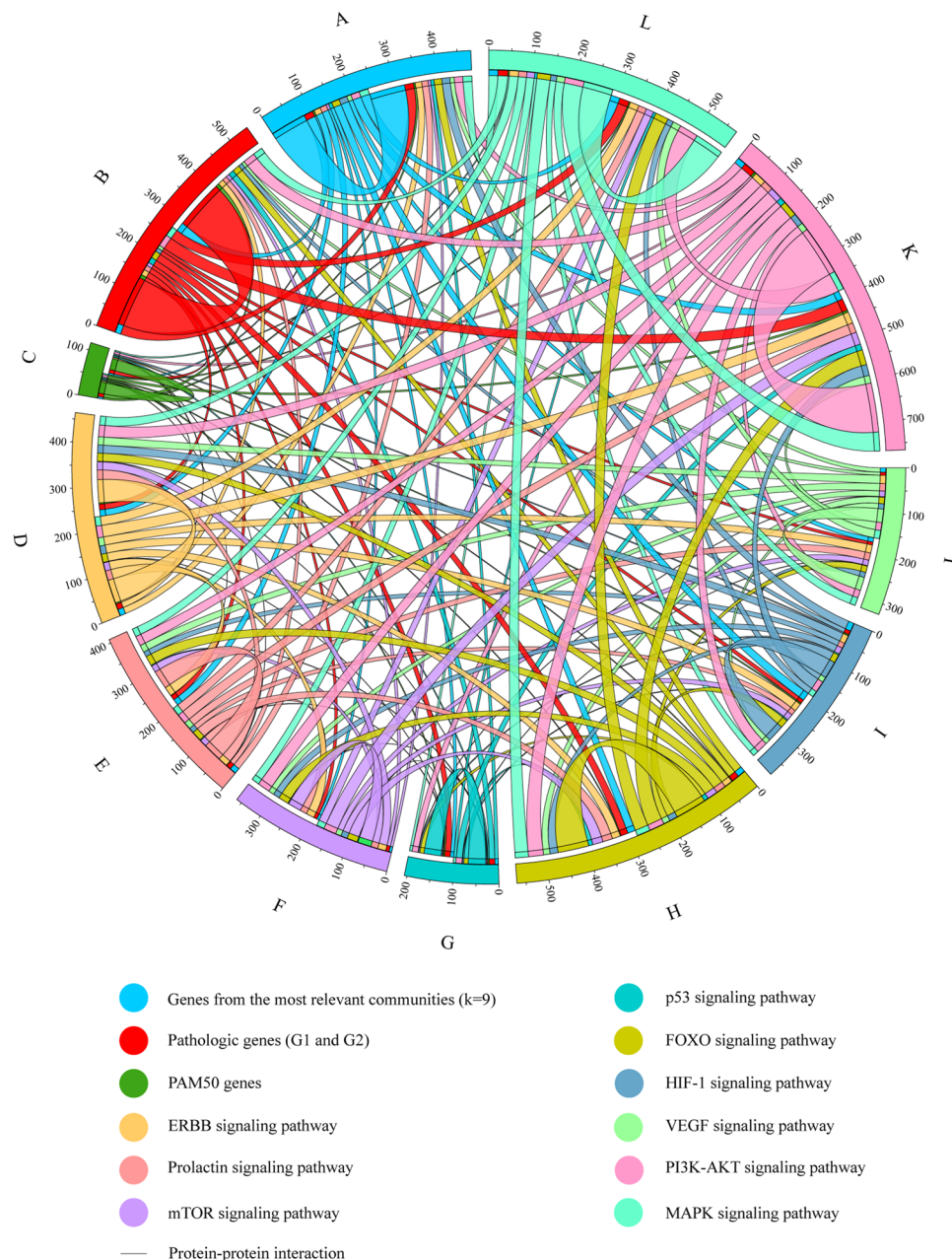


Figure 3. Circular chord diagram of the BC integrated network. PPI among the most relevant communities (k-clique 9), pathogenic genes (G1 + G2), PAM50 genes and genes of the most relevant KEGG signaling pathways in BC.

The degree centrality allowed to establish a significant correlation ($p < 0.001$) between the OncoPPI BC network and genes of this network present in our String Database. Similarly, the degree centrality allowed to establish a significant correlation ($p < 0.05$) between the OncoPPI BC network and our BC integrated network. Finally, the sub-network that shares 62 breast cancer-associated PPI between the OncoPPI BC network and our BC integrated network is shown in Fig. 4 and Table S12. The 20 genes of the most relevant communities present in this sub-network were CBL, NFKB1, STAT3, CTNNA1, INS, MAPK8, MAPK14, FYN, JUN, PIK3CA, AKT1, FOS, RELA, TP53, RAC1, CDC42, CDK4, CCND1, SRC and ERBB3.

The CS was effective in the prioritization of genes involved in the expression of BC intrinsic molecular subtypes. The CS identified 31 of 50 (62%) PAM50 genes. The best average ranking corresponded to HER2-enriched (98), followed by luminal B (106.2), normal-like (624.8), luminal A (637.1) and basal-like (738.6). The correlation between average rankings and intrinsic molecular subtypes could be observed in the heatmap dendrogram of the centroid models of subtype of Parker *et al.*³. On the other side, our String network allowed to identify 26 of 50 (52%) PAM50 genes. Of these, 8 were tier 1 in the Cancer Gene Census and 7 were cancer hallmarks^{61–63}.

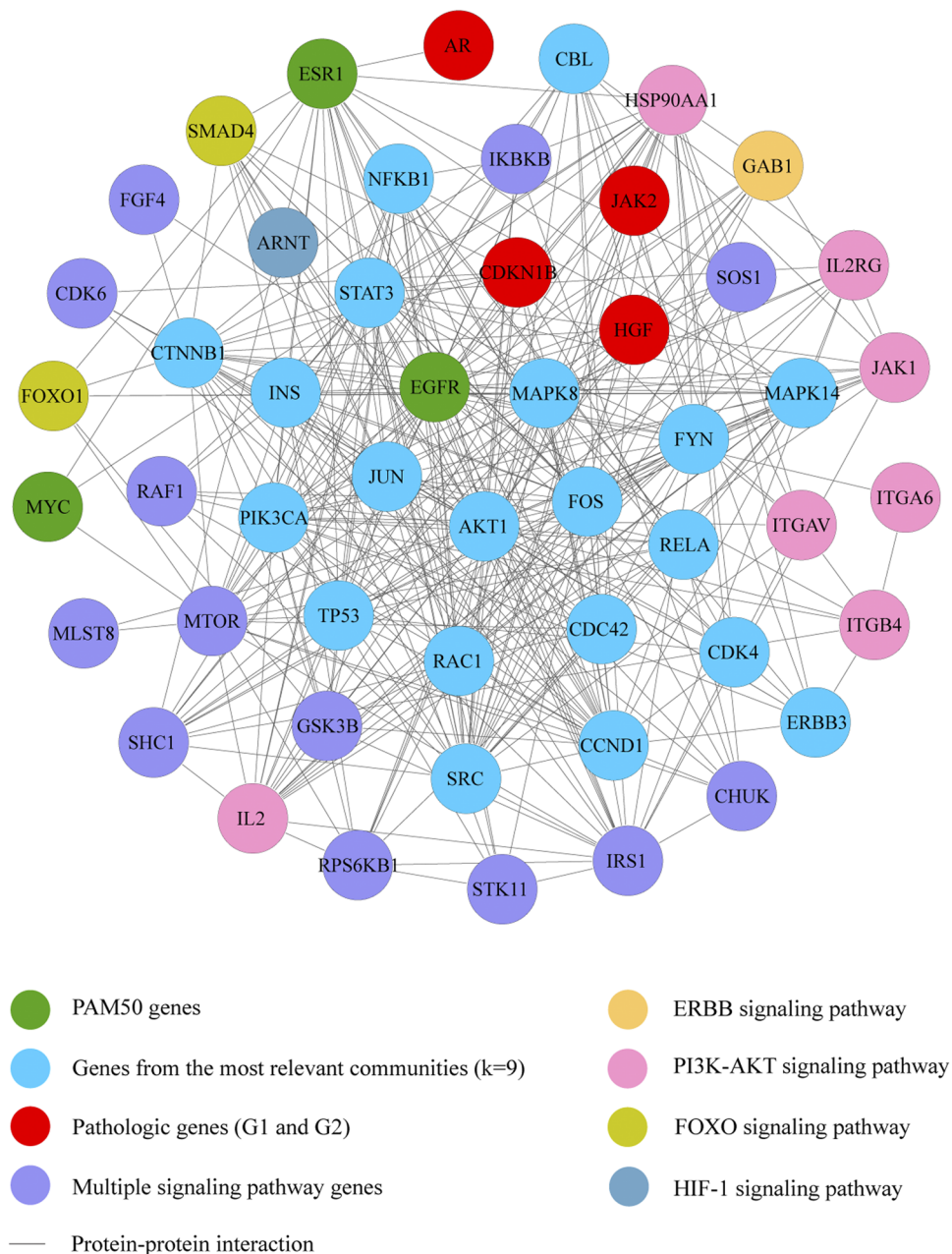


Figure 4. Significant correlation of degree centrality between the OncoPpi BC network and our BC integrated network ($p < 0.05$), ($r^2 = 0.23688$). This sub-network is conformed by genes of the most relevant communities (k-clique 9), pathogenic genes (G1 + G2), PAM50 genes, and genes of the ERBB, PI3K-AKT, FOXO, and HIF-signaling pathways in BC.

Table S11 details the PPI between PAM50 and genes from the most relevant communities. These interactions could be a guide to enrich future experimental studies related to find breast cancer-focused PPI per each molecular subtype. Finally, the circular chord diagram of the BC integrated network showed that PAM50 was most associated with the PI3K-AKT, ERBB, HIF-1, p53 and MAPK signaling pathways.

According to McDonald *et al.*, DRIVE is the larger-scale gene knockdown experiment to discover functional gene requirements across 398 cancer cell lines and 24-25 BC cell lines³⁹. The sensitivity score analysis was performed on the genes that make up the Consensus, communities, BC integrated network, pathogenic genes and OncoPpi BC network (Fig. 5a,b). In all these groups, a higher percentage of genes with significant sensitivity score (≤ -3) could be observed in BC cell lines than in all cancer cell lines. This means that the CS and CNA in BC pathogenesis have been effective and corroborated by DRIVE. Hence, the 4.15% (54 genes) of the Consensus has significant sensitivity score in $>50\%$ of BC cell lines and 6.33% (5 genes) of genes from the most relevant communities has significant sensitivity score in $>50\%$ of BC cell lines.

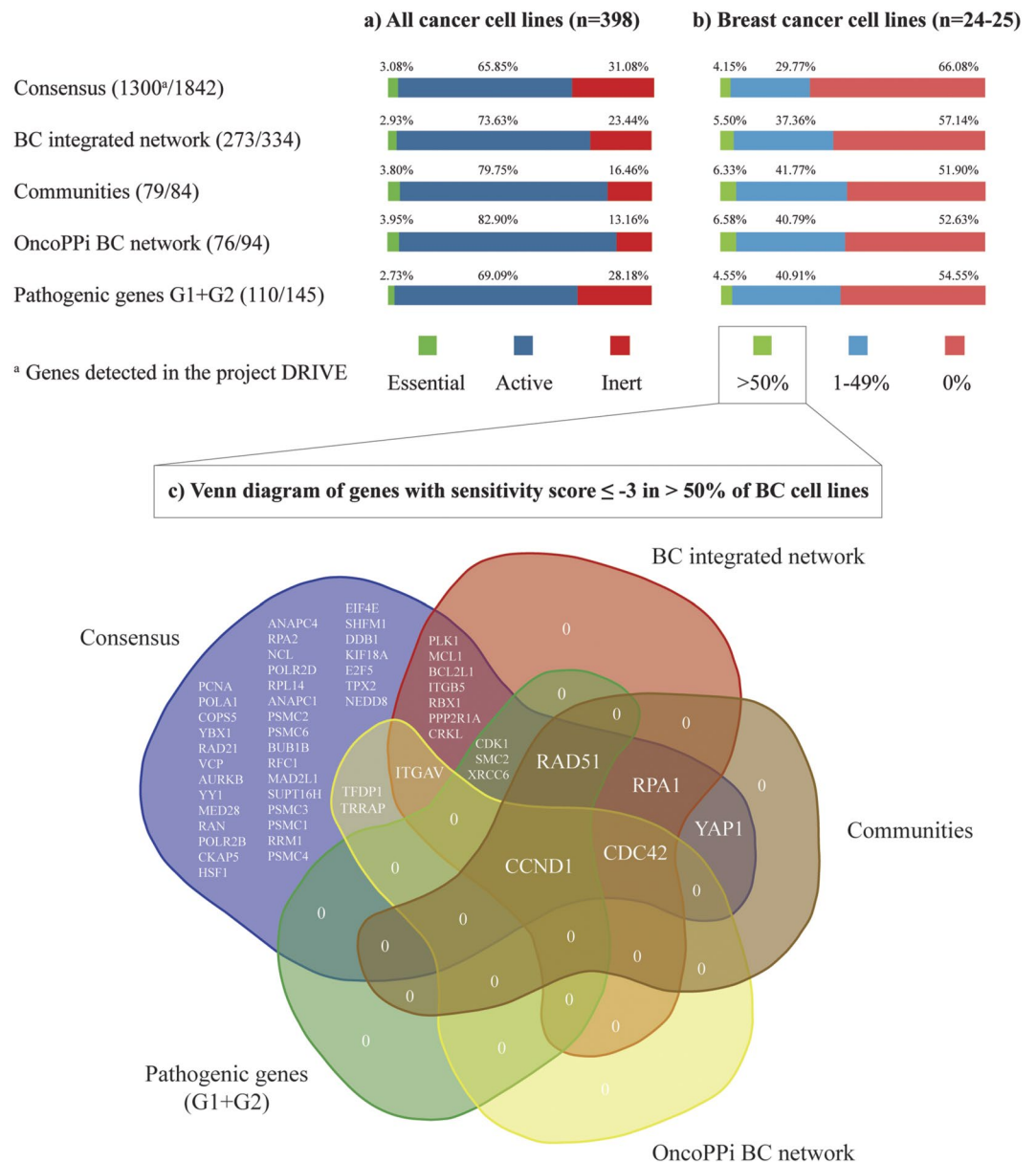


Figure 5. Oncogenomics validation with the DRIVE project. **(a)** Percentage of essential, active and inert genes in all cancer cell lines. **(b)** Percentage of genes with sensitivity score ≤ -3 in >50%, 1–40%, and 0% of BC cell lines. **(c)** Venn diagram of genes with significant sensitivity score in >50% of BC cell lines.

CCND1, CDC42, RAD51, RPA1 and YAP1 were genes with significant sensitivity score in >50% of BC cell lines present not only in the communities but also in the Consensus, BC integrated network, pathogenic genes and OncoPPi BC network (Fig. 5c)^{37,38}. Regarding those genes, high expression of the CCND1 oncogene is associated to high proliferation rate and increased risk of mortality in ER-positive women⁸⁰. CDC42 is a protein kinase that controls cell migration and progression through G1 to S phase for DNA synthesis⁸¹. RAD51 is a key player in DNA double-strand break repair. Lack of RAD51 nuclear expression is associated with poor prognostic parameters in invasive BC⁸². RPA1 is upregulated in BC tumors and plays an essential role in DNA replication and repair⁸³. Finally, YAP1, a major downstream effector of the Hippo pathway, has an important role in tumor growth. Elevated oncogenic activity of YAP1 contributes to BC cell survival⁸⁴.

The expanded integrated metabolic network (Model 3) (Fig. S4) incorporates 66 compounds that act as connectors according to the Human Metabolome Database⁸⁵, giving us more information related to pharmacogenomics⁸⁶. The metabolic species with the highest connectivity in our network were biophosphate, deoxyguanosine diphosphate (dGDP), cyclic GMP (cGMP), phosphatidate, glutathione (GSH), hydrogen carbonate (HCO₃-), lecithin and benzo[a]pyrene-4,5-oxide. Biophosphate participates in phosphatidylinositol biosynthesis. According to Clarke *et al.*, phosphatidylinositol is critical for intracellular signaling and anchoring of carbohydrates and proteins to outer cellular membranes⁸⁷. dGDP is involved in pyrimidine and purine metabolisms. cGMP acts on the purine metabolism. According to Fajardo *et al.*, altered cGMP signaling has been observed in

BC⁸⁸. GSH and benzo[a]pyrene-4,5-oxide are involved in glutathione metabolism. According to Lien *et al.*, oncogenic PI3K-AKT stimulates glutathione biosynthesis in mammary human cells by activating Nrf2 to upregulate the GSH biosynthesis genes⁸⁹. HCO3⁻ is involved in propanoate and pyruvate metabolisms. According to Zhu *et al.*, the dysfunction of propanoate and pyruvate metabolisms can trigger the BC progression⁹⁰. Finally, phosphatidate and lecithin are involved in the glycerophospholipid metabolism. According to Huang and Freter, the glycerophospholipids are the main component of biological membranes⁹¹.

The contribution of each individual approach on the whole consensus was analyzed according to the pathogenic genes G1 + G2 as shown in Fig. S5. The CS was evaluated between several prioritization strategies guiding us to genes with pathogenic involvement in BC. Subsequently, the PPI network and the communality network analyses allowed us to obtain a group of genes increasingly associated with BC. For instance, 0.074 was the ratio between the 145 pathogenic genes (G1 + G2) and the CS genes (n = 1842), 0.083 was the ratio between the 124 pathogenic genes and the PPI network (n = 1484), 0.127 was the ratio between the 63 pathogenic genes and all communities (n = 496), and 0.262 was the ratio between the 22 pathogenic genes with the 14 most relevant communities (n = 84 genes). On the other hand, 0.235 was the ratio between the 22 pathogenic genes and the OncoPPI BC network (n = 51), 0.116 was the ratio between the 45 pathogenic genes and the active genes (n = 387) of the DRIVE BC cell lines, lastly, 0.093 was the ratio between the 5 pathogenic genes and the essential genes (n = 54) of the DRIVE BC cell lines. The oncogenomics validations showed that BC is a complex disease whose development and progression is due in large part to the alteration of genes, metabolites and pathways analyzed in this research and leading us towards reasonable discussion in agreement with our scientific knowledge of the disease. However, the proposed strategies need to be further improved in several topics: 1) the inclusion of other network processing methods to reduce the gene lost, 2) the inclusion of prioritization algorithms based on learning strategies, and 3) the differentiation among BC intrinsic molecular subtypes by bioinformatics tools. Finally, overlapping the barriers previously mentioned we would improve the gene prioritization strategy and the validation of the predicted subtype-specific drug targets such as Zaman *et al.* study⁹².

Data Availability Statement

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

References

- Espinal-Enrriquez, J., Fresno, C., Anda-Jáuregui, G. & Hernández-Lemus, E. RNA-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Sci. Rep.* **7**, 1760 (2017).
- Guerrero, S. *et al.* Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. *Sci. Rep.* **8**, 13978 (2018).
- Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
- Kumar. *Robbins Basic Pathology*. 10.1007/s13398-014-0173-7.2 Elsevier, (2007).
- Malhotra, G. K., Zhao, X., Band, H. & Band, V. Histological, molecular and functional subtypes of breast cancers. *Cancer Biol. Ther.* **10**, 955–60 (2010).
- Kumar, R., Sharma, A. & Tiwari, R. K. Application of microarray in breast cancer: An overview. *J. Pharm. Bioallied Sci.* **4**, 21–6 (2012).
- Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
- López-Cortés, A. *et al.* Breast cancer risk associated with gene expression and genotype polymorphisms of the folate-metabolizing MTHFR gene: a case-control study in a high altitude Ecuadorian mestizo population. *Tumour Biol.* **36**, 6451–61 (2015).
- Prat, A., Ellis, M. J. & Perou, C. M. Practical implications of gene-expression-based assays for breast oncologists. *Nature Reviews Clinical Oncology* **9**, 48–57 (2012).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
- Castro, M. A. A. *et al.* Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.* **48**, 12–21 (2016).
- Kitano, H. Opinion: Cancer as a robust system: implications for anticancer therapy. *Nat. Rev. Cancer* **4**, 227–235 (2004).
- Croce, C. M. Oncogenes and Cancer. *N. Engl. J. Med.* **358**, 502–511 (2008).
- Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* **10**, 789–799 (2004).
- Börnigen, D. *et al.* An unbiased evaluation of gene prioritization tools. *Bioinformatics* **28**, 3081–3088 (2012).
- Tranchevent, L.-C. *et al.* A guide to web tools to prioritize candidate genes. *Brief. Bioinform.* **12**, 22–32 (2011).
- Tejera, E. *et al.* Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis. *BMC Med. Genomics* **10**, 50 (2017).
- Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).
- Jourquin, J., Duncan, D., Shi, Z. & Zhang, B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics* **13**(Suppl 8), S20 (2012).
- Piñero, J. *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)*. **2015**, bav028 (2015).
- Fontaine, J.-F., Priller, F., Barbosa-Silva, A. & Andrade-Navarro, M. A. Génie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.* **39**, W455–61 (2011).
- Yue, P., Melamud, E. & Moulton, J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7**, 166 (2006).
- Guney, E., Garcia-Garcia, J. & Oliva, B. GUILDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. *Bioinformatics* **30**, 1789–90 (2014).
- Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Mol. Syst. Biol.* **4**, 189 (2008).
- Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841–3 (2015).
- Cheng, D. *et al.* PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.* **36**, W399–405 (2008).
- Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C. & Greene, C. S. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Brief. Bioinform.* **17**, 33–42 (2016).

28. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
29. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
30. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
31. Guala, D. & Sonnhhammer, E. L. L. A large-scale benchmark of gene prioritization methods. *Sci. Rep.* **7**, 46598 (2017).
32. Antonov, A. V., Schmidt, E. E., Dietmann, S., Krestyaninova, M. & Hermjakob, H. R spider: a network-based analysis of gene lists by combining signaling and metabolic pathways from Reactome and KEGG databases. *Nucleic Acids Res.* **38**, W78–83 (2010).
33. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**, 29–34 (1999).
34. Szklarczyk, D. *et al.* STRINGv10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–52 (2015).
35. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).
36. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–8 (2005).
37. Li, Z. *et al.* The OncoPPI network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nat. Commun.* **8** (2017).
38. Ivanov, A. A. *et al.* The OncoPPI Portal: an integrative resource to explore and prioritize protein-protein interactions for cancer target discovery. *Bioinformatics* 1–9, <https://doi.org/10.1093/bioinformatics/btx743> (2017).
39. McDonald, E. R. *et al.* Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**, 577–592.e10 (2017).
40. König, R. *et al.* A probability-based approach for the analysis of large-scale RNAi screens. *Nat. Methods* **4**, 847–849 (2007).
41. Tejera, E., Bernardes, J. & Rebelo, I. Co-expression network analysis and genetic algorithms for gene prioritization in preeclampsia. *BMC Med. Genomics* **6**, 51 (2013).
42. Montenegro, M. F. *et al.* Targeting the epigenetics of the DNA damage response in breast cancer. *Cell Death Dis.* **7**, e2180 (2016).
43. Liu, L. *et al.* Identification of STAT3 as a specific substrate of breast tumor kinase. *Oncogene* **25**, 4904–12 (2006).
44. Ali, R. & Wendt, M. K. The paradoxical functions of EGFR during breast cancer progression. *Signal Transduct. Target. Ther.* **2**, 16042 (2017).
45. Paplomata, E. & O’regan, R. The PI3K/AKT/mTOR pathway in breast cancer: Targets, trials and biomarkers. *Therapeutic Advances in Medical Oncology* **6**, 154–166 (2014).
46. Bullock, M. FOXO factors and breast cancer: outfoxing endocrine resistance. *Endocr. Relat. Cancer* **23**, R113–30 (2016).
47. Mestres, J. A., Mateo, M. M. & Gascón, P. ErbB tyrosine kinase receptor inhibitors in breast cancer. *Rev. Oncol.* **6**, 12–21 (2004).
48. Eckert, L. B. *et al.* Involvement of Ras Activation in Human Breast Cancer Cell Signaling, Invasion, and Anoikis Involvement of Ras Activation in Human Breast Cancer Cell Signaling. *Invasion*, 4585–4592, <https://doi.org/10.1158/0008-5472.CAN-04-0396> (2004).
49. Santen, R. J. *et al.* The role of mitogen-activated protein (MAP) kinase in breast cancer. *J. Steroid Biochem. Mol. Biol.* **80**, 239–256 (2002).
50. Mancini, M. L., Lien, E. C. & Toker, A. Oncogenic AKT1(E17K) mutation induces mammary hyperplasia but prevents HER2-driven tumorigenesis. *Oncotarget* **7** (2016).
51. Roberts, P. J. & Der, C. J. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene* **26**, 3291–3310 (2007).
52. Nielsen, T. O. *et al.* A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res.* **16**, 5222–5232 (2010).
53. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* **8**, 54 (2015).
54. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
55. Cruz-Monteaugado, M. *et al.* Efficient and biologically relevant consensus strategy for Parkinson’s disease gene prioritization. *BMC Med. Genomics* **9**, 12 (2016).
56. Yu, D. & Hung, M.-C. Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. *Oncogene* **19**, 6115–6121 (2000).
57. Davis, J. D. & Lin, S.-Y. DNA damage and breast cancer. *World J. Clin. Oncol.* **2**, 329–38 (2011).
58. Baselga, J. Targeting the Phosphoinositide-3 (PI3) Kinase Pathway in Breast Cancer. *Oncologist* **16**, 12–19 (2011).
59. Masuda, H. & Zhang, D. Role of epidermal growth factor receptor in breast cancer. *Breast cancer Res. ...* **136**, 1–21 (2012).
60. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
61. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
62. Forbes, S. A. *et al.* COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
63. Forbes, S. A. *et al.* Europe PMC Funders Group The Catalogue of Somatic Mutations in Cancer (COSMIC), <https://doi.org/10.1002/0471142905.hg1011s57.The> (2009).
64. Fouad, Y. A. & Aanei, C. Revisiting the hallmarks of cancer. *Am. J. Cancer Res.* **7**, 1016–1036 (2017).
65. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–8 (2013).
66. Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* 1–16, <https://doi.org/10.1016/j.ccell.2018.03.014> (2018).
67. Paz-y-Miño, C. *et al.* Incidence of the L858R and G719S mutations of the epidermal growth factor receptor oncogene in an Ecuadorian population with lung cancer. *Cancer Genetics and Cytogenetics* **196**, 201–203 (2010).
68. López-ozuna, V. M., Hac, I. Y., Hachim, M. Y., Lebrun, J. & Ali, S. Prolactin Pro-Differentiation Pathway in Triple Negative Breast Cancer: Impact on Prognosis and Potential Therapy. *Nat. Publ. Gr.* 1–13, <https://doi.org/10.1038/srep30934> (2016).
69. O’Leary, K. A., Rugowski, D. E., Sullivan, R. & Schuler, L. A. Prolactin cooperates with loss of p53 to promote claudin-low mammary carcinomas. *Oncogene* **33**, 3075–3082 (2014).
70. Vivanco, I. & Sawyers, C. L. The phosphatidylinositol 3-Kinase-AKT pathway in human cancer. *Nat. Rev. Cancer* **2**, 489–501 (2002).
71. Woo, S.-U. *et al.* Vertical inhibition of the PI3K/Akt/mTOR pathway is synergistic in breast cancer. *Oncogenesis* **6**, e385 (2017).
72. Murphy, M. E. *et al.* A functionally significant SNP in TP53 and breast cancer risk in African-American women. *npj Breast Cancer* **3**, 5 (2017).
73. Xie, B. *et al.* Benzyl Isothiocyanate potentiates p53 signaling and antitumor effects against breast cancer through activation of p53-LKB1 and p73-LKB1 axes. *Sci. Rep.* **7** (2017).
74. Fu, Z. & Tindall, D. J. FOXOs, cancer and regulation of apoptosis. *Oncogene* **27**, 2312–9 (2008).
75. Gilkes, D. M. & Semenza, G. L. Role of hypoxia-inducible factors in breast cancer metastasis. *Futur. Oncol.* **9**, 1623–1636 (2013).
76. Goel, H. L. & Mercurio, A. M. VEGF targets the tumour cell. *Nat. Rev. Cancer* **13**, 871–882 (2013).
77. Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nat. Rev. Cancer* **3**, 11–22 (2003).
78. Wellbrock, C., Karasarides, M. & Marais, R. The RAF proteins take centre stage. *Nat. Rev. Mol. Cell Biol.* **5**, 875–85 (2004).
79. Dhillon, A. S., Hagan, S., Rath, O. & Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **26**, 3279–3290 (2007).
80. Ahlin, C. *et al.* High expression of cyclin D1 is associated to high proliferation rate and increased risk of mortality in women with ER-positive but not in ER-negative breast cancers. *Breast Cancer Res. Treat.* **164**, 667–678 (2017).

81. Chrysanthou, E. *et al.* Phenotypic characterisation of breast cancer: the role of CDC42. *Breast Cancer Res. Treat.* **164**, 317–325 (2017).
82. Alshareeda, A. T. *et al.* Clinical and biological significance of RAD51 expression in breast cancer: a key DNA damage response protein. *Breast Cancer Res. Treat.* **159**, 41–53 (2016).
83. Hass, C. S., Gakhar, L. & Wold, M. S. Functional characterization of a cancer causing mutation in human replication protein A. *Mol. Cancer Res.* **8**, 1017–1026 (2010).
84. Li, L. *et al.* The deubiquitinase USP9X promotes tumor cell survival and confers chemoresistance through YAP1 stabilization. *Oncogene* **1**, <https://doi.org/10.1038/s41388-018-0134-2> (2018).
85. Wishart, D. S. *et al.* HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.* **41**, D801–D807 (2013).
86. López-Cortés, A., Guerrero, S., Redal, M. A., Alvarado, A. T. & Quiñones, L. A. State of art of cancer pharmacogenomics in Latin American populations. *International Journal of Molecular Sciences* **18** (2017).
87. Clarke, O. B. *et al.* Structural basis for phosphatidylinositol-phosphate biosynthesis. *Nat. Commun.* **6**, 8505 (2015).
88. Fajardo, A. M., Piazza, G. A. & Tinsley, H. N. The role of cyclic nucleotide signaling pathways in cancer: targets for prevention and treatment. *Cancers (Basel)* **6**, (436–58 (2014).
89. Lien, E. C. *et al.* Glutathione biosynthesis is a metabolic vulnerability in PI(3)K/Akt-driven breast cancer. *Nat. Cell Biol.* **18**, 572–8 (2016).
90. Zhu, X. *et al.* Identification of collaboration patterns of dysfunctional pathways in breast cancer. *Int. J. Clin. Exp. Pathol.* **7**, 3853–64 (2014).
91. Huang, C. & Freter, C. Lipid metabolism, apoptosis and cancer therapy. *Int. J. Mol. Sci.* **16**, 924–49 (2015).
92. Zaman, N. *et al.* Signaling Network Assessment of Mutations and Copy Number Variations Predict Breast Cancer Subtype-Specific Drug Targets. *Cell Rep.* **5**, 216–223 (2013).

Acknowledgements

This work was supported by Universidad UTE (Quito, Ecuador), Universidad de las Américas (Quito, Ecuador), University of Coruna (Coruña, Spain), University of the Basque Country (Bilbao, Spain), and McGill University (Montreal, Canada). Additionally, this work was supported by “Collaborative Project in Genomic Data Integration (CICLOGEN)” PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER).

Author Contributions

A.L.C. and E.T. conceived the subject and the conceptualization of the study. A.L.C. wrote the manuscript. E.T., S.J.B., C.R.M., H.G.D. and C.Py.M. supervised the project. A.L.C. and C.Py.M. did founding acquisition. A.L.C. and A.C.A. did data curation and supplementary data. E.T., S.J.B., C.R.M. and H.G.D. reviewed the manuscript. E.T., A.C.A., S.J.B., C.R.M., H.G.D., A.P. and Y.P.C. gave conceptual advice and valuable scientific input.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-35149-1>.

Competing Interests: The authors declare no competing interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

OPEN

OncoOmics approaches to reveal essential genes in breast cancer: a panoramic view from pathogenesis to precision medicine

Andrés López-Cortés^{1,2,3*}, César Paz-y-Miño¹, Santiago Guerrero¹, Alejandro Cabrera-Andrade^{2,4,5}, Stephen J. Barigye⁶, Cristian R. Munteanu^{2,7,8}, Humberto González-Díaz^{9,10}, Alejandro Pazos^{2,7,8}, Yunierkis Pérez-Castillo^{5,11} & Eduardo Tejera^{5,12*}

Breast cancer (BC) is the leading cause of cancer-related death among women and the most commonly diagnosed cancer worldwide. Although in recent years large-scale efforts have focused on identifying new therapeutic targets, a better understanding of BC molecular processes is required. Here we focused on elucidating the molecular hallmarks of BC heterogeneity and the oncogenic mutations involved in precision medicine that remains poorly defined. To fill this gap, we established an OncoOmics strategy that consists of analyzing genomic alterations, signaling pathways, protein-protein interactome network, protein expression, dependency maps in cell lines and patient-derived xenografts in 230 previously prioritized genes to reveal essential genes in breast cancer. As results, the OncoOmics BC essential genes were rationally filtered to 140. mRNA up-regulation was the most prevalent genomic alteration. The most altered signaling pathways were associated with basal-like and Her2-enriched molecular subtypes. *RAC1*, *AKT1*, *CCND1*, *PIK3CA*, *ERBB2*, *CDH1*, *MAPK14*, *TP53*, *MAPK1*, *SRC*, *RAC3*, *BCL2*, *CTNNB1*, *EGFR*, *CDK2*, *GRB2*, *MED1* and *GATA3* were essential genes in at least three OncoOmics approaches. Drugs with the highest amount of clinical trials in phases 3 and 4 were paclitaxel, docetaxel, trastuzumab, tamoxifen and doxorubicin. Lastly, we collected ~3,500 somatic and germline oncogenic variants associated with 50 essential genes, which in turn had therapeutic connectivity with 73 drugs. In conclusion, the OncoOmics strategy reveals essential genes capable of accelerating the development of targeted therapies for precision oncology.

Breast cancer (BC) is a complex and heterogeneous disease characterized by an intricate interplay between different biological aspects such as ethnicity, genomic alterations, gene expression deregulation, hormone disruption, signaling pathway alterations, hypoxia, and environmental determinants^{1,2}. Over the last years, prevention, treatment and survival strategies have evolved favorably; however, there are BC profiles that remain incurable³.

¹Centro de Investigación Genética y Genómica, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Mariscal Sucre Avenue, Quito, 170129, Ecuador. ²RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, A Coruña, 15071, Spain. ³Red Latinoamericana de Implementación y Validación de Guías Clínicas Farmacogenómicas (RELIVAF-CYTED), Quito, Ecuador. ⁴Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de Las Américas, Avenue de los Granados, Quito, 170125, Ecuador. ⁵Grupo de Bio-Quimioinformática, Universidad de Las Américas, Avenue de los Granados, Quito, 170125, Ecuador. ⁶Department of Chemistry, McGill University, 801 Sherbrooke Street West, Montreal, QC, H3A 0B8, Canada. ⁷Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), A Coruña, 15006, Spain. ⁸Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Campus de Elviña s/n, A Coruña, 15071, Spain. ⁹Department of Organic Chemistry II, University of the Basque Country UPV/EHU, Leioa, 48940, Biscay, Spain. ¹⁰IKERBASQUE, Basque Foundation for Science, Bilbao, 48011, Biscay, Spain. ¹¹Escuela de Ciencias Físicas y Matemáticas, Universidad de Las Américas, Avenue de los Granados, Quito, 170125, Ecuador. ¹²Facultad de Ingeniería y Ciencias Agropecuarias, Universidad de Las Américas, Avenue de los Granados, Quito, 170125, Ecuador. *email: aalc84@gmail.com; eduardo.tejera@udla.edu.ec

Nowadays, BC is the leading cause of cancer-related death among women (627,000; 15% cases) and the most commonly diagnosed cancer (2,088,849; 24% cases) worldwide⁴.

The development of large-scale DNA sequencing, gene expression, proteomics, large-scale RNA interference (RNAi) screens, large-scale CRISPR-Cas9 screens and patient-derived xenografts (PDXs) has allowed us to better understand the molecular landscape of oncogenesis. Considerable progress has been made in discovering coding and non-coding somatic drivers^{5,6}, cancer driver genes^{7,8}, cancer driver mutations^{9,10}, germline variants¹¹, driver fusion genes^{12,13}, alternatively spliced transcripts¹⁴, expression-based stratification¹⁵, molecular subtyping¹⁶, biomarkers¹⁷, druggable enzymes¹⁸, cancer dependencies^{19–22}, and drug resistance²³.

Scientific advances made to date mark the era called the “end of the beginning” of cancer omics. In other words, each approach that was previously mentioned needs to be fully understood as a part of a complex network, analyzing the mechanistic interplay of signaling pathways, protein-protein interactome (PPI) networks, enrichment maps, gene ontology (GO), deep learning, molecular dependencies and genomic alterations per intrinsic molecular subtype: basal-like (estrogen receptor (ER)⁻, progesterone receptor (PR)⁻, human epidermal growth factor receptor 2 (Her2)⁻, cytokeratin 5/6⁺ and/or EGFR⁺); Her2-enriched (ER⁻, PR⁻, Her2⁺); luminal A (ER⁺ and/or PR⁺, Her2⁻, low Ki67); luminal B with Her2⁻ (ER⁺ and/or PR⁺, Her2⁻, low Ki67); luminal B with Her2⁺ (ER⁺ and/or PR⁺, Her2⁺, any Ki67); and normal like^{24–30}.

Here we focus on elucidating the molecular hallmarks of BC essential genes and the oncogenic mutations applied in precision medicine that remains poorly defined. To fill this gap, we propose the OncoOmics strategy that consists in the analysis of genomic alterations (mRNA up-regulation, mRNA down-regulation, putative driver mutation, copy number variant (CNV) amplification, CNV deep deletion, and fusion gene), signaling pathways, PPI network, protein expression, BC dependencies in cell lines and patient-derived xenografts in a set of previously prioritized genes. These genes will come from our Consensus Strategy (CS) study²⁹, the Pan-Cancer Atlas (PCA) project^{3,13,31–37}, the Cancer Genome Interpreter (CGI) study³⁸, and the Pharmacogenomics Knowledgebase (PharmGKB)³⁹.

In our previous studies, López-Cortés *et al.*, Tejera *et al.*, and Cabrera-Andrade *et al.*, developed a Consensus Strategy that was proved to be highly efficient in the recognition of gene-disease association^{29,40,41}. The main objective was to apply several bioinformatics methods to explore BC pathogenic genes. On the other hand, The Cancer Genome Atlas (TCGA) has concluded the most sweeping cross-cancer analysis yet undertaken, namely the PCA project³². PCA reveals how genomic alterations and protein expression collaborate in BC progression, providing insights to prioritize the development of new treatments^{3,13,31–37}. The CGI flags genomic biomarkers of drug response with different levels of clinical relevance³⁸. Lastly, PharmGKB is a comprehensive resource that curates and spreads knowledge of the impact of clinical annotations on drug response^{39,42}. PharmGKB collects the precise guidelines for the application of precision medicine and pharmacogenomics in clinical practice published by the European Society for Medical Oncology (ESMO), the National Comprehensive Cancer Network (NCCN), the Royal Dutch Association for the Advancement of Pharmacy (DPWG), the Canadian Pharmacogenomics Network for Drug Safety (CPNDS) and the Clinical Pharmacogenetics Implementation Consortium (CPIC)^{43–46}. Hence, we identified essential genes, oncogenic mutations and potential therapeutic targets that could be incorporated into strategies aimed at improving novel drug development and precision medicine in BC.

Results

OncoPrint of genomic alterations according to the Pan-Cancer Atlas. PCA has reported the clinical data of 1084 individuals with BC and it can be visualized in the Genomic Data Commons of the National Cancer Institute and in the cBioPortal^{47,48}. In regard to molecular subtypes and tumor stages, 46% were luminal A, 18% luminal B, 7% Her2-enriched, 16% basal-like and 3% normal-like, whereas 17% were tumor stage 1 (T1), 58% T2 stage, 23% T3 stage and 2% T4 stage (Supplementary Table S1).

Figure 1a shows the frequency mean of genomic alterations per gene set. The frequency mean of the PCA gene set was 1.3, followed by the CS gene set (1.2), the PharmGKB/CGI gene set (1.0), BC driver genes (0.8), and non-cancer genes (0.4) (Supplementary Table S2). Consequently, we performed a multiple comparison of the genomic alteration frequencies using the Bonferroni correction in order to determine statistical significance among gene sets. There were significant differences between BC driver genes and non-cancer genes ($P < 0.001$), the PCA gene set and BC driver genes ($P < 0.001$), and the CS gene set and BC driver genes ($P < 0.001$). Hence, the fact that gene sets of interest (CS and PCA) presented significant differences in the amount of genomic alterations versus BC driver genes could indicate that we are analyzing potentially essential genes in BC. Figure 1b shows the percentage of genomic alterations per type. The most common genomic alterations were mRNA up-regulation (55.8%), CNV amplification (17.1%), and missense mutations (8.4%). Figure 1c shows the ratio of genomic alterations in the 230 genes per sample and molecular subtype. Basal-like had the highest ratio ($n = 33$), followed by Her2-enriched (29), luminal B (24), normal-like (17), and luminal A (15). The ratio of all BC samples was 19.6. Figure 1d shows the ratio of genomic alterations in the 230 genes per sample and tumor stage. T2 stage had the highest ratio (23), followed by T3 (22), T1 (17) and T4 (8). Figure 1e,f show the percentage of genomic alterations per subtype and tumor stage, respectively. mRNA up-regulation and CNV amplification were the most common alterations in all molecular subtypes and tumor stages.

Figure 2 shows the ranking of genes with the highest amount of genomic alterations per molecular subtype and tumor stage. Regarding molecular subtypes, *PIK3CA* was the most altered gene in luminal A, *CCND1* in luminal B, *TP53* in basal-like and normal-like, and *ERBB2* in Her2-enriched (Fig. 2a). Figure 2b–f show genes with the highest ratio of mutations, CNV amplifications, CNV deep deletions, mRNA up-regulations, and mRNA down-regulations per molecular subtype (Tables S3–S7). After Bonferroni correction, we obtained statistically significant differences ($P < 0.05$) regarding CNV amplifications, CNV deep deletions, mRNA up-regulations, and mRNA down-regulations among molecular subtypes. On the other hand, the most altered genes per tumor stage were *PIK3CA* in T1 stage, *TP53* in T2 and T3, and *ERBB2* in T4 (Fig. 2g). Figure 2h–l show genes with the

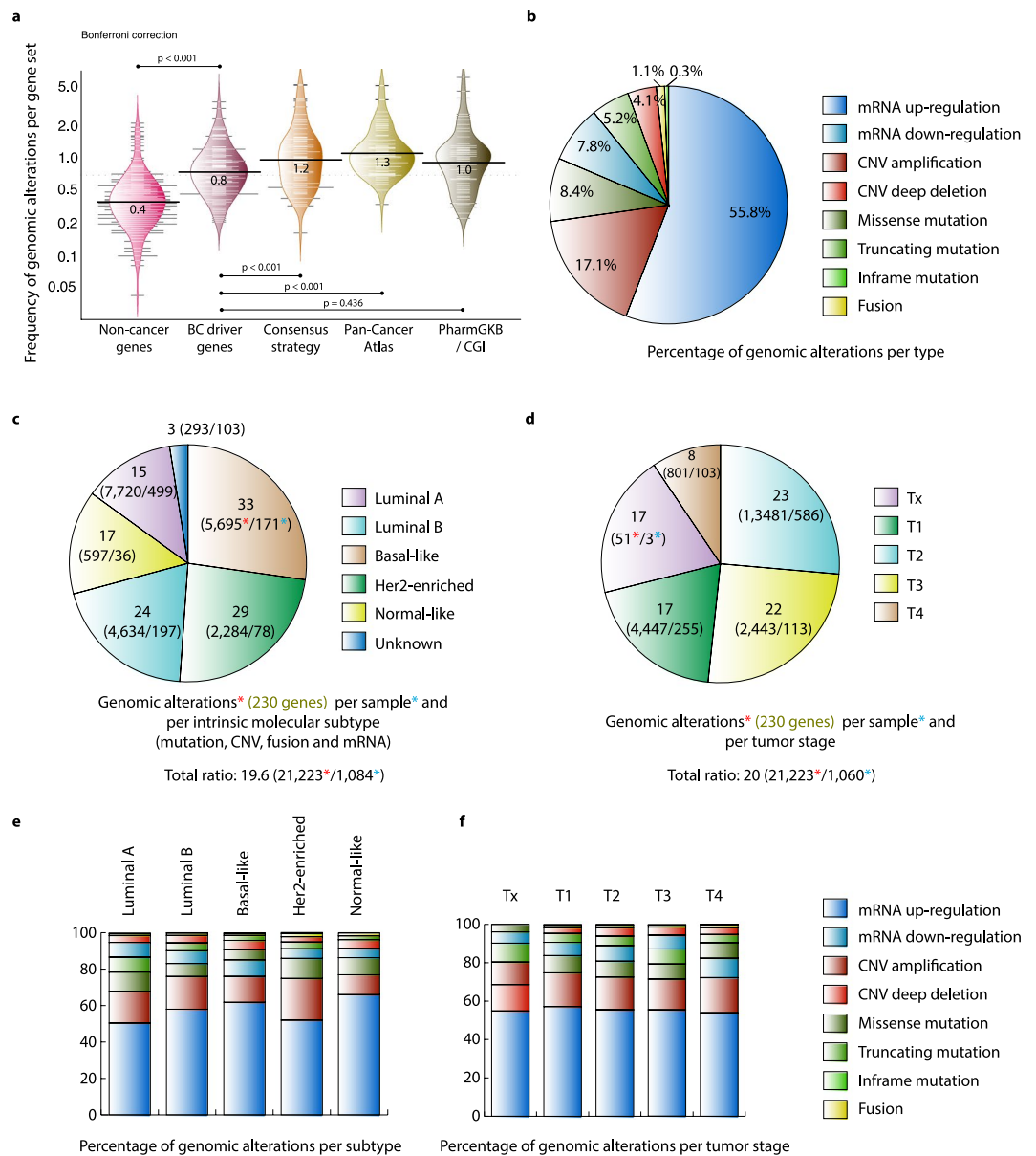


Figure 1. Genomic alterations of the breast cancer cohort according to PCA. **(a)** Frequency of genomic alterations per gene set (non-cancer genes, BC driver genes according to the Network of Cancer Genes, Consensus Strategy, BC genes according to PCA, BC biomarkers according to the PharmGKB and CGI). Bonferroni correction with significant level of $P < 0.05$ and a 95% confidence interval was performed. **(b)** Percentage of genomic alterations per type. **(c)** Ratio of genomic alterations per intrinsic molecular subtype. **(d)** Ratio of genomic alterations per tumor stage. **(e)** Percentage of genomic alterations per type and molecular subtype. **(f)** Percentage of genomic alterations per type and tumor stage.

highest percentage of mutations, CNV amplifications, CNV deep deletions, mRNA up-regulations, and mRNA down-regulations per tumor stage (Tables S8–S12). We found statistically significant differences ($P < 0.05$) regarding all genomic alterations among tumor stages using the Bonferroni correction test.

The first OncoOmics approach was focused on genes with the highest amount of genomic alterations (more than the average). The panoramic landscape of genomic alterations was termed OncoPrint and is shown in Fig. 3a. Putative driver mutations were taken into account for this analysis, discarding passenger mutations (Figure S1 and Supplementary Table S13). Figure 3b,c show circos plots of interactions among molecular subtypes, tumor stages, and genomic alterations of the most altered genes (Supplementary Table S14). Highest amount of fusion genes were in Her2-enriched subtype and T4 stage, highest amount of mRNA down-regulation + CNV deep deletion were in basal-like subtype and T4 stage, highest amount of mRNA up-regulation + CNV amplification were in basal-like subtype and T4 stage, lastly, highest amount of putative driver mutations were in Her2-enriched subtype and T3 stage. As result, the first OncoOmics approach revealed 73 essential genes with highest frequencies of genomic alterations.

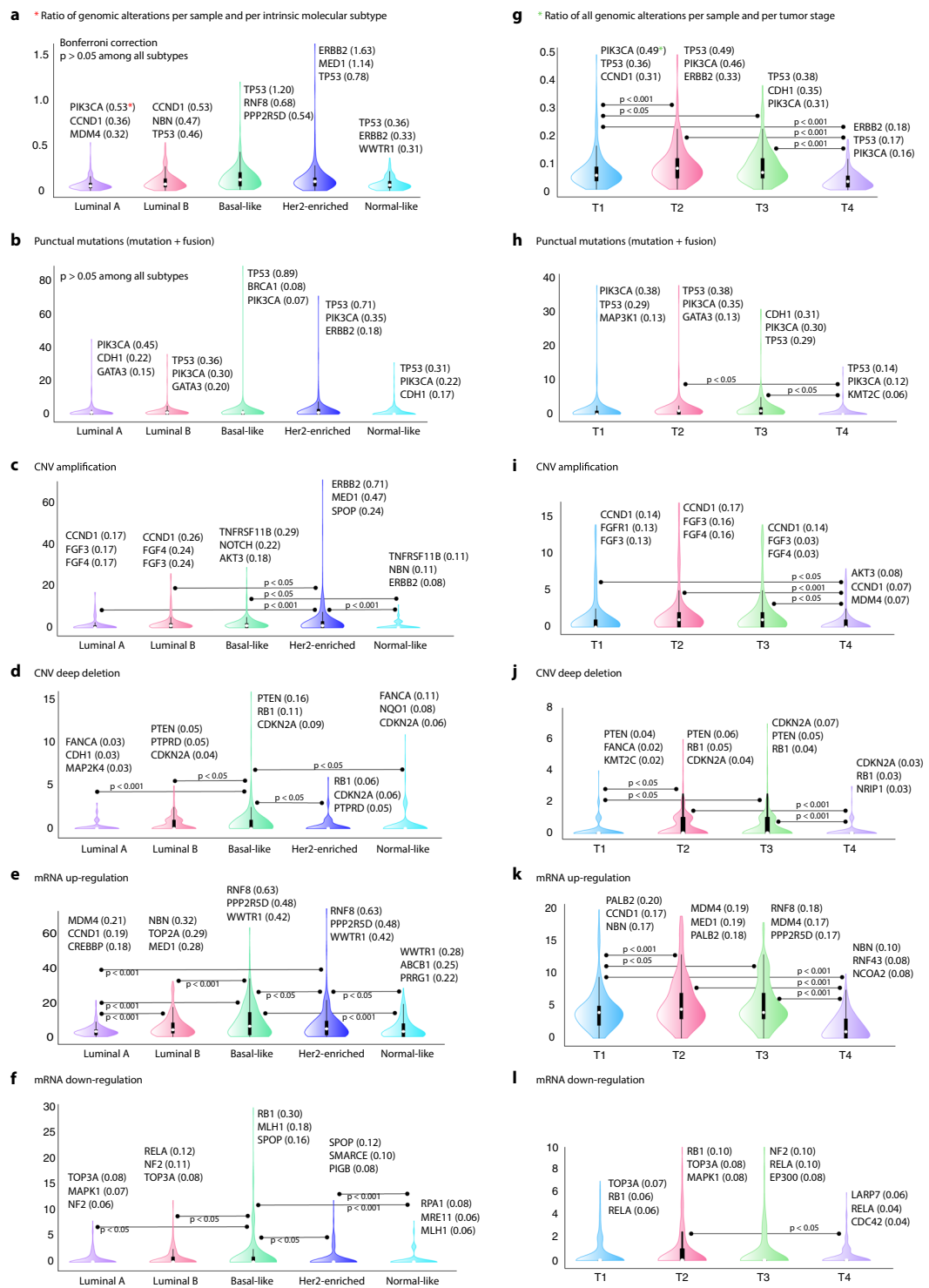


Figure 2. Ranking of genes with the highest amount of genomic alterations per molecular subtype and tumor stage. **(a)** Frequency of genomic alterations (punctual mutations, copy number variants and mRNA expression) per molecular subtype. **(b)** Frequency of genomic alterations per tumor stage. **(c)** Frequency of punctual mutations per molecular subtype. **(d)** Frequency of punctual mutations per tumor stage. **(e)** Frequency of CNV amplifications per molecular subtype. **(f)** Frequency of CNV amplifications per tumor stage. **(g)** Frequency of CNV deep deletions per molecular subtype. **(h)** Frequency of CNV deep deletions per tumor stage. **(i)** Frequency of mRNA up-regulation per molecular subtype. **(j)** Frequency of mRNA up-regulation per tumor stage. **(k)** Frequency of mRNA down-regulation per molecular subtype. **(l)** Frequency of mRNA down-regulation per tumor stage.

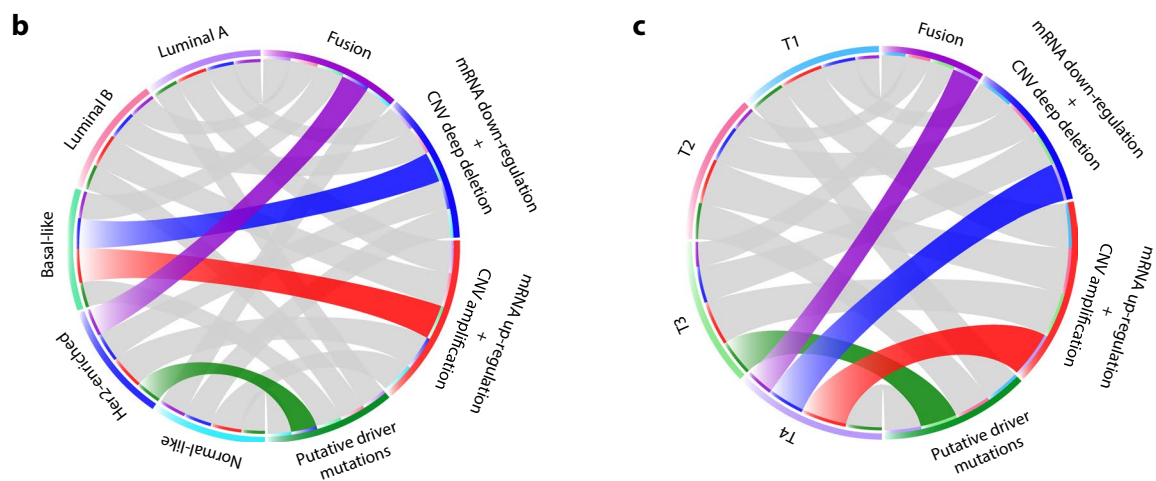
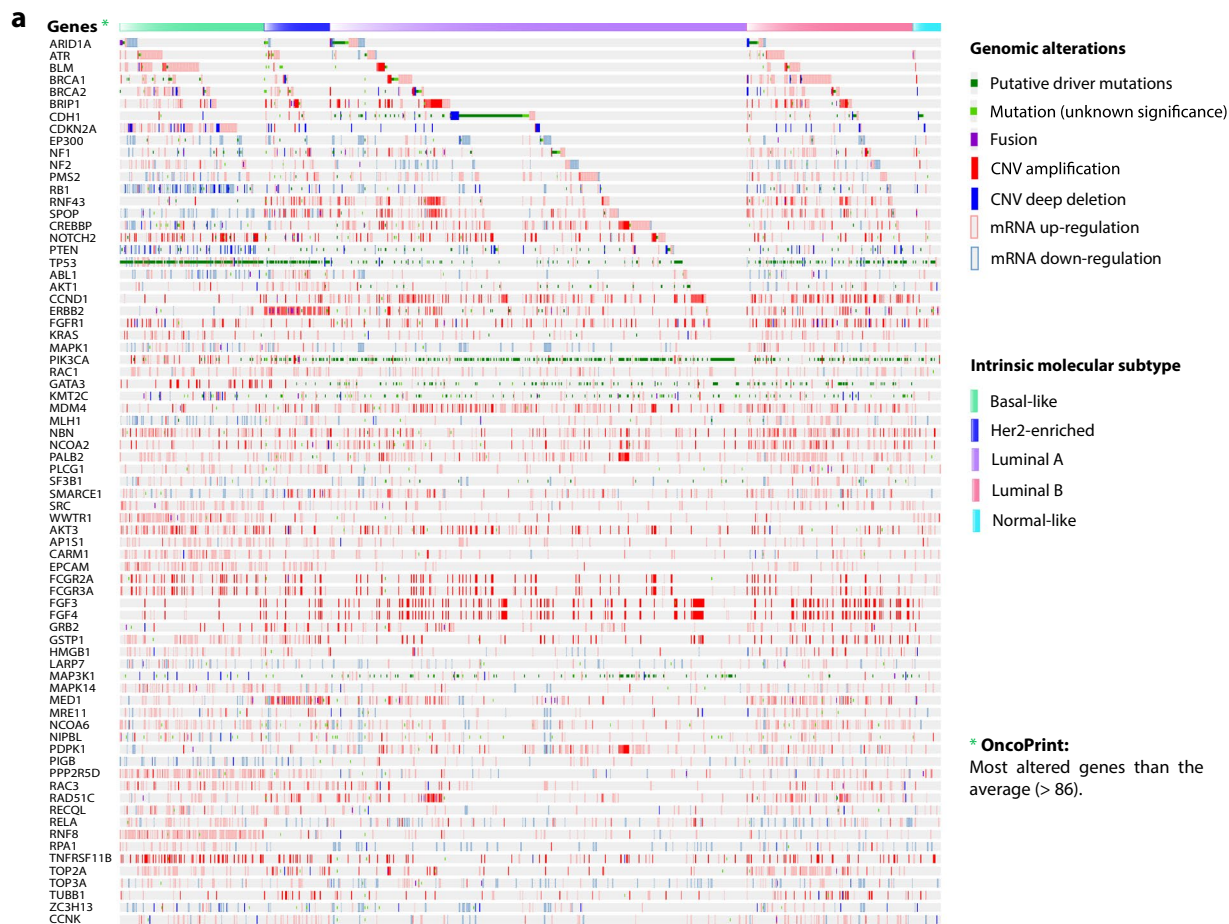


Figure 3. OncoPrint of genomic alterations according to the Pan-Cancer Atlas. (a) OncoPrint of genes with more genomic alterations than the average (>86) per molecular subtype. (b) Circos plot between molecular subtypes and the highest amount of genomic alterations (fusion genes, mRNA down-regulation plus CNV deep deletion, mRNA up-regulation plus CNV amplification, and driver mutations). (c) Circos plot between tumor stages and the highest amount of genomic alterations.

Pathway enrichment analysis. This enrichment analysis was performed using David Bioinformatics Resource to obtain integrated information from the Kyoto Encyclopedia of Genes and Genomes (KEGG)^{49–52}. The enrichment analysis of signaling pathways was carried on in the 230 genes, obtaining more than 50 terms with a Benjamini-Hochberg - false discovery rate (FDR) < 0.01 (Supplementary Table S15). Subsequently, genomic alterations of genes that make up each signaling pathway were analyzed according to the molecular subtype and tumor stage. Figure 4a shows a circos plot correlating molecular subtypes with signaling pathways

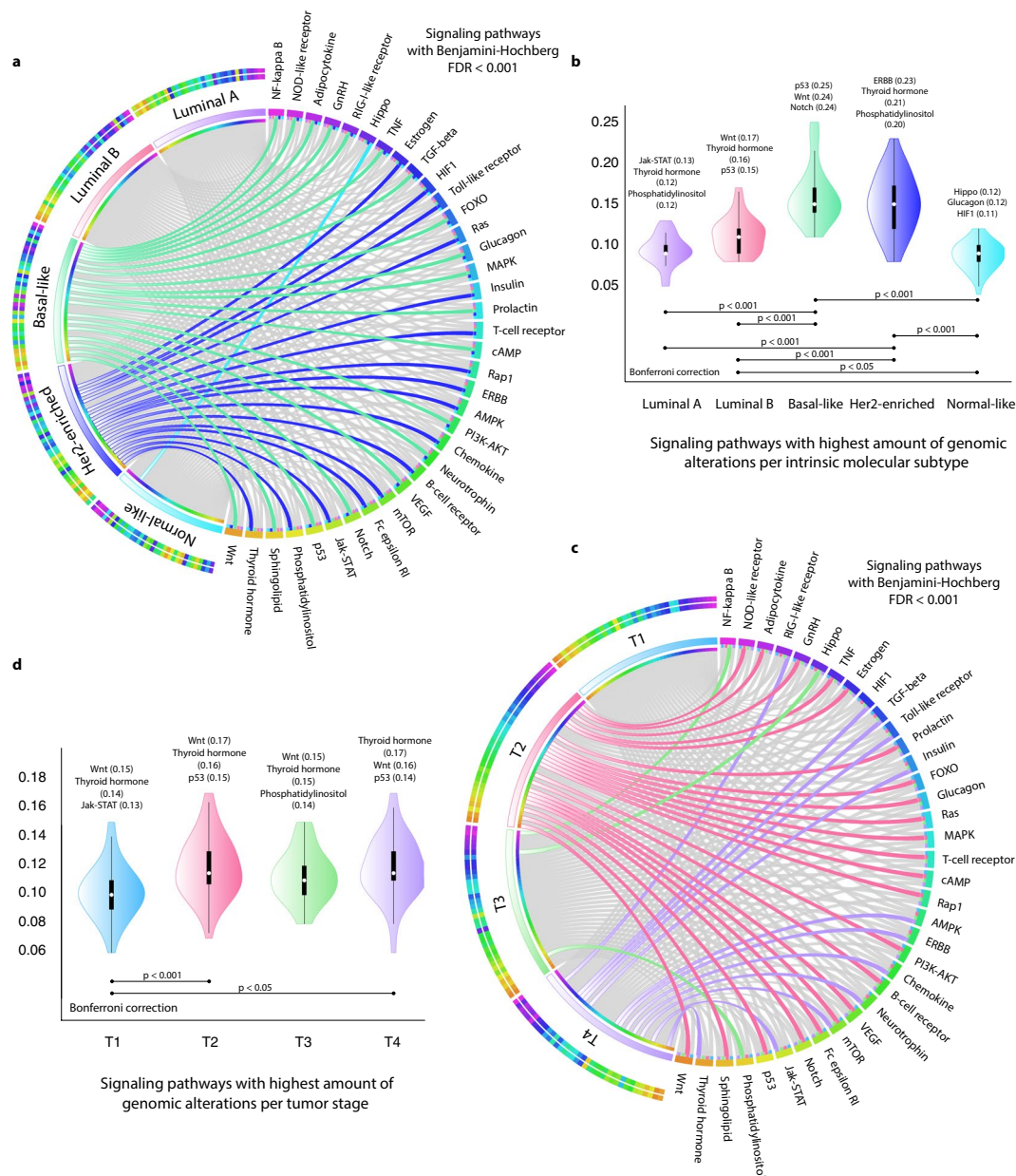


Figure 4. Pathway enrichment analysis per molecular subtype and tumor stage. **(a)** Circos plot between molecular subtypes and the most altered signaling pathways. **(b)** Violin plots showing the frequency of the most altered signaling pathways per molecular subtype. **(c)** Circos plot between tumor stages and the most altered signaling pathways. **(d)** Violin plots showing the frequency of the most altered signaling pathways per tumor stage.

(Supplementary Table S16). NF-kappa β , NOD-like receptor, adipocytokine, GnRH, RIG-like receptor, TNF, TGF β , FOXO, glucagon, MAPK, prolactin, cAMP, PI3K-AKT, neurotrophin, VEGF, notch, p53, sphingolipid and Wnt signaling pathways were more altered in basal-like; estrogen, HIF1, toll-like receptor, ras, insulin, T-cell receptor, rap1, ERBB, AMPK, chemokine, B-cell receptor, mTOR, Fc-epsilon RI, Jak-STAT, phosphatidylinositol and thyroid hormone pathways were more altered in Her2-enriched; and Hippo pathway in normal-like. On the other hand, Fig. 4b shows the ranking of the most altered signaling pathways per molecular subtype. Jak-STAT pathway was more altered in luminal A; Wnt pathway in luminal B; p53 pathway in basal-like; ERBB pathway in Her2-enriched; and Hippo pathway in normal-like (Supplementary Table S17). After Bonferroni correction, we observed statistically significant differences ($P < 0.001$) regarding the amount of genomic alterations in signaling pathways among molecular subtypes.

Figure 4c shows a circos plot correlating tumor stages with signaling pathways according to the frequency of genomic alterations (Supplementary Table S16). NOD-like receptor, adipocytokine, GnRH, TNF, estrogen, prolactin, FOXO, glucagon, ras, MAPK, T-cell receptor, cAMP, rap1, PI3K-AKT, B-cell receptor, VEGF, mTOR, Fc

epsilon RI, NOTCH, p53, sphingolipid and Wnt pathways were more altered in stage T2; NF-kappa β , Hippo and phosphatidylinositol pathways were more altered in T3 stage; and RIG-like receptor, HIF1, TGF β , toll-like receptor, insulin, AMPK, ERBB, chemokine, neurotrophin, mTOR, jak-STAT and thyroid hormone pathways were more altered in T4 stage. On the other hand, Fig. 4d shows the ranking of the most altered signaling pathways per tumor stage. Wnt pathway was more altered in T1, T2 and T3 stages; and thyroid hormone pathway was more altered in T4 stage (Supplementary Table S18). We found statistically significant differences ($P < 0.001$) regarding the amount of genomic alterations in signaling pathways among different tumor stages using the Bonferroni correction test.

Protein-protein interactome network. The second OncoOmics approach was focused on proteins with the highest degree centrality and consensus score in the String PPI network. The PPI network was performed to better understand BC behavior using the String Database and Cytoscape^{53,54}. With the indicated cutoff of 0.9, the final interactome network had 258 nodes conformed by 198 (86%) proteins from the CS, PCA and PharmGKB/CGI sets. Regarding nodes with the highest amount of genomic alterations showed previously in the OncoPrint, 65 (89%) of them integrated this network (Fig. 5a). On the other hand, out of the 258 proteins that make up our String PPI network, 16 (6%) proteins and 18 edges were part of the OncoPPI BC network^{55,56}. The degree centrality made it possible to establish a significant correlation (Spearman test, $P < 0.05$) between our String PPI network and the OncoPPI BC network (Fig. 5b).

Considering degree centrality and consensus scores from our previous study²⁹, there was enrichment among sub-networks (Fig. 5a,b). The degree centrality average in the whole network was 48.8, and out of the OncoPPI BC network was 124.4. Meanwhile, the average of consensus score of the whole network was 0.803, and out of the OncoPPI BC network was 0.885. As result, the second OncoOmics approach revealed 40 proteins with both the highest degree centrality and consensus score, as shown in Supplementary Table S19.

Protein expression analysis. The third OncoOmics approach was focused on proteins with considerable high and low expressions in BC. Figure 6a shows 43 proteins with significant high expression (Z -scores ≥ 2) and low expression (Z -scores ≤ -2) analyzed with the reverse-phase protein array (RPPA) and mass spectrometry, in a cohort of 994 individuals according to TCGA (Supplementary Table S20). On the other hand, the Human Protein Atlas (HPA) presented a map of the human tissue proteome based on tissue microarray-based immunohistochemistry. HPA has analyzed 202 (88%) of the 230 proteins of our study, classifying the protein expression in high, medium, low and non-detected. As results, RAC1, GJB2, MED1, PIK3CA, PIK3R3, FGFR2, HCFC2, MAP2K4, NQO2 and RAC3 were proteins with high/medium expression in normal tissue, and low/non-detected expression in BC tissue. Meanwhile, CDK2, CYP2D6, NCOR1, RRM1, FOXA1 and TOP2A were proteins with high/medium expression in BC tissue, and low/non-detected expression in normal tissue (Fig. 6b and Supplementary Table S21)^{57,58}. As result, the third OncoOmics approach revealed 60 proteins with significant altered expression levels as shown in Tables S20 and S21.

Breast cancer dependency map. The first analysis of the fourth OncoOmics approach consisted in identifying genes that are essential for breast cancer cell proliferation and survival performing systematic loss-of-function screens in a large number of well-annotated cell lines representing the tumor heterogeneity^{19–22}. Figure 7a shows the distribution of dependency scores of 227 genes through DEMETER2, an analytical framework for analyzing genome-scale RNAi loss-of-function screens in 73 BC cell lines (Supplementary Table S22). Our results showed 563 dependencies with at least one score ≤ -1 in 57 (25%) essential genes. At the same time, Fig. 7a shows the distribution of dependency scores of 217 genes through CERES, an analytical framework for analyzing genome-scale CRISPR-Cas9 loss-of-function screens in 28 BC cell lines (Supplementary Table S23). Our results showed 310 dependencies with at least one score ≤ -1 in 34 (16%) essential genes. Figure 7b shows the distribution of dependency scores of DEMETER2 and CERES per molecular subtype. The genome-scale RNAi loss-of-function screens detected 165 (29%) dependencies in 19 Her2-enriched cell lines (ratio = 8.7), 110 (20%) in 13 luminal A cell lines (8.5), 57 (10%) in 7 luminal B cell lines (8.1), and 231 (41%) in 34 basal-like cell lines (6.8), whereas the genome-scale CRISPR-Cas9 loss-of-function screens detected 85 (27%) dependencies in 7 luminal A cell lines (ratio = 12.1), 176 (15%) in 16 basal-like cell lines (11), and 49 (16%) in 5 Her2-enriched cell lines (9.8). Figure 7c shows violin plots of dependencies per molecular subtype. DEMETER2 has detected a greatest number of substantial dependencies in basal-like, followed by Her2-enriched, luminal A and luminal B, whereas CERES has detected a greatest number of substantial dependencies in basal-like, followed by luminal A and Her2-enriched. Figure 7d shows a Venn diagram of 22 strongly selective genes, 26 common essential genes, and 5 strongly selective and common essential genes in breast and other cancer cell lines.

Patient-derived xenografts. The second analysis of the fourth OncoOmics approach consisted in identifying proteins with significant expression in PDXs. According to Woo *et al.*, PDXs are *in vivo* models of human cancer that are useful for translational cancer research and therapy selection for individual patient. We analyzed the 66 strongly selective and common essential genes of BC cell lines using the Jackson Laboratory PDX resource⁵⁹. Figure 7e shows 7 proteins with significant high expression (Z -score ≥ 2) and 33 proteins with significant low expression (Z -scores ≤ -2) with its respective mice model ID. As result, the fourth OncoOmics approach revealed 38 proteins with significant expression in both BC cell lines and patient-derived xenografts (Supplementary Tables S22 and S23).

OncoOmics approaches to reveal essential genes in BC. After analyses of the four OncoOmics approaches (genomic alterations, String PPI network, protein expression and BC dependencies/patient-derived xenografts), we used a Venn diagram to integrate essential genes, termed OncoOmics BC essential genes. Consequently, we could observe 140 essential genes in at least one OncoOmics approach; of them, 92 were

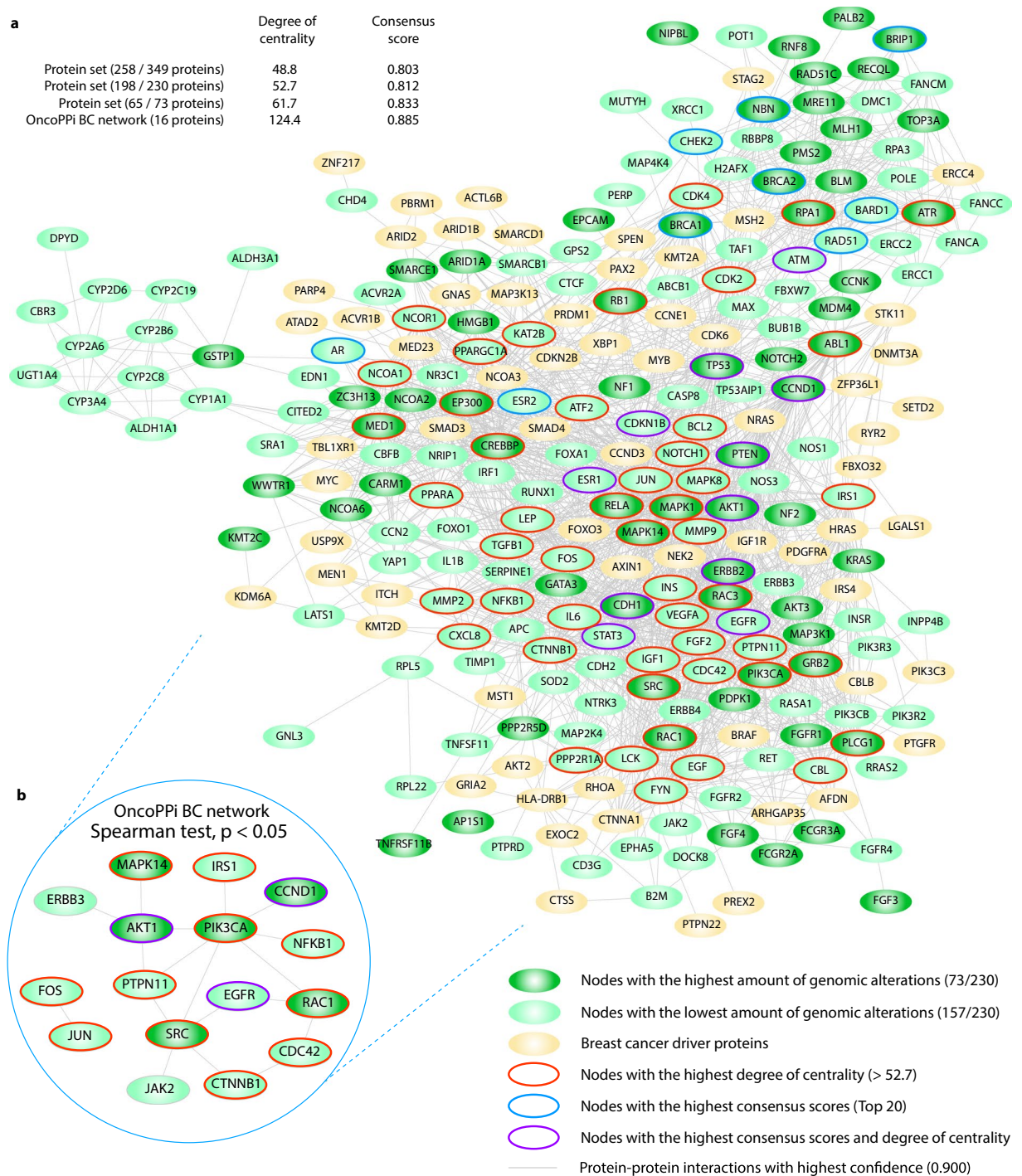


Figure 5. Protein-protein interactome network. (a) Network composed of BC driver genes and genes of our study (PCA gene set, consensus strategy gene set and PharmGKB gene set). (b) Significant correlation ($P < 0.05$) of degree centrality and consensus score between the OncoPPi BC network and our String PPI network.

essential in one OncoOmics approach, 30 were essential in two OncoOmics approaches, 13 were essential in three OncoOmics approaches, and 5 were essential in all OncoOmics approaches as shown in Fig. 8a and Supplementary Table S24.

The 140 OncoOmics BC essential genes were conformed by oncogenes (21%), tumor suppressor genes (24%) and driver genes in other cancer types (59%)⁶⁰ (Fig. 8b). Additionally, some of these OncoOmics BC essential genes were involved in cancer immunotherapy⁶¹, kinome signaling⁶², cell cycle⁶³, DNA repair⁶⁴ and RNA-binding as shown in Fig. 8c and Supplementary Table S25⁶⁵.

Figure 8d shows a circos plot detailing the correlation between 48 (34%) OncoOmics BC essential genes and hallmarks of cancer. Suppression of growth was promoted by *AKT1*, *CTNNB1*, *PTEN*, *RB1* and *TP53*; escaping

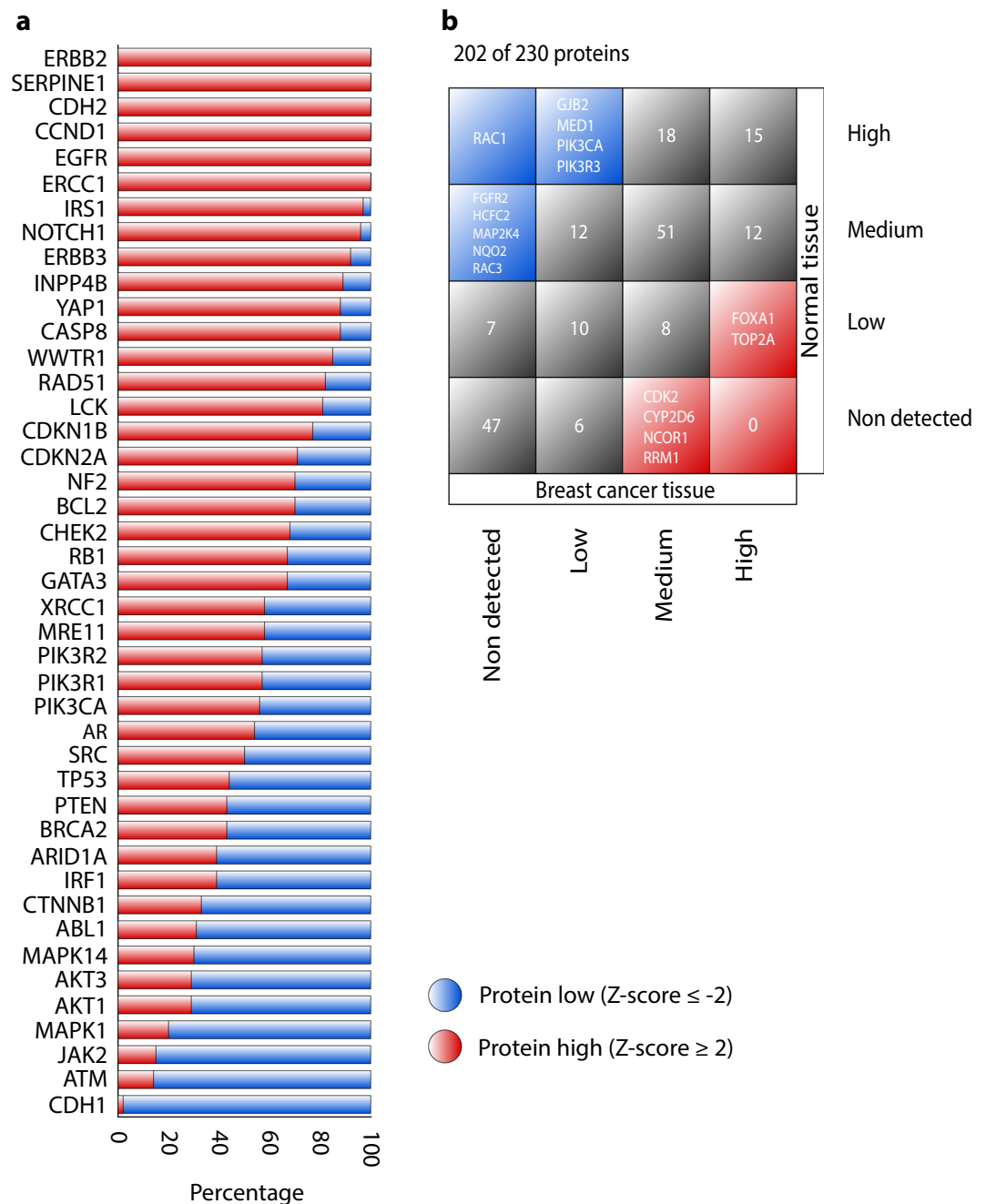


Figure 6. Protein expression analyses. (a) Proteins ($n = 43$) with alterations in the expression levels. Low expression proteins with Z-score ≤ -2 and high expression proteins with Z-score ≥ 2 according to TCGA. (b) Comparison of protein expression levels ($n = 202$) by immunohistochemistry between BC tissue and normal tissue according to The Human Protein Atlas.

immune response to cancer was promoted by *CTNNB1*, *EGFR* and *RAC1*; cell replicative immortality was promoted by *CTNNB1*, *KRAS* and *NOTCH1*; tumor promoting inflammation was promoted by *KRAS*; metastasis was promoted by *ABL1*, *CTNNB1*, *EGFR*, *KRAS*, *RAC1* and *RB1*; angiogenesis was promoted by *ABL1*, *CTNNB1*, *EGFR*, *KRAS*, *NOTCH1* and *RAC1*; genome instability was promoted by *ABL1* and *RB1*; escaping programmed cell death was promoted by *AKT1*, *CTNNB1*, *EGFR*, *NOTCH1*; change of cellular energetics was promoted by *ABL1*, *AKT1*, *CTNNB1*, *EGFR*, *KRAS*, *NOTCH1*, *PTEN*, *RB1* and *TP53*; finally, proliferative signaling was promoted by *ABL1*, *AKT1*, *CTNNB1*, *EGFR*, *KRAS*, *NOTCH* and *RAC1* (Supplementary Table S26).

Enrichment map of the OncoOmics BC essential genes. Figure 8e shows the enrichment map of the 140 OncoOmics BC essential genes. g:Profiler searches for a collection of genes representing GO terms, pathways and disease phenotypes⁶⁶. The most significant GO: biological processes with a FDR < 0.001 was positive

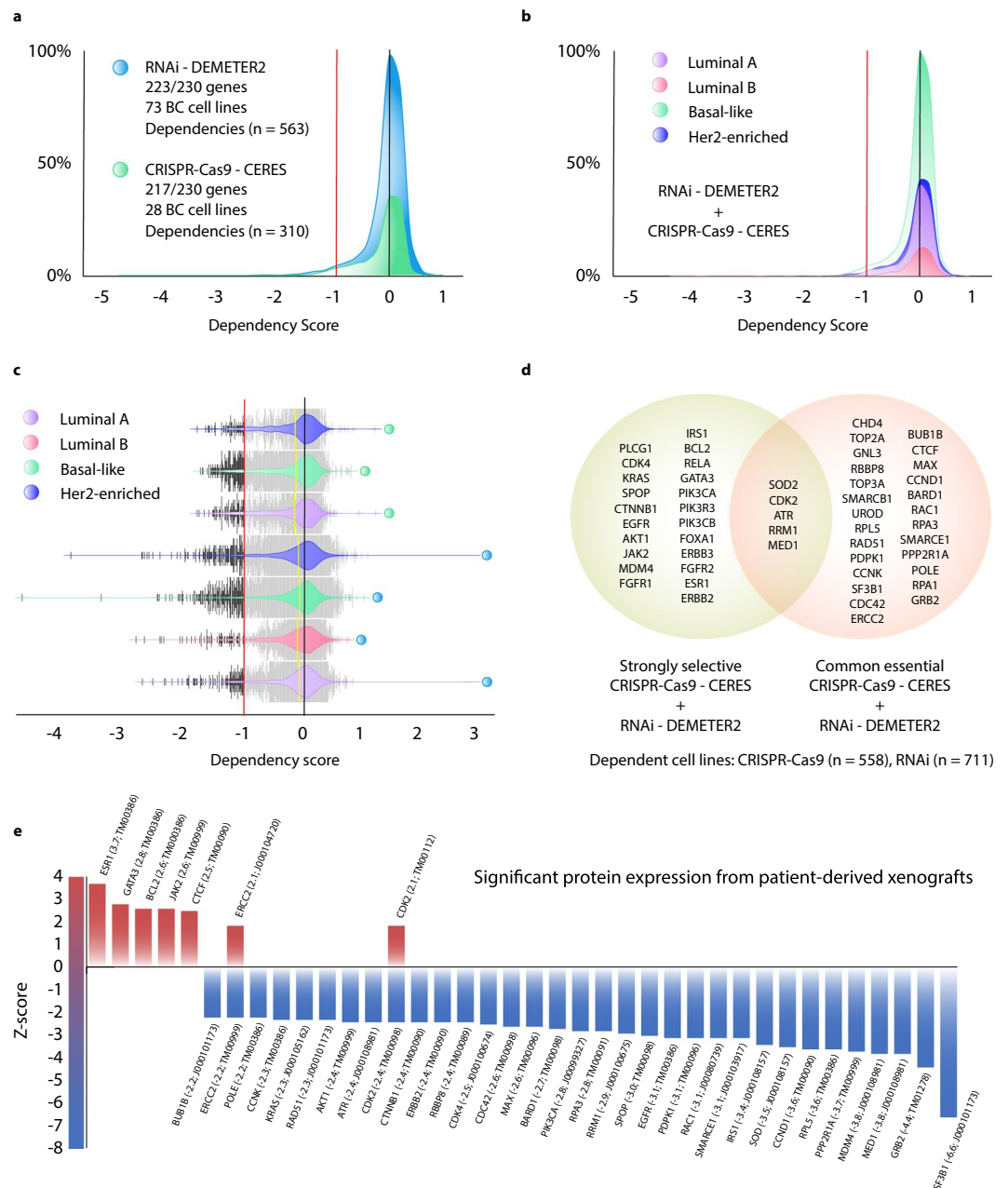


Figure 7. BC dependency maps in cell lines and patient-derived xenografts. **(a)** Dependency score of gene sets using RNAi DIMETER2 and CRISPR-Cas9 CERES algorithms in BC cell lines. **(b)** Dependency score of BC gene sets per molecular subtypes. **(c)** Violin plots of dependencies per molecular subtypes. All substantial dependencies < -1 are in black. **(d)** Venn diagram of strongly selective and common essential genes in all cancer cell lines. **(e)** Significant protein expression from patient-derived xenografts.

regulation of macromolecule metabolic process (Supplementary Table S27); the most significant GO: molecular function was phosphatidylinositol 3-kinase activity (Supplementary Table S28); the most significant Reactome pathway was generic transcription pathway (Supplementary Table S29)⁶⁷; additionally, the most relevant disease, according to the Human Phenotype Ontology, was breast carcinoma (Supplementary Table S30)⁶⁸. Subsequently, g:Profiler annotations were analyzed with the EnrichmentMap software and visualized using Cytoscape, in order to generate network interactions of the most relevant GO: biological processes (Supplementary Fig. S2) and Reactome pathways (Fig. 9) related to immune system, tyrosine kinase, cell cycle and DNA repair pathways^{54,66}.

Clinical trials. Figure 10 and Supplementary Table S31 details the current status of clinical trials regarding OncoOmics BC essential proteins, according to the Open Targets Platform⁶⁹. There are 98 drugs that are being analyzed in 2,904 clinical trials in 28 of 140 OncoOmics BC essential proteins (Fig. 10a). The top 10 drugs with the highest number of clinical trials in process or completed were paclitaxel (370), trastuzumab (315), docetaxel

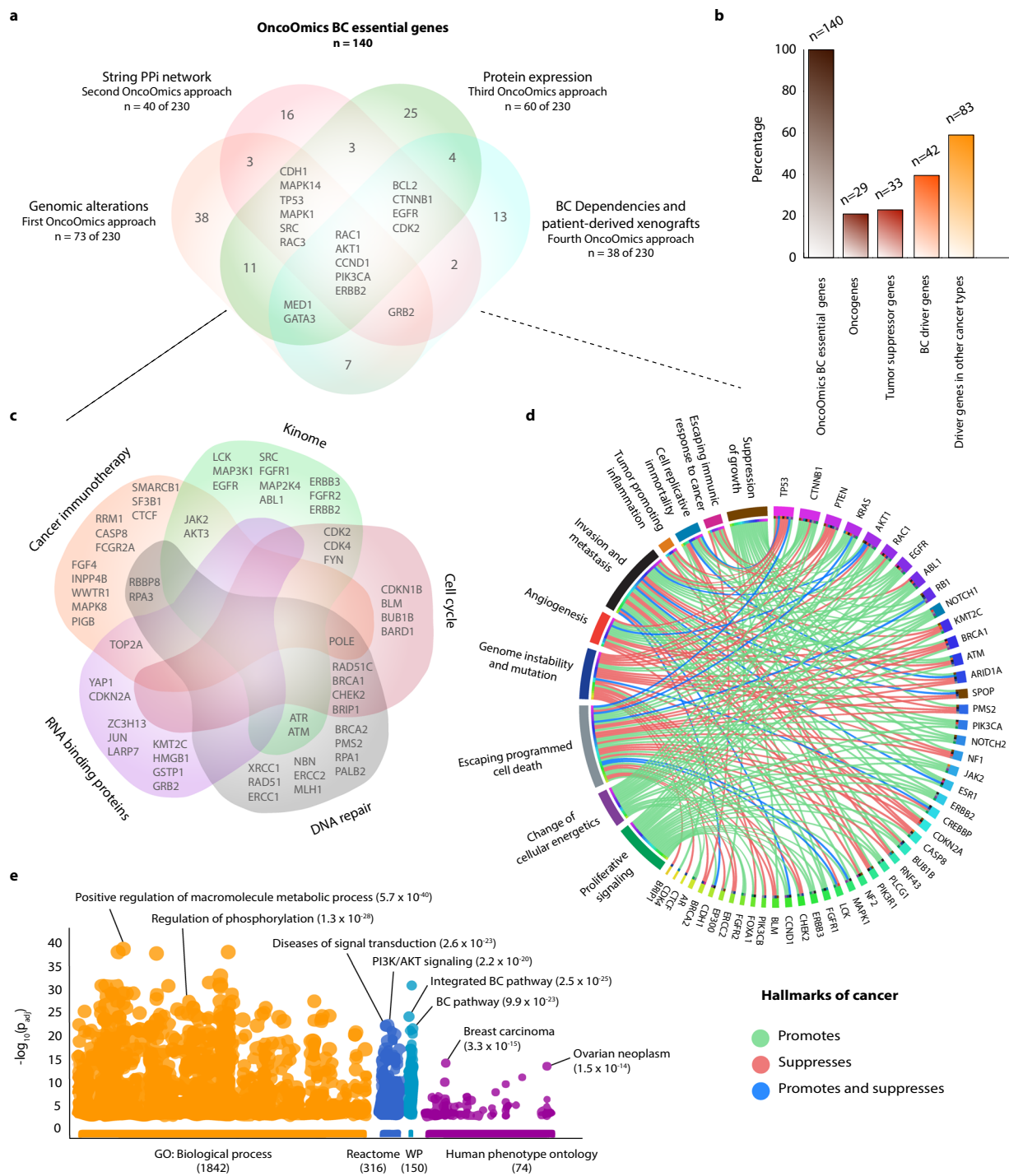


Figure 8. OncoOmics BC essential genes. **(a)** Venn diagram of the most essential genes per genomics approach (genomic alterations, String PPI network, protein expression, and BC dependencies/patient-derived xenografts). **(b)** Percentage of oncogenes, tumor suppressor genes and driver genes in other cancer types. **(c)** Venn diagram of the most essential genes related to cancer immunotherapy, kinome signaling, cell cycle, DNA repair and RNA-binding proteins. **(d)** Circos plot of genes with hallmarks of cancer. **(e)** Most significant g:Profiler features of the OncoOmics BC essential genes according to GO: biological processes, Reactome pathways, WikiPathways and the human phenotype ontology.

(262), doxorubicin (204), gemcitabine (196), lapatinib (152), tamoxifen (131), fulvestrant (129), bevacizumab (120) and neratinib (110). Regarding drugs, 94% were antagonists, 79% were small molecules, and 35% were protein kinases as shown in Fig. 10b–d, respectively. Additionally, drugs with the highest number of clinical trials in phases 3 and 4 were paclitaxel (111), docetaxel (105), trastuzumab (80), tamoxifen (69) and doxorubicin (60) as shown in a Sankey plot detailed in Fig. 10e.

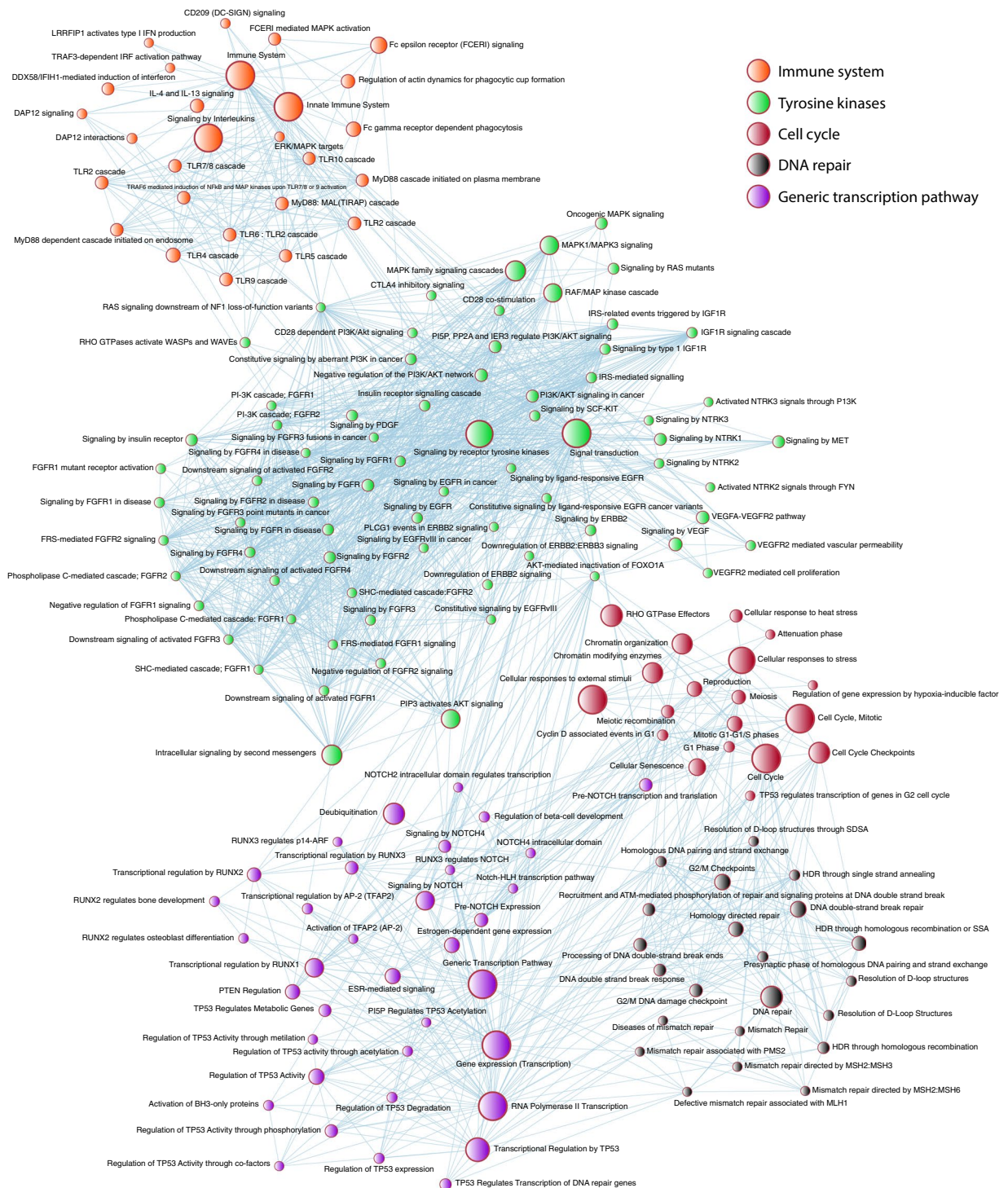


Figure 9. Pathway enrichment analysis of the OncoOmics BC essential genes using g:Profiler and EnrichmentMap. Most significant Reactome pathways related to immune system, kinase signaling, cell cycle, DNA repair and genetic transcription.

Precision medicine. Precision oncology focuses on matching the most effective and safe treatment based on the ‘omics’ profile of each individual or population^{70,71}. However, the identification of driver mutational events remains the biggest challenge⁷². There are some consortiums and studies that have robustly identified variants associated with BC. Tamborero *et al.* detailed a compendium of 62 somatic and 398 germline validated oncogenic mutations in 14 OncoOmics BC essential genes (Supplementary Table S32)³⁸. Huang *et al.* identified 87 pathogenic germline variants in 22 OncoOmics BC essential genes⁷³ (Supplementary Table S33). Long *et al.*^{74,75}, Cai *et al.*⁷⁶, Michailidou *et al.*⁷⁷, and the Breast Cancer Association Consortium performed genome-wide association

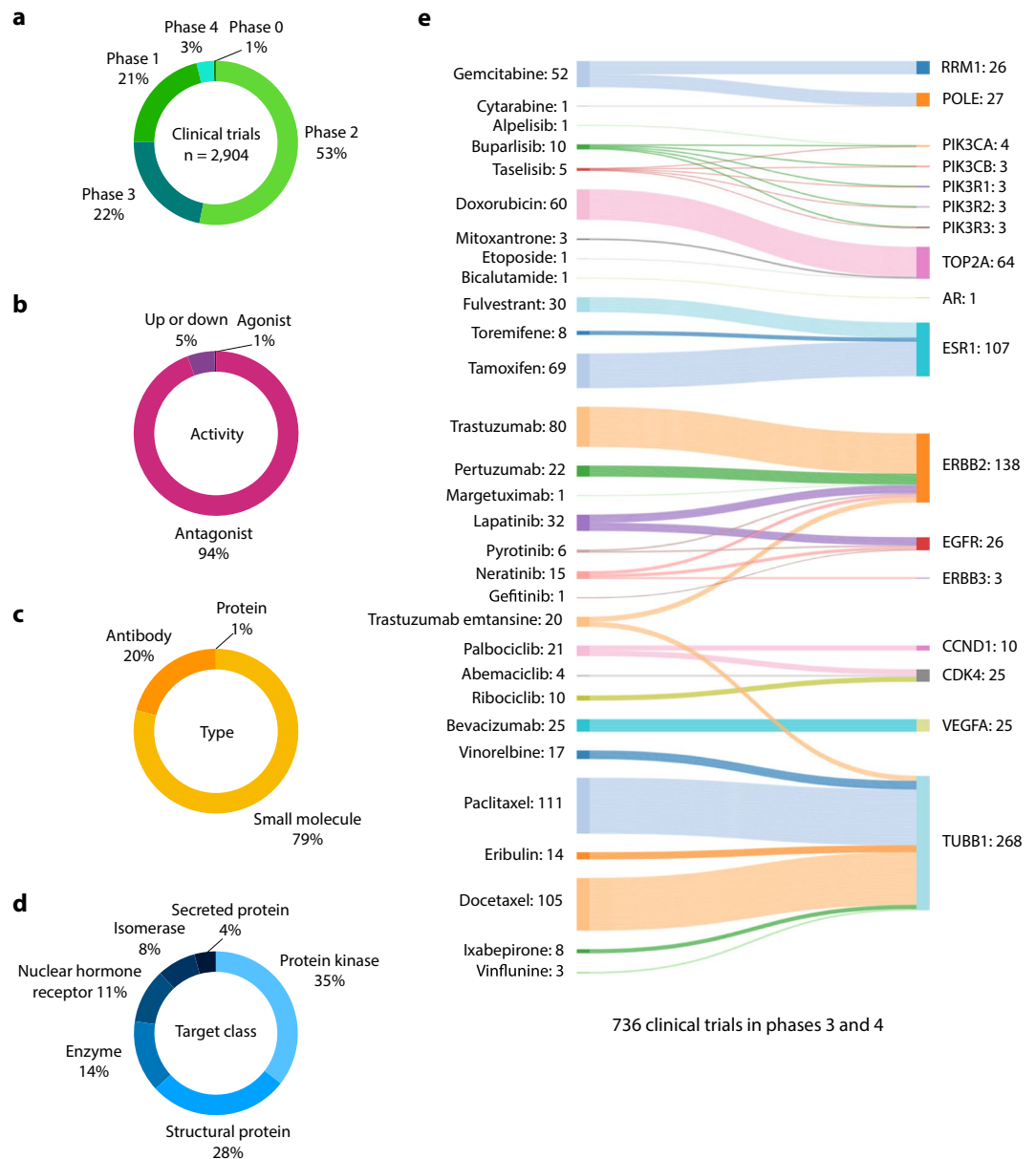


Figure 10. Current status of clinical trials in the OncoOmics BC essential proteins. **(a)** Clinical trials per phase. **(b)** Clinical trials per activity. **(c)** Clinical trials per type. **(d)** Clinical trials per target class. **(e)** Correlation of drugs with proteins in advanced stages of clinical trials (3 and 4) using a Sankey plot.

studies identifying 172 germline variations related to BC development (Supplementary Table S34). The Precision Medicine Knowledgebase (PreMedKB) detailed a compendium of 2791 germline variants in 7 OncoOmics BC essential genes (Supplementary Table S35)⁷¹. PharmGKB enriched clinical guidelines with 59 well-known clinical annotations related to 29 OncoOmics BC essential genes (Supplementary Table S36)^{42,78,79}. Finally, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium identified 19 non-coding somatic mutations and 17 coding somatic mutations in BC (Supplementary Table S37)⁶.

Regarding the Ensembl Variant Effect Predictor⁸⁰, 1,102 of 3,565 variants were processed, being 24% intron variants, 16% missense variants, 15% downstream gene variants, 10% stop gained, 7% upstream gene variants, 7% NMD transcript variants, 4% splice region variants, 4% 3' untranslated region variants, and 2% splice acceptor variants (Supplementary Table S38).

Consequently, based on the aforementioned somatic and germline oncogenic variants, the Cancer Genome Interpreter and PreMedKB platforms provided a comprehensive *in silico* list of biological therapy drugs aimed to improve precision medicine in breast cancer (Fig. 11, Tables S35 and S39).

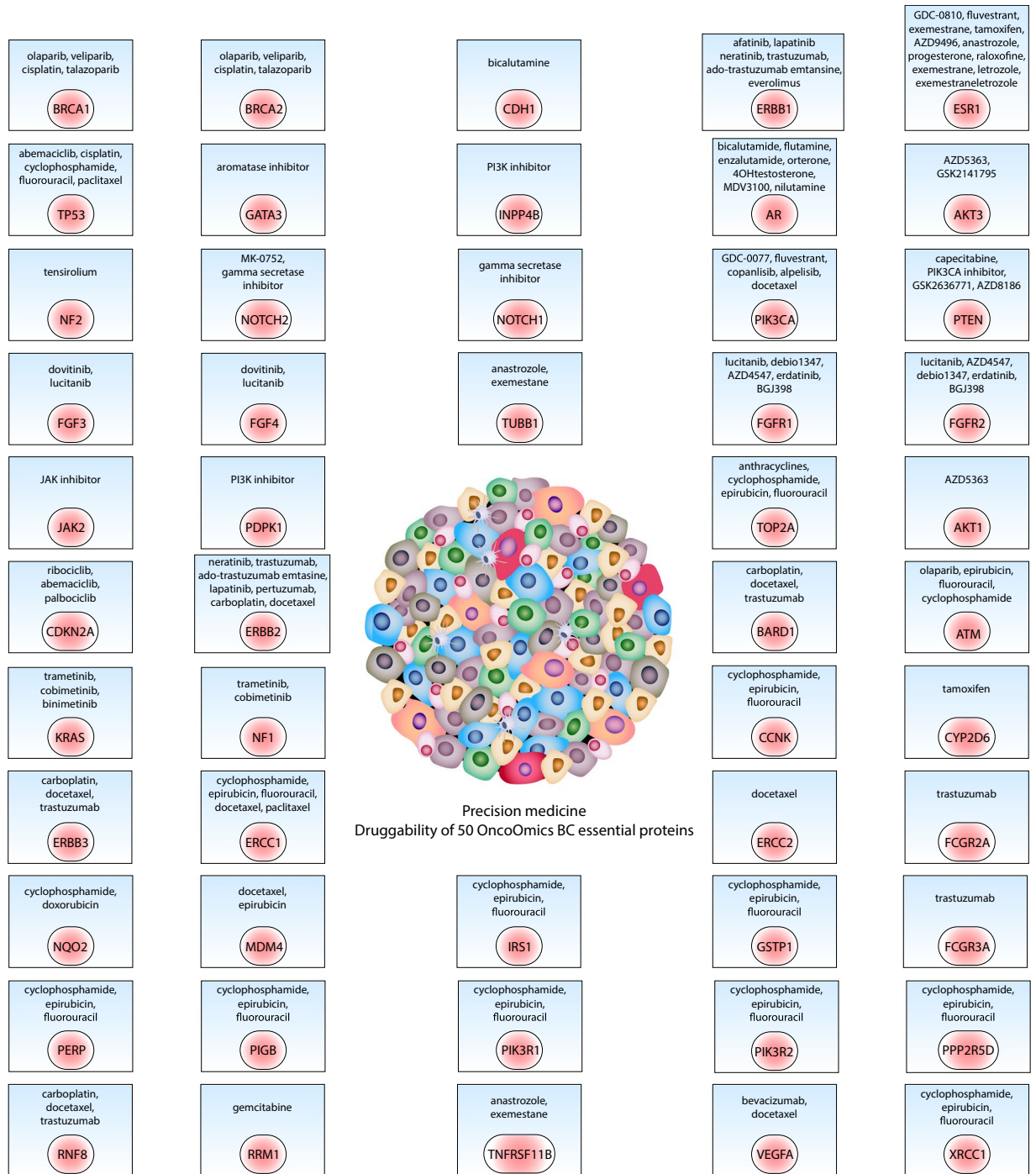


Figure 11. Precision medicine. Interaction between drugs and 50 OncoOmics BC essential proteins.

Discussion

In this study we reveal essential genes in breast cancer through an OncoOmics strategy that analyzes genomic alterations, PPI networking, protein expression, dependency maps and patient-derived xenografts in three gene sets. The first gene set was taken from our previous study where we developed a Consensus Strategy that was proved to be highly efficient in the recognition of BC pathogenic genes^{29,41}. The second gene set was taken from several studies of PCA, which provides a panoramic view of the oncogenic processes that contributes to BC pathogenesis^{3,13,31–37}. The third gene set was taken from the CGI and PharmGKB. On the one hand, the CGI flags genomic biomarkers of drug response with different levels of clinical relevance³⁸. On the other hand, PharmGKB collects clinical annotations applied in BC patients and taken from the NCCN, ESMO, CPNDS, DPWG and CPIC guidelines^{43–46}. Finally, the compendium of these 230 genes was analyzed through four different OncoOmics approaches.

The first OncoOmics approach consisted in the analysis of genomic alterations using the PCA data^{47,48}. The frequency mean of genomic alterations in the CS (1.2) and PCA (1.3) gene sets were significantly higher than both the non-cancer genes (0.4) and the well-known BC driver genes (0.8), with a significant Bonferroni correction of $P < 0.001$. This means that the analyzed set of genes might be strongly associated with BC (Fig. 1a).

The most common genomic alterations in a cohort of 994 individuals were mRNA up-regulation, CNV amplification and missense mutations. Regarding molecular subtypes, basal-like showed the highest amount of genomic alterations. *PIK3CA* was the most altered gene in luminal A, *CCND1* in luminal B, *TP53* in basal-like and normal-like, and *ERBB2* in Her2-enriched (Fig. 2a). A multiple comparison through Bonferroni correction found significant differences ($P < 0.05$) of CNV amplifications, CNV deep deletions, mRNA up-regulations, and mRNA down-regulations among molecular subtypes (Figs. 2c–f). Regarding tumor stages, T2 showed the highest amount of genomic alterations. *PIK3CA* was the most altered gene in T1, *TP53* in T2 and T3, and *ERBB2* in T4 (Fig. 2g). Bonferroni correction found significant differences ($P < 0.05$) in punctual mutations, CNV amplifications, CNV deep deletions, mRNA up-regulations, and mRNA down-regulations among tumor stages (Fig. 2h–l). Lastly, the first OncoOmics approach revealed that 73 essential genes presented frequencies of alteration higher than the average (Fig. 3a)^{3,13,31–37}.

Subsequently, the enrichment analysis of signaling pathways was carried on taking into account all genomic alterations in the 230 genes using David Bioinformatics Resource and KEGG^{49,52}. Pathways with the highest amount of genomic alterations per molecular subtype were Jak-STAT in luminal A, Wnt in luminal B, p53 in basal-like, ERBB in Her2-enriched and Hippo in normal-like. Bonferroni correction showed significant differences ($P < 0.05$) among several subtypes as shown in Fig. 4b. On the other hand, pathways with the highest amount of genomic alterations per tumor stage were Wnt in T1, T2 and T3, and thyroid hormone in T4. Bonferroni correction showed significant differences ($P < 0.05$) comparing T1 with T2 and T4 as shown in Fig. 4d.

Regarding previously mentioned signaling pathways, Jak-STAT is involved in inflammatory response, stem cell maintenance, and hematopoiesis⁸¹. The Wnt signaling pathway actively functions in embryonic development and helps in homeostasis in mature tissues by regulating cell survival, migration, proliferation, and polarity⁸². The p53 signaling pathway plays an essential role into inhibition of growth, programmed cell death, cell migration and angiogenesis⁸³. The ERBB pathway mediates signal transduction events that control cell survival, migration and proliferation in BC⁸⁴. The Hippo pathway plays important roles in tumor suppression and immune response. However, alterations in this pathway are involved in the BC tumorigenesis and metastasis⁸⁵. Lastly, the thyroid hormone pathway plays an important role as regulator of growth and metabolism. Nevertheless, dysfunction of the T3 hormone promotes cancer progression in mammary epithelial cells⁸⁶.

The second OncoOmics approach was focused on proteins with the highest degree centrality and consensus score in the String PPI network. In accordance with Li *et al.* and Ivanov *et al.*^{56,87}, PPI with therapeutic significance can be revealed by the integration of cancer proteins into networks. PPI regulate essential oncogenic signals to cell proliferation and survival, and thus, represents potential targets for drug development and drug discovery. Regarding our networking analysis, the final interaction network consisted in 258 nodes with a degree centrality average of 48.8 and a consensus score average of 0.803²⁹; the sub-network integrated by 198 of 230 nodes had 52.7 of degree centrality and 0.812 of consensus scoring; finally, the sub-network integrated by 65 of 73 proteins with the highest amount of genomic alterations had 61.7 of degree centrality and 0.833 of consensus score. Hence, a sub-network of nodes with the highest amount of genomic alterations presented a highest degree centrality and consensus score, suggesting that there is strong correlation between these proteins and BC. Additionally, the oncogenomics validation showed a substantial correlation between our String PPI network (Fig. 5a) and the OncoPPI BC network (Fig. 5b), identifying 16 nodes strongly associated with BC²⁹. The second OncoOmics approach revealed 40 essential proteins with the highest degree centrality and consensus scoring.

The third OncoOmics approach was focused on proteins with significant high and low expression in BC proteome. More than 500 proteins have been identified as strongly involved in oncogenesis. Loss of expression, overexpression or expression of dysfunctional proteins contribute to uncontrolled tumor growth, causing chromosomal rearrangements, gene amplification and uncontrolled methylation⁸⁸. Regarding our 230 proteins, 43 showed significant high (Z -scores ≥ 2) and low (Z -scores ≤ -2) expression according to TCGA⁸⁹ (Fig. 6a); and 16 proteins showed opposite expression between healthy and affected tissues after microarray-based immunohistochemistry according to the Human Protein Atlas (Fig. 6b)^{57,58}. The compendium of 60 proteins with significant high and low expressions made up the third OncoOmics approach.

The fourth OncoOmics approach was related to the BC dependency map in cell lines and patient-derived xenografts. According to Tsherniak *et al.*, mutations that trigger the growth of cancer cells also confer specific vulnerabilities that normal cells lack, and these dependencies are compelling therapeutic targets¹⁹. The cancer dependency map identifies essential genes in proliferation and survival of well-annotated cell lines through systematic loss-of-function screens^{19–22}. On the one hand, DETEMER2 analyzed the genome-scale RNAi loss-of-function screens, and on the other hand, CERES analyzed the genome-scale CRISPR-Cas9 loss-of-function screens as shown in Fig. 7a. In addition to the loss-of-function screens in a large number of well-annotated BC cell lines, the patient-derived xenografts are *in vivo* models of human tumors engrafted in a mouse host and emerging as a powerful tool for understanding tumor hallmarks and predicting drug efficacy⁹⁰. Consequently, we validated the genomic expression of the strongly selective and common essential genes (dependencies in BC cell lines) in breast tumors from PDXs provided by the Jackson Laboratory⁵⁹. The fourth OncoOmics approach was made up of 38 essential proteins in BC (Fig. 7e).

Subsequently, the compendium of essential genes per approach reveals the 140 OncoOmics BC essential genes (Fig. 8a). *RAC1*, *AKT1*, *CCND1*, *PIK3CA* and *ERBB2* were essential genes in all the OncoOmics approaches. *CDH1*, *MAPK14*, *TP53*, *MAPK1*, *SRC* and *RAC3* showed genomic alterations, highest degree centrality and consensus scores in the String PPI network, and significant protein expression. *GRB2* showed genomic alterations,

highest degree centrality and consensus scores in the String PPI network, and substantial relevance in BC cell lines and PDXs. *MED1* and *GATA3* showed genomic alterations, significant protein expression, and considerable relevance in BC cell lines and PDXs. Lastly, *BCL2*, *CTNNB1*, *EGFR* and *CDK2* showed significant protein expression, highest degree centrality and consensus scores in the String PPI network, and substantial relevance in BC cell lines and PDXs.

Relevant studies worldwide have identified OncoOmics BC essential genes. For instance, genome-wide association studies performed by the Breast Cancer Association Consortium showed that *BRCA2*, *CHEK2*, *ESR1*, *FGFR2*, *MDM4* and *PIK3R3* carry germline variants associated with BC development^{74–77}. According to Bailey *et al.*, identifying molecular cancer drivers is critical for precision oncology³². Their final consensus list was conformed by 29 BC driver genes, of them, 22 were OncoOmics BC essential genes (*AKT1*, *ARID1A*, *BRCA1*, *CASP8*, *CDH1*, *CDKN1B*, *CTCF*, *ERBB2*, *FOXA1*, *GATA3*, *KMT2C*, *KRAS*, *MAP2K4*, *MAP3K1*, *NCOR1*, *NF1*, *PIK3CA*, *PIK3R1*, *PTEN*, *RB1*, *SF3B1* and *TP53*). According to Gonzalez-Perez *et al.*, the IntOGen-mutation platform summarizes somatic mutations involved in tumorigenesis⁹¹. Their final consensus list was conformed by 99 mutational BC driver genes, of them, 34 were identified by the OncoOmics strategy (*TP53*, *PIK3CA*, *KMT2C*, *GATA3*, *CDH1*, *MAP3K1*, *ESR1*, *PTEN*, *AKT1*, *NCOR1*, *ARID1A*, *MAP2K4*, *FOXA1*, *NF1*, *ERBB2*, *RB1*, *SF3B1*, *ERBB3*, *CTCF*, *PIK3R1*, *ATM*, *FGFR2*, *BRCA1*, *CASP8*, *CREBBP*, *BRCA2*, *CDKN2A*, *KRAS*, *CDKN1B*, *NOTCH2*, *MAX*, *MDM4*, *EGFR* and *JAK2*). Finally, the PCAWG Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas reported an integrative analysis of 2,658 whole-cancer genomes across 38 tumor types⁹². Regarding breast cancer, PCAWG identified 27 mutational BC driver genes, of them, 15 were OncoOmics BC essential genes (*TP53*, *PIK3CA*, *MAP3K1*, *KMT2C*, *NOTCH2*, *SF3B1*, *PTEN*, *ARID1A*, *MAP2K4*, *AKT1*, *CTCF*, *FOXA1*, *RB1*, *CDKN2A* and *ATM*).

According to Reimand *et al.*, g:Profiler lets us know the enrichment map of the 140 OncoOmics BC essential genes⁶⁶. The most significant GO: biological process was the positive regulation of macromolecule metabolic process, the GO: molecular function was phosphatidylinositol 3-kinase activity, the Reactome pathway was generic transcription pathway, and the most significant Human Phenotype Ontology term was breast carcinoma⁶⁸. Subsequently, the most relevant network interactions of the GO: biological process and the Reactome pathways were related to immune system, tyrosine kinase, cell cycle and DNA repair terms (Figs. 9 and S2)^{54,66}.

There is currently great enthusiasm about immunotherapeutic strategies to treat BC⁹³. The first approval of an immune checkpoint blockade agent for treatment of BC came in March 2019 when the anti-PD-L1 antibody atezolizumab was approved to be used with nab-paclitaxel in triple-negative BC patients^{94,95}. 16 OncoOmics BC essential genes were associated with immunotherapy^{61,96} as shown in Fig. 8C. Kinases have been recognized as therapeutic targets due to their druggability and play a critical role in cell migration, differentiation, growth and survival⁹⁷. 15 OncoOmics BC essential genes were kinomes⁶². Cell cycle comprises a series of events that drive cell division and DNA replication⁹⁸. 12 OncoOmics BC essential genes were involved in cell cycle⁶³. DNA repair signaling pathways work in concert to correct DNA lesions and maintain genome stability. Nevertheless, a defective DNA repair machinery causes BC development and progression⁹⁹. 17 OncoOmics BC essential genes were involved in DNA repair⁶⁴. RBPs are key players in post-transcriptional events and are emerging as critical modulators in BC^{100–102}. Bioinformatics profiling of tumors have revealed the landscape of alterations in RBPs across cancer types^{103–106}. Lastly, 10 OncoOmics BC essential genes were RBPs⁶⁵.

Regarding clinical trials reported on the OncoOmics BC essential proteins, the Open Targets Platform is an available resource for the integration of genomics and chemical data to aid systematic drug target identification and prioritization⁶⁹. There are 98 drugs that are being analyzed in 2,904 clinical trials in 28 of 140 OncoOmics BC essential proteins. Additionally, there are 30 drugs involved in 736 clinical trials in phases 3 and 4. The top five drugs with the highest number of clinical trials in process or completed are paclitaxel (111), docetaxel (105), trastuzumab (80), tamoxifen (69), and doxorubicin (60)⁶⁹ (Fig. 10e).

Tumor-related genomic alterations predict tumor prognosis, drug response, and toxicity¹⁰⁷. Precision medicine provides patients with the most appropriate diagnostics and targeted therapies based on the 'omics' profile and other predictive and prognostic tests¹⁰⁸. Therefore, precision medicine aims to deliver the right medicine to the right patient at the right dose at the right time, minimizing adverse effects and maximizing drug efficacy^{109,110}. Figure 11 shows comprehensive interactions between directed biological drugs and 50 OncoOmics BC essential proteins aimed to improve precision medicine in breast cancer.

In conclusion, since BC is a complex and heterogeneous disease, the study of different OncoOmics approaches is an effective way to reveal essential genes to better understand the molecular landscape of processes behind oncogenesis, and to develop better therapeutic treatments focused on pharmacogenomics and precision medicine.

Methods

OncoPrint of genomic alterations according to the Pan-Cancer Atlas. PCA has reported the clinical data of 1084 individuals with BC and it can be visualized in the Genomic Data Commons of the National Cancer Institute (<https://gdc.cancer.gov/>) and in the cBioPortal (<http://www.cbioportal.org/>)^{47,48}. The clinical annotations were age, pTNM classification, tumor type, tumor stage and race/ethnicity.

Additionally, PCA has reported genomic alterations (mRNA up-regulation, mRNA down-regulation, CNV amplification, CNV deep deletion, putative driver mutations and fusion gene) of 994 individuals. Putative mutations were analyzed through exome sequencing, CNVs through the Genomic Identification of Significant Targets in Cancer (GISTIC 2.0)^{111,112}, and mRNA expression through RNA Seq V2. We analyzed five gene sets in order to compare the frequency mean of genomic alterations among them. The first gene set (n = 177) was integrated by the non-cancer genes¹¹³. We calculated the OncoScore of non-cancer genes, taking out all genes from our study. The second gene set (n = 119) was the BC driver genes, according to The Network of Cancer Genes⁶⁰. The third gene set (n = 84) was taken from our previous study where we developed a Consensus Strategy of prioritized

genes related to BC pathogenesis²⁹. The fourth gene set ($n = 85$) was made up of genes associated with BC development, according to several PCA studies^{31,32,114}. Finally, the fifth gene set ($n = 91$) consisted of BC biomarkers and druggable enzymes taken from PharmGKB and the CGI (Supplementary Table S2)^{38,39,42}.

The OncoOmics approaches were performed in 230 genes conformed by the CS, PCA and PharmGKB/CGI gene sets. We calculated the percentage and ratio of genomic alterations per intrinsic molecular subtype and tumor stage, and then we established a ranking of genes with the highest amount of genomic alterations (OncoPrint). The OncoPrint conformed the first OncoOmics approach.

Pathway enrichment analysis. The enrichment analysis of signaling pathways was performed using David Bioinformatics Resource to obtain integrated information from KEGG^{49–52}. It was carried on in the 230 genes, taking into account terms with a significant FDR < 0.01 . After that, genomic alterations that comprise each signaling pathway were analyzed, taking into account the molecular subtype and tumor stage of individuals from PCA. Circos plots and violin plots were designed to visualize all data. Lastly, in order to compare the ratio of genomic alterations among subtypes and tumor stages, normalization was carried out dividing the number of genomic alterations by the number of individuals per subtype and tumor stage. Regarding molecular subtypes, 499 individuals were luminal A, 197 were luminal B, 171 were basal-like, 78 were Her2-enriched and 36 were normal-like, and regarding tumor stage, 255 were T1, 586 were T2, 113 were T3, and 103 were T4.

Protein-protein interactome network. The PPI network with a highest confidence cutoff of 0.9 and zero node addition was created using the String Database, which takes into account predicted and known interactions⁵³. The confidence scoring is the approximate probability that a predicted link exists between two enzymes in the same metabolic map, whereas the degree centrality of a node means the number of edges the node has to other nodes in a network. The centrality indexes calculation and network visualization were analyzed through the Cytoscape software⁵⁴. Proteins with the highest degree centrality, consensus score and sub-networks were differentiated by colors in the PPI network. On the other hand, OncoPPI (<http://oncoppi.emory.edu/>) reports the development of a cancer-focused PPI network, identifying more than 260 high-confidence cancer-associated PPI^{55,56}. In addition, the OncoPPI BC network consisted of 16 proteins and 18 PPI experimentally analyzed in BC cell lines^{55,56}. The correlation of the degree centrality by means of Spearman P-value test between our String PPI network and the OncoPPI BC network allowed for the validation of all the high-confidence BC-focused PPI analyzed in cell lines²⁹. Lastly, proteins with the highest degree centrality and consensus scoring made up the second OncoOmics approach.

Protein expression analysis. TCGA has reported the protein expression data of 994 individuals with BC through RPPA and mass spectrometry by the Clinical Proteomic Tumor Analysis Consortium (CPTAC), and it can be visualized in the cBioPortal^{47,48}. We analyzed the protein expression of 230 protein where Z-scores ≥ 2 mean a significant high protein expression and Z-scores ≤ -2 mean a significant low protein expression.

On the other hand, the Human Protein Atlas (<https://www.proteinatlas.org/>) explains the diverse molecular signatures of proteomes in human tissues based on an integrated 'omics' approach that involves quantitative transcriptomics and tissue microarray-based immunohistochemistry^{58,88,115}. We compared the protein expression levels (high, medium, low and non-detected) of our 230 proteins between normal and BC tissues. Finally, all genes with the altered protein expression made up the third OncoOmics approach.

Breast cancer dependency map. The DepMap project (<https://depmap.org/portal/>) is collaboration between the Broad Institute and the Wellcome Sanger Institute. Multiple genetic or epigenetic changes provide cancer cells with specific vulnerabilities that normal cells lack. Even though the landscape of genomic alterations has been extensively studied to date, we have limited understanding of the biological impact of these alterations in the development of specific tumor vulnerabilities, which triggers a limited use of precision medicine in the clinical practice worldwide. Therefore, the main goal of DepMap is to create a comprehensive preclinical reference map connecting tumor features with tumor dependencies to accelerate the development of precision treatments^{19–22}.

In order to identify essential genes for BC cell proliferation and survival, DepMap performed systematic loss-of-function screens in a large number of well-annotated BC cell lines representing the tumor heterogeneity and their molecular subtypes. The DEMETER2 algorithm was applied to analyze genome-scale RNAi loss-of-function screens in 73 BC cell lines and 711 cancer cell lines, whereas the CERES algorithm was applied to analyze genome-scale CRISPR-Cas9 loss-of-function screens in 28 BC cell lines and 558 cancer cell lines^{20,22}. In addition to existing cell lines, the Cancer Cell Line Encyclopedia (CCLE) project will greatly expand the collection of characterized cell lines to improve precision treatments¹¹⁶.

Regarding dependency scores, a lower score means that a gene is more likely to be dependent in a specific cancer cell line. A score of 0 means that a gene is not essential, whereas a score of -1 corresponds to the median of all common essential genes. A strongly selective gene means that its dependency is at least 100 times more likely to have been sampled from a skewed distribution than a normal distribution. A common essential gene is when in a pan-cancer screen its gene ranks in the top most depleting genes in at least 90% of cell lines¹⁹. All genes or proteins with a dependency score ≤ -1 were subsequently analyzed with patient-derived xenografts.

Patient-derived xenografts. The Jackson Laboratory PDX resource (<http://tumor.informatics.jax.org/mtbwi/pdxSearch.do>) comprises 455 PDX models originating from 34 different primary sites⁵⁹. Even though, we analyzed expression levels of strongly selective and common essential proteins in breast cancer obtained from the analysis of BC dependency map in cell lines. Significant high protein expression has a Z-score ≥ 2 and significant low protein expression has a Z-scores ≤ -2 .

Enrichment map of the OncoOmics BC essential genes. The pathway enrichment analysis gives scientists curated interpretation of gene lists generated from genome-scale experiments⁶⁶. The OncoOmics essential genes in BC were analyzed by using g:Profiler (<https://biit.cs.ut.ee/gprofiler/>) in order to obtain significant annotations (FDR < 0.001) related to GO terms, pathways, networks and disease phenotypes. Subsequently, g:Profiler annotations were analyzed with the EnrichmentMap software in order to generate network interactions of the most relevant GO: biological processes and Reactome pathways, and these networks were visualized using Cytoscape^{54,66}.

Clinical trials. The Open Targets Platform (<https://www.targetvalidation.org>) is comprehensive and robust data integration for access to and visualization of drugs involved in clinical trials associated with BC proteins, detailing its phase, status, type and target class⁶⁹. In addition, we created a Sankey plot to better understand which drugs are involved in the most advanced phases (3 and 4) of clinical trials.

Precision medicine. Precision oncology focuses on matching the most effective treatment based on the 'omics' profile of each individual or population^{70,71}. The CGI (<https://www.cancergenomeinterpreter.org/home>) flags genomic biomarkers of drug response with different levels of clinical relevance³⁸. Huang *et al.* and the Pan-Cancer Atlas project conducted the largest investigation of pathogenic germline variants in cancer⁷³. Long *et al.*^{74,75}, Cai *et al.*⁷⁶, and Michailidou *et al.*⁷⁷, performed genome-wide association studies identifying germline variations related to BC development. PreMedKB (<http://www.fudan-pgx.org/premedkb/index.html#/home>) is a bioinformatics tool that facilitates the interpretation of the clinical meaning of a patient's genetic variants⁷¹. PharmGKB (<https://www.pharmgkb.org/>) collected complete guidelines for application of pharmacogenomics in clinical practice, according to several consortiums worldwide^{43–46}. Finally, PCAWG Consortium (<https://dcc.icgc.org/>) revealed an integrative analysis of genomic alterations in coding and non-coding regions^{6,92}.

Based on the aforementioned somatic and germline oncogenic variants we performed two analyses. On the one hand, we analyzed the consequence type of variants with the Ensembl Variant Effector Predictor (<https://www.ensembl.org/Multi/Tools/VEP?db=core>), which is a powerful toolset for the annotation of genomic variants in coding and non-coding regions⁸⁰. On the other hand, we analyzed oncogenic variants through the Cancer Genome Interpreter and PreMedKB platforms to provide a comprehensive *in silico* list of biological therapy drugs^{38,71}.

Statistical analyses. We performed a multiple comparison using the Bonferroni correction test (significant level of $P < 0.05$ and a 95% confidence interval) to analyze: 1) significant differences of genomic alteration frequencies among non-cancer genes, BC driver genes, Consensus Strategy, Pan-Cancer Atlas and PharmGKB/CGI genes; 2) significant differences of genomic alteration frequencies among intrinsic molecular subtypes and tumor stages; 3) significant differences of genomic alteration frequencies of signaling pathways among molecular subtypes and tumor stages. A significant correlation of the degree centrality between the String PPI network and the OncoPPI BC network was performed using the Spearman p-value test with a $P < 0.05$. The significant high and low protein expression in human tissues and patient-derived xenografts was considered using the Z-score. Z-score ≥ 2 means significant high protein expression and Z-scores ≤ -2 means significant low protein expression. Lastly, the enrichment map of OncoOmics BC essential genes was performed using g:Profiler that determines the most significant GO: biological processes, GO: molecular functions, Reactome pathways, WikiPathways, KEGG pathways and human phenotype ontology with a false discovery rate < 0.001 .

Data availability

All data generated or analyzed during this study are included in this published article (and its Supplementary Information files).

Received: 15 July 2019; Accepted: 2 March 2020;

Published online: 24 March 2020

References

1. Espinal-Enríquez, J., Fresno, C., Anda-Jáuregui, G. & Hernández-Lemus, E. RNA-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Sci. Rep.* **7**, 1760 (2017).
2. Guerrero, S. *et al.* Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. 1–8. <https://doi.org/10.1038/s41598-018-32264-x> (2018).
3. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**, 305–320.e10 (2018).
4. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Canc. Clin. Oncol.* <https://doi.org/10.3322/caac.21492> (2018).
5. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* (80-). <https://doi.org/10.1126/science.1133427> (2006).
6. Rheinbay, E., Nielsen, M. M., Abascal, F. & Wala, J. A. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*. <https://doi.org/10.1038/s41586-020-1965-x> (2020).
7. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature*. <https://doi.org/10.1038/nature12634> (2013).
8. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. <https://doi.org/10.1038/nature12912> (2014).
9. Porta-Pardo, E. *et al.* Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat. Methods*. <https://doi.org/10.1038/nmeth.4364> (2017).
10. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
11. Lu, C. *et al.* Patterns and functional implications of rare germline variants across 12 cancer types. *Nat. Commun.* <https://doi.org/10.1038/ncomms10086> (2015).
12. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.3080> (2015).

13. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* <https://doi.org/10.1016/j.celrep.2018.03.050> (2018).
14. Oltean, S. & Bates, D. O. Hallmarks of alternative splicing in cancer. *Oncogene.* <https://doi.org/10.1038/onc.2013.533> (2014).
15. Stricker, T. P. *et al.* Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression. *PLoS Genet.* <https://doi.org/10.1371/journal.pgen.1006589> (2017).
16. Lawrence, R. T. *et al.* The Proteomic Landscape of Triple-Negative Breast Cancer. *Cell Rep.*, <https://doi.org/10.1016/j.celrep.2015.03.050> (2015).
17. Sogawa, K. *et al.* Identification of a novel serum biomarker for pancreatic cancer, C4b-binding protein α -chain (C4BPA) by quantitative proteomic analysis using tandem mass tags. *Br. J. Cancer.* <https://doi.org/10.1038/bjc.2016.295> (2016).
18. Rubio-Perez, C. *et al.* In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell* **27**, 382–396 (2015).
19. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell.* <https://doi.org/10.1016/j.cell.2017.06.010> (2017).
20. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* <https://doi.org/10.1038/ng.3984> (2017).
21. Stransky, N. *et al.* Pharmacogenomic agreement between two cancer cell line data sets. *Nature.* <https://doi.org/10.1038/nature15736> (2015).
22. McFarland, J. M. *et al.* Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* <https://doi.org/10.1038/s41467-018-06916-5> (2018).
23. Shah, P. *et al.* Integrated Proteomic and Glycoproteomic Analyses of Prostate Cancer Cells Reveal Glycoprotein Alteration in Protein Abundance and Glycosylation. *Mol. Cell. Proteomics.* <https://doi.org/10.1074/mcp.M115.047928> (2015).
24. Bernard, P. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
25. Kumar, R., Sharma, A. & Tiwari, R. K. Application of microarray in breast cancer: An overview. *J. Pharm. Bioallied Sci.* **4**, 21–6 (2012).
26. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
27. López-Cortés, A. *et al.* Breast cancer risk associated with gene expression and genotype polymorphisms of the folate-metabolizing MTHFR gene: a case-control study in a high altitude Ecuadorian mestizo population. *Tumor Biol.* **36**, 6451–6461 (2015).
28. Prat, A., Ellis, M. J. & Perou, C. M. Practical implications of gene-expression-based assays for breast oncologists. *Nature Reviews Clinical Oncology* **9**, 48–57 (2012).
29. López-Cortés, A. *et al.* Gene prioritization, communality analysis, networking and metabolic integrated pathway to better understand breast cancer pathogenesis. *Sci. Rep.* **8**, 16679 (2018).
30. López-cortés, A. *et al.* Mutational Analysis of Oncogenic AKT1 Gene Associated with Breast Cancer Risk in the High Altitude Ecuadorian Mestizo Population. **2018** (2018).
31. Huang, K. L. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* 355–370, <https://doi.org/10.1016/j.cell.2018.03.039> (2018).
32. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018).
33. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* 1–19, <https://doi.org/10.1016/j.immuni.2018.03.023> (2018).
34. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell.* <https://doi.org/10.1016/j.cell.2018.02.052> (2018).
35. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).
36. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018).
37. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* <https://doi.org/10.1016/j.cels.2018.03.002> (2018).
38. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. 1–8, <https://doi.org/10.1186/s13073-018-0531-8> (2018).
39. Thorn, C. F., Klein, T. E. & Altman, R. B. PharmGKB: The pharmacogenomics knowledge base. *Methods Mol. Biol.* **1015**, 311–320 (2013).
40. Tejera, E. *et al.* Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis. *BMC Med. Genomics* **10**, 50 (2017).
41. Cabrera-andrade, A. Gene Prioritization through Consensus Strategy, Enrichment Methodologies Analysis, and Networking for Osteosarcoma Pathogenesis. *Int. J. Mol. Sci.* **21**, 1–21 (2020).
42. Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine.* <https://doi.org/10.1002/wsbm.1417> (2018).
43. Ross, C. J. D. *et al.* The Canadian Pharmacogenomics Network for Drug Safety: a model for safety pharmacology. *Thyroid* **20**, 681–7 (2010).
44. Saito, Y. *et al.* CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clinical Pharmacology and Therapeutics* **99**, 36–37 (2016).
45. Swen, J. J. *et al.* Pharmacogenetics: From bench to byte an update of guidelines. *Clin. Pharmacol. Ther.* **89**, 662–673 (2011).
46. European Society for Medicinal Oncology. Breast Cancer: A guide for patients. *European Society for Medical Oncology* (2018).
47. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* <https://doi.org/10.1158/2159-8290.CD-12-0095> (2012).
48. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* <https://doi.org/10.1126/scisignal.2004088> (2013).
49. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
50. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
51. Antonov, A. V., Schmidt, E. E., Dietmann, S., Krestyaninova, M. & Hermjakob, H. R spider: a network-based analysis of gene lists by combining signaling and metabolic pathways from Reactome and KEGG databases. *Nucleic Acids Res.* **38**, W78–83 (2010).
52. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**, 29–34 (1999).
53. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–52 (2015).
54. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).
55. Ivanov, A. A. *et al.* The OncoPPI Portal: an integrative resource to explore and prioritize protein-protein interactions for cancer target discovery. *Bioinformatics* 1–9, <https://doi.org/10.1093/bioinformatics/btx743> (2017).
56. Li, Z. *et al.* The OncoPPI network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nat. Commun.* **8** (2017).
57. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* (80-). <https://doi.org/10.1126/science.aan2507> (2017).
58. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* (80-), <https://doi.org/10.1126/science.1260419> (2015).

59. Woo, X. Y. *et al.* Genomic data analysis workflows for tumors from patient-derived xenografts (PDXs): Challenges and guidelines. *BMC Med. Genomics*. <https://doi.org/10.1186/s12920-019-0551-2> (2019).
60. Repana, D. *et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.* <https://doi.org/10.1186/s13059-018-1612-0> (2019).
61. Patel, S. J. *et al.* Identification of essential genes for cancer immunotherapy. *Nature*. <https://doi.org/10.1038/nature23477> (2017).
62. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science*. <https://doi.org/10.1126/science.1075762> (2002).
63. Bar-Joseph, Z. *et al.* Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl. Acad. Sci.* <https://doi.org/10.1073/pnas.0704723105> (2008).
64. Chae, Y. K. *et al.* Genomic landscape of DNA repair genes in cancer. *Oncotarget*. <https://doi.org/10.18632/oncotarget.8196> (2016).
65. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*. <https://doi.org/10.1038/nrm.2017.130> (2018).
66. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* <https://doi.org/10.1038/s41596-018-0103-9> (2019).
67. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkx1132> (2018).
68. Posey, J. E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N. Engl. J. Med.* <https://doi.org/10.1056/nejmoa1516767> (2016).
69. Carvalho-Silva, D. *et al.* Open Targets Platform: New developments and updates two years on. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gky1133> (2019).
70. Shin, S. H., Bode, A. M. & Dong, Z. Precision medicine: the foundation of future cancer therapeutics. *npj Precis. Oncol.* <https://doi.org/10.1038/s41698-017-0016-z> (2017).
71. Yu, Y. *et al.* PreMedKB: An integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gky1042> (2019).
72. Arnedos, M. *et al.* Precision medicine for metastatic breast cancer—limitations and solutions. *Nature Reviews Clinical Oncology*. <https://doi.org/10.1038/nrclinonc.2015.123> (2015).
73. Huang, K. lin *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*. <https://doi.org/10.1016/j.cell.2018.03.039> (2018).
74. Long, J. *et al.* Genome-wide association study in East Asians identifies novel susceptibility loci for breast cancer. *PLoS Genet.* <https://doi.org/10.1371/journal.pgen.1002532> (2012).
75. Long, J. *et al.* A common deletion in the APOBEC3 genes and breast cancer risk. *J. Natl. Cancer Inst.* <https://doi.org/10.1093/jnci/djt018> (2013).
76. Cai, Q. *et al.* Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat. Genet.* <https://doi.org/10.1038/ng.3041> (2014).
77. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature*. <https://doi.org/10.1038/nature24284> (2017).
78. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*. <https://doi.org/10.1038/clpt.2012.96> (2012).
79. Amstutz, U. *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Dihydropyrimidine Dehydrogenase Genotype and Fluoropyrimidine Dosing: 2017 Update. *Clin. Pharmacol. Ther.* **103**, 210–216 (2018).
80. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* <https://doi.org/10.1186/s13059-016-0974-4> (2016).
81. Thomas, S. J., Snowden, J. A., Zeidler, M. P. & Danson, S. J. The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. *Br. J. Cancer*. <https://doi.org/10.1038/bjc.2015.233> (2015).
82. Kazi, M., Trivedi, T., Kobawala, T. & Ghosh, N. The Potential of Wnt Signaling Pathway in Cancer: A Focus on Breast Cancer. *Cancer Transl. Med.* <https://doi.org/10.4103/2395-3977.181437> (2016).
83. Xie, B. *et al.* Benzyl Isothiocyanate potentiates p53 signaling and antitumor effects against breast cancer through activation of p53-LKB1 and p73-LKB1 axes. *Sci. Rep.* **7** (2017).
84. Paz-y-Miño, C. *et al.* Incidence of the L858R and G719S mutations of the epidermal growth factor receptor oncogene in an Ecuadorian population with lung cancer. *Cancer Genet. Cytogenet.* **196** (2010).
85. Wu, L. & Yang, X. Targeting the Hippo Pathway for Breast Cancer Therapy. *Cancers (Basel)*. <https://doi.org/10.3390/cancers10110422> (2018).
86. Uzair, I. D., Conte Grand, J., Flamini, M. I. & Sanchez, A. M. Molecular Actions of Thyroid Hormone on Breast Cancer Cell Migration and Invasion via Cortactin/N-WASP. *Front. Endocrinol. (Lausanne)*. <https://doi.org/10.3389/fendo.2019.00139> (2019).
87. Ivanov, A. A. *et al.* The OncoPPI Portal: An integrative resource to explore and prioritize protein-protein interactions for cancer target discovery. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btx743> (2018).
88. Uhlén, M. *et al.* A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics. *Mol. Cell. Proteomics*. <https://doi.org/10.1074/mcp.M500279-MCP200> (2005).
89. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
90. Murayama & Gotoh. Patient-Derived Xenograft Models of Breast Cancer and Their Application. *Cells*. <https://doi.org/10.3390/cells8060621> (2019).
91. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
92. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*. <https://doi.org/10.1038/s41586-020-1969-6> (2020).
93. De Mattos-Arruda, L. *et al.* The Genomic and Immune Landscapes of Lethal Metastatic Breast Cancer. *Cell Rep.* **27**, 2690–2708.e10 (2019).
94. Adams, S. *et al.* Current Landscape of Immunotherapy in Breast Cancer. *JAMA Oncol.* **1–10**. <https://doi.org/10.1001/jamaoncol.2018.7147> (2019).
95. Lopez-Cortes, A. *et al.* Prediction of breast cancer proteins using molecular descriptors and artificial neural networks: a focus on cancer immunotherapy proteins, metastasis driver proteins, and RNA-binding proteins. *bioRxiv Bioinforma.*, <https://doi.org/10.1101/840108> (2019).
96. López-Cortés, A. *et al.* Prediction of druggable proteins using machine learning and functional enrichment analysis: a focus on cancer-related proteins and RNA-binding proteins. *bioRxiv*. <https://doi.org/10.1101/825513> (2019).
97. Miller, S. M., Goulet, D. R. & Johnson, G. L. Targeting the Breast Cancer Kinome. *J. Cell. Physiol.* <https://doi.org/10.1002/jcp.25427> (2017).
98. Caldon, C. E., Daly, R. J., Sutherland, R. L. & Musgrove, E. A. Cell cycle control in breast cancer cells. *Journal of Cellular Biochemistry*. <https://doi.org/10.1002/jcb.20690> (2006).
99. Majidinia, M. & Yousefi, B. DNA repair and damage pathways in breast cancer development and therapy. *DNA Repair*. <https://doi.org/10.1016/j.dnarep.2017.03.009> (2017).
100. Pereira, B., Billaud, M. & Almeida, R. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends in Cancer*. <https://doi.org/10.1016/j.trecan.2017.05.003> (2017).
101. Wurth, L. *et al.* UNR/CSDE1 Drives a Post-transcriptional Program to Promote Melanoma Invasion and Metastasis. *Cancer Cell* **30**, 694–707 (2016).

102. Guerrero, S. *et al.* *In silico* analyses reveal new putative Breast Cancer RNA-binding proteins. *bioRxiv*. <https://doi.org/10.1101/2020.01.08.898965> (2020).
103. Kechavarzi, B. & Janga, S. C. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol.* <https://doi.org/10.1186/gb-2014-15-1-r14> (2014).
104. Wang, J., Liu, Q. & Shyr, Y. Dysregulated transcription across diverse cancer types reveals the importance of RNA-binding protein in carcinogenesis. *BMC Genomics.* <https://doi.org/10.1186/1471-2164-16-S7-S5> (2015).
105. Sebestyén, E. *et al.* Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* <https://doi.org/10.1101/gr.199935.115> (2016).
106. García-cárdenas, J. M. *et al.* Post-transcriptional Regulation of Colorectal Cancer: A Focus on RNA-Binding. *Proteins.* **6**, 1–18 (2019).
107. López-Cortés, A. *et al.* Pharmacogenomics, biomarker network, and allele frequencies in colorectal cancer. *Pharmacogenomics J.* <https://doi.org/10.1038/s41397-019-0102-4> (2019).
108. Harris, E. E. R. Precision Medicine for Breast Cancer: The Paths to Truly Individualized Diagnosis and Treatment. *Int. J. Breast Cancer.* <https://doi.org/10.1155/2018/4809183> (2018).
109. López-Cortés, A., Guerrero, S., Redal, M. A., Alvarado, A. T. & Quiñones, L. A. State of art of cancer pharmacogenomics in Latin American populations. *International Journal of Molecular Sciences* **18** (2017).
110. Quiñones, L. *et al.* Perception of the Usefulness of Drug/Gene Pairs and Barriers for Pharmacogenomics in Latin America. *Curr. Drug Metab.* **15**, 202–208 (2014).
111. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature.* <https://doi.org/10.1038/nature08822> (2010).
112. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* <https://doi.org/10.1186/gb-2011-12-4-r41> (2011).
113. Rocco, P. *et al.* OncoScore: A novel, Internet-based tool to assess the oncogenic potential of genes. *Sci. Rep.* <https://doi.org/10.1038/srep46290> (2017).
114. Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* 1–16. <https://doi.org/10.1016/j.ccell.2018.03.014> (2018).
115. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt1210-1248> (2010).
116. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–7 (2012).

Acknowledgements

This work was supported by Universidad UTE (Ecuador), Universidad de Las Américas (Ecuador), University of A Coruna (Spain), University of the Basque Country (Spain), and McGill University (Canada). Additionally, this work was supported by “Collaborative Project in Genomic Data Integration (CICLOGEN)” PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER).

Author contributions

A.L.C. conceived the subject, the conceptualization of the study and wrote the manuscript. E.T., S.J.B., C.R.M., H.G.D. and C.Py.M. supervised the project. A.L.C. and C.Py.M. did founding acquisition. A.L.C., S.G. and A.C.A. did data curation and supplementary data. E.T., S.G., A.C.A., S.J.B., C.R.M., H.G.D., A.P., Y.P.C. and C.Py.M. gave conceptual advice and valuable scientific input. Finally, all authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-62279-2>.

Correspondence and requests for materials should be addressed to A.L.-C. or E.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020



OPEN

Prediction of breast cancer proteins involved in immunotherapy, metastasis, and RNA-binding using molecular descriptors and artificial neural networks

Andrés López-Cortés^{1,2,3,12}✉, Alejandro Cabrera-Andrade^{2,4,5,12},
José M. Vázquez-Naya^{2,6,7}, Alejandro Pazos^{2,6,7}, Humberto González-Díaz^{8,9},
César Paz-y-Miño¹, Santiago Guerrero¹, Yunierkis Pérez-Castillo^{4,10}, Eduardo Tejera^{4,11}
& Cristian R. Munteanu^{2,6,7}

Breast cancer (BC) is a heterogeneous disease where genomic alterations, protein expression deregulation, signaling pathway alterations, hormone disruption, ethnicity and environmental determinants are involved. Due to the complexity of BC, the prediction of proteins involved in this disease is a trending topic in drug design. This work is proposing accurate prediction classifier for BC proteins using six sets of protein sequence descriptors and 13 machine-learning methods. After using a univariate feature selection for the mix of five descriptor families, the best classifier was obtained using multilayer perceptron method (artificial neural network) and 300 features. The performance of the model is demonstrated by the area under the receiver operating characteristics (AUROC) of 0.980 ± 0.0037 , and accuracy of 0.936 ± 0.0056 (3-fold cross-validation). Regarding the prediction of 4,504 cancer-associated proteins using this model, the best ranked cancer immunotherapy proteins related to BC were RPS27, SUPT4H1, CLPSL2, POLR2K, RPL38, AKT3, CDK3, RPS20, RASL11A and UBD1; the best ranked metastasis driver proteins related to BC were S100A9, DDA1, TXN, PRNP, RPS27, S100A14, S100A7, MAPK1, AGR3 and NDUFA13; and the best ranked RNA-binding proteins related to BC were S100A9, TXN, RPS27L, RPS27, RPS27A, RPL38, MRPL54, PPA, RPS20 and CSRP1. This powerful model predicts several BC-related proteins that should be deeply studied to find new biomarkers and better therapeutic targets. Scripts can be downloaded at <https://github.com/muntisa/neural-networks-for-breast-cancer-proteins>.

The intricate interplay between several biological aspects such as environmental determinants, gene expression deregulation, genetic alterations, signaling pathway alterations and ethnicity causes the development of breast

¹Centro de Investigación Genética y Genómica, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Mariscal Sucre Avenue, Quito, 170129, Ecuador. ²RNASA-IMEDIR, Computer Science Faculty, University of Coruna, Coruna, 15071, Spain. ³Red Latinoamericana de Implementación y Validación de Guías Clínicas Farmacogenómicas (RELIVAF-CYTED), Quito, Ecuador. ⁴Grupo de Bio-Quimioinformática, Universidad de Las Américas, Avenue de los Granados, Quito, 170125, Ecuador. ⁵Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de Las Américas, Avenue de los Granados, Quito, 170125, Ecuador. ⁶Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Campus de Elviña s/n 15071, A Coruña, Spain. ⁷Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), 15006, A Coruña, Spain. ⁸Department of Organic Chemistry II, University of the Basque Country UPV/EHU, Leioa 48940, Biscay, Spain. ⁹IKERBASQUE, Basque Foundation for Science, Bilbao, 48011, Biscay, Spain. ¹⁰Escuela de Ciencias Físicas y Matemáticas, Universidad de Las Américas, Avenue de los Granados, Quito, 170125, Ecuador. ¹¹Facultad de Ingeniería y Ciencias Agropecuarias, Universidad de Las Américas, Avenue de los Granados, Quito, 170125, Ecuador. ¹²These authors contributed equally: Andrés López-Cortés and Alejandro Cabrera-Andrade. ✉e-mail: aalc84@gmail.com

cancer (BC), a heterogeneous disease^{1,2}. Over the last years, multi-omics studies, pharmacogenomics treatments and precision medicine strategies have evolved favorably; however, there are still biases such as the significant inclusion of minority populations in cancer research³⁻⁷. Nowadays, BC is the most commonly diagnosed cancer (2,088,849; 24% cases), and the leading cause of cancer-related deaths among women (626,679; 15% cases) worldwide⁸.

In our previous study, López-Cortés *et al.* developed the OncoOmics strategy to reveal essential genes in BC⁹. This strategy was a compendium of approaches that analyzed genomic alterations, protein expression, protein-protein interactome (PPI) network, dependency maps in cell lines and patient-derived xenografts of BC genes / proteins using relevant databases such as the Pan-Cancer Atlas project^{3,10-12}, The Cancer Genome Atlas (TCGA)¹³, The Human Protein Atlas (HPA)¹⁴⁻¹⁶, the DepMap project¹⁷⁻¹⁹, and the OncoPPI network²⁰.

Gene sets were taken from the Consensus Strategy²¹, the Pan-Cancer Atlas^{3,11,12,22}, the Pharmacogenomics Knowledgebase (PharmGKB)^{23,24}, and the Cancer Genome Interpreter²⁵. The Consensus Strategy, developed by López-Cortés *et al.*, Tejera *et al.*, and Cabrera-Andrade *et al.*, was proved to be highly efficient in the recognition of genes associated with BC pathogenesis^{21,26,27}. The Pan-Cancer Atlas reveals how genomic alterations, such as protein expression, copy number alterations (CNAs), mRNA expression, and putative mutations collaborate in BC progression^{11,22,28-32}. PharmGKB is a comprehensive resource that collects the precise guidelines for the application of pharmacogenomics in clinical practice^{23,24}. Lastly, the Cancer Genome Interpreter flags genomic biomarkers of drug response with different levels of clinical relevance²⁵.

The OncoOmics BC essential genes were rationally filtered to 140. *RAC1*, *AKT1*, *CCND1*, *PIK3CA*, *ERBB2*, *CDH1*, *MAPK14*, *TP53*, *MAPK1*, *SRC*, *RAC3*, *BCL2*, *CTNNB1*, *EGFR*, *CDK2*, *GRB2*, *MED1*, and *GATA3* were significant in at least three OncoOmics approaches⁹. On the other hand, g:Profiler lets us know the enrichment map of the 140 essential genes in BC³³. The most significant gene ontologies (GO) related to biological process and molecular function were the positive regulation of macromolecule metabolic process and the phosphatidylinositol 3-kinase activity, respectively. The most significant term, according to the Human Phenotype Ontology, was breast carcinoma³⁴. Subsequently, the most relevant network interactions of the GO: biological process and the Reactome pathways were related to the immune system³⁵, tyrosine kinase³⁶, cell cycle³⁷, DNA repair³⁸, and RNA-binding proteins³⁹. The Open Targets Platform has a largest number of drugs involved in clinical trials to treat BC with a direct focus on the OncoOmics BC essential genes were small molecules that correspond most likely to tyrosine kinases⁴⁰. Hence, the essential proteins with signaling function are the interesting drug targets to modify any biological activity.

Starting a screening applying theoretical methods could save economic resources and time. Therefore, machine-learning (ML) techniques could obtain classification models that links signaling activity to protein structure. ML encodes molecular features into invariant descriptors based on physical and chemical properties of the amino acids, 3D protein conformation, graph topology, and protein sequences. The classification model is a quantitative structure-activity relationship (QSAR) between the biological function and the protein structure⁴¹. Different classification models have been published for prediction of protein activities: anti-oxidant⁴², lectins⁴³, signaling⁴⁴, anti-angiogenic⁴⁵, anti-cancer⁴⁶, and enzyme class⁴⁷. Vilar *et al.* developed a QSAR model for alignment-free prediction of BC biomarkers using a linear discriminant analysis method, electrostatic potentials of protein pseudofolding HP-lattice networks as features, and 122 proteins related to BC and a control group of 200 proteins with classifications above 80%⁴⁸. Our group proposed an improved multi-target classification model for human breast and colon cancer-related proteins by using a similar molecular graph theory for descriptors: star graph topological indices⁴⁹. The accuracy of the models was 90.0% for a linear forward stepwise model. Both models presented linear relationships between graph-based protein sequence descriptors and BC, and unbalanced datasets. Thus, the aim of this study was to obtain an effective machine-learning classification model to predict BC-related proteins screening cancer immunotherapy proteins (CIPs), metastasis driver proteins (MDPs) and RNA-binding proteins (RBPs), using non-graph protein sequence descriptors and additional non-linear machine-learning techniques.

Methods

Figure 1 presents the general flow chart of the methodology to obtain a classifier for BC proteins. In the first step, we constructed a database with BC essential proteins and non-cancer proteins. In the second step, five families of RcpI (R package)⁵⁰ molecular descriptors have been used: 20 amino acid composition (AC), 400 di-amino acid composition (DC), 8000 tri-amino acid composition (TC), 80 amphiphilic pseudo-amino acid composition (APAAC), and 240 normalized Moreau-Broto autocorrelation (MB). The six sets of descriptors were constructed by mixing all the five-descriptor families, resulting 8,708 total descriptors (Mix).

Jupyter notebooks with python/sklearn⁵¹ were used to test 13 types of machine-learning classifiers for each set of descriptors, without feature selection, with univariate feature selection, or using principal component analysis (PCA)⁵². The classifiers were Gaussian Naive Bayes (NB)⁵³, k-nearest neighbors algorithm (KNN)⁵⁴, linear discriminant analysis (LDA)⁵⁵, support vector machine (SVM) linear and non-linear based on radial basis functions (RBF), support vector classification (SVC) kernel = linear, and SVC kernel = RBF⁵⁶, logistic regression (LR)⁵⁷, multilayer perceptron (MLP) / neural network with 20 neurons in one hidden layer⁵⁸, decision tree (DT)⁵⁹, random forest (RF)⁶⁰, XGBoost (XGB) is an optimized and distributed gradient boosting library⁶¹, Gradient Boosting for classification (GB)⁶², AdaBoost classifier (AdaB)⁶³, and Bagging classifier (Bagging)⁶⁴. The feature selection method was univariate filter such as SelectKBest (chi2, k), and the dimension reduction technique was PCA⁵².

Gaussian Naive Bayes is based on Bayes' theorem and considers all the features are independent⁵³. k-nearest neighbors algorithm assigns an unclassified sample using the nearest of k samples in the training set⁵⁴. Linear discriminant analysis is a basic linear classifier⁵⁵. SVM linear is using a higher dimensionality space to map the input features⁵⁶. For non-linear problems, SVM uses Gaussian radial basis as non-linear kernels.

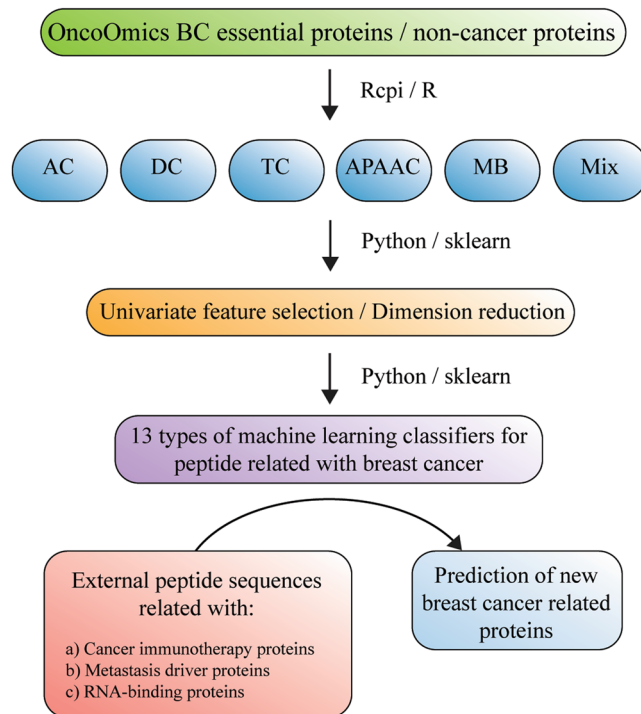


Figure 1. Flow chart of methodology for breast cancer (BC) protein prediction. AC, amino acid composition; DC, di-amino acid composition; TC, tri-amino acid composition; APAAC, amphiphilic pseudo-amino acid composition; MB, Moreau–Broto autocorrelation; Mix, total descriptors.

Logistics regression is another linear classifier that is able to calculate probability of a binary response using weights⁵⁷. Multilayer perceptron represents a basic neural network with one hidden layer and with an ability to combine linear and nonlinear functions inside artificial neurons⁵⁸. Decision tree represents a tree-type structure of decision rules obtained from the inputs⁵⁹. Random forest is an ensemble method that combines parallel decision trees⁶⁰. XGBoost uses sequential weak trees to improve the classification performance⁶¹. Gradient Boosting for classification is a basis boost method using sequential weak classifiers⁶². AdaBoost classifier is mixing different classifiers: it starts the fitting with a classifier based on the original dataset and adds additional copies of the original classifier with adjusted weights for the incorrectly classified instances⁶³. Bagging classifier is a modified version of AdaB: the additional classifiers are based on subsets of the original dataset⁶⁴.

The machine-learning prediction model was constructed from two protein sets. On the one hand, the positive set named OncoOmics BC essential proteins was made up of 140 strongly associated proteins to BC pathogenesis, according to López-Cortés *et al.*⁹. On the other hand, the negative protein set was constructed as follows: non-cancer proteins from Piazza *et al.*⁶⁵, without BC-related proteins, were reanalyzed using Piazza's OncoScore algorithm (<http://www.galseq.com/oncoscore.html>), giving a final list of 233 non-cancer proteins. Supplementary Tables 1 and 2 detail the sets and FASTA sequences of the OncoOmics BC essential proteins and the non-cancer proteins, respectively.

Three lists of cancer-related proteins were scanned with the final machine-learning prediction model: 1,232 CIPs were taken from Patel *et al.*,³⁵ 1,903 MDPs were taken from the Human Cancer Metastasis Database (HCMDB) (<http://hcmdb.i-sanger.com/index>)⁶⁶, and 1,369 RBPs were taken from Hentze *et al.*,³⁹ (Supplementary Tables 3 to 5).

After the calculation of amino acid composition descriptors, the datasets contained 373 proteins. The BC class was labeled with 1 and non-cancer class with 0. Several preprocessing was done before any calculation: elimination of doubled examples, elimination of data with NA values, and elimination of features with zero variance. All feature values were normalized to values between 0 and 1 using MinMax() scaler. A SMOTE filter was used to balance the dataset⁶⁷. The performance of the models used Area Under the Receiver Operating Characteristics (AUROC) metrics⁶⁸, and 3-fold cross-validation (CV) method.

The best model to be used for predictions was chosen using criteria such as mean AUROC, standard deviation (SD) of AUROC, and the number of features. All the results obtained can be reproduced by using the scripts at <https://github.com/muntisa/neural-networks-for-breast-cancer-proteins>. The scaler, selected features and the best model were saved as files too. These are used to make predictions with another notebook for any new data (see 2-Predictions-BreastCancerPeptides.ipynb). We used these automatic scripts to predict the breast cancer activity for a 4,504 external proteins by using their molecular descriptors: 1,232 CIPs, 1,903 MDPs, and 1,369 RBPs.

After the screening of the 4,504 external proteins through the machine-learning model, complementary analyses were done to compare the amount of genomic alterations between BC related proteins (prediction 1)

and BC non-related proteins (prediction 0). Firstly, we selected the study ‘Breast Invasive Carcinoma (TCGA, PanCancer Atlas)’ from the cBioPortal (<https://www.cbioportal.org/>)^{69,70}, then, we downloaded and analyzed a matrix of CNAs (amplifications and deep deletions), putative mutations (inframe, truncating and missense), mRNA alterations (mRNA high and mRNA down), and protein alterations (high and low expression) related to the 4,504 proteins queried in a cohort of 1,066 individuals according to the Pan-Cancer Atlas^{3,11,12,22}. Lastly, a Mann-Whitney U test was performed to obtain significant differences ($p < 0.001$) on the amount of genomic alterations between CIPs related and non-related to BC, MDPs related and non-related to BC, and RBPs related and non-related to BC.

Results and Discussion

The current work proposes innovative classification models to predict new breast cancer proteins by using 6 sets of protein sequence descriptors calculated with Rcp1: AC, DC, TC, APAAC, MB and Mix. Python was used to build 13 types of machine-learning classifiers (NB, KNN, LDA, SVM linear, SVM, LR, MLP, DT, RF, XGB, GB, AdaB and Bagging), univariate filter as feature selection method, and PCA transformation of features. All the models used AUROC (mean values using 3-fold CV) to quantify the classification performance. Details about feature selection methods and parameters of machine-learning classifiers are included in the Supplementary_ML_Details.pdf.

For the first models, we used the pool of features for the six sets of descriptors without any feature selection or dimension reduction with 12 machine-learning methods (Fig. 2). We can observe that with a big number of descriptors in TC and Mix (over 8000), it is possible to obtain mean AUROC values greater than 0.9 with SVM linear, LR, and MLP. Even with 20 AC descriptors and XGB it is possible to obtain a mean AUROC of 0.857. But we tried to improve this performance and we applied univariate feature selection or PCA dimension reduction to diminish the number of inputs to a maximum of 300 features (due to the small number of instances).

Therefore, we selected models based on 20, 100, 200, and 300 features (see 1-ML-BreastCancerPeptides.ipynb). Figure 3 presents mean AUROC values for classifiers based on only 20 features: AC, DS-Best20, DC-PCA20, TC-Best20, TC-PCA20, APAAC-Best20, APAAC-PCA20, MB-Best20, MB-PCA20, Mix-Best20 and Mix-PCA20 (Best = univariate filter, PCA = feature transformation). DS-Best20 with only 20 di-amino acid composition descriptors and Mix-Best20 with a mixture of descriptors are able to offer mean AUROC values over 0.84 with non-linear SVM, XGB and GB. Additional results could be found in Supplementary Table 6.

If the number of features increased to 100 (5 times from 20), better AUROC values are obtained in Fig. 4: DC-Best100, DC-PCA100, TC-Best100, TC-PCA100, MB-Best100, MB-PCA100, Mix-Best100, and Mix-PCA100. Two sets of descriptors with four machine-learning methods are able to provide mean AUROC values greater than 0.9: TC-Best100 and Mix-Best100 with SVM linear, non-linear SVM, LR and MLP. Thus, LR and TC-Best100 (100 descriptors of tri-amino acid composition) generate a classifier with mean AUROC of 0.917. The increasing of AUROC values is important from 20 to 100 best descriptors. In the next step, the number of selected descriptors was increased to 200. The PCA transformed sets using the same number of components, as the selected features are not able to provide similar classification performance.

Figure 5 presents the AUROC values for classifiers based on 200 selected features (a double number of inputs from 100): DC-Best200, DC-PCA200, TC-Best200, TC-PCA200, MB-Best200, MB-PCA200, Mix-Best200, and Mix-PCA200. We can observe that the same TC and Mix-based sets are providing mean AUROC values between 0.90 and 0.95 with five machine-learning methods: NB, SVM linear, LR, MLP, and RF. The maximum mean AUROC value was 0.950 using TC-Best200 and the simple linear LR method.

In Fig. 6 the AUROC values for classifiers based on 300 selected features are presented: DC-Best300, DC-PCA300, TC-Best300, TC-PCA300, Mix-Best300, and Mix-PCA300. With 300 features, it is possible to provide more accurate classifier for BC proteins. The same TC and Mix subsets can generate classifiers with mean AUROC from 0.963 to 0.980 using SVM linear, SVM, LR and MLP.

The best AUROC of 0.980 ± 0.0037 was obtained with MLP and Mix-Best300. The same AUROC value was generated by TC-Best300 and LR but with a double SD of 0.0077. In the best model with the mixed descriptors, between the 300 descriptors, seven DC (LR, QI, NK, EM, QM, MM and EY) and two APAAC descriptors (Pc1.N and Pc1.M) were selected for BC function. The rest is TC descriptors without any MP descriptor selected (see Supplementary Table 7). The accuracy of the best model was 0.936 ± 0.0056 . No methodology is perfect, and; therefore, our method/model has few weak spots: a) our dataset could be bigger: more examples/instances mean more accurate models. We were limited by the available database data; b) the best model has a relatively high number of descriptors: a model should use the minimum number of features because of simplicity, model explanation power, and to not overfit the dataset; c) our best model is an MLP with 300 descriptors and AUROC of 0.98, but in Figs. 3–6 we showed other different models obtained with other machine-learning methods, based on a smaller number of features. Thus, we can observe that it is possible to obtain a prediction model with an AUROC > 0.84 with only 20 descriptors. If the interest is the number of descriptors, the user could reproduce the models with the available notebooks and save any model; d) the best model is a black box such any neural network. If the explanation of the machine learning is the most important aspect, there are models with AUROC > 0.84 that could be explained better such as tree-based methods or linear models; e) our results could be improved by an extensive grid search of the hyperparameters of each machine-learning method. We did not consider this step because of the very high values of AUROC, which are fine for the purpose of this study.

In order to check if the best model is overfitted, we tried different CV folds (data splits) with the same MLP method (see CVs.ipynb for details). Thus, in the case of 5-fold CV, the mean AUROC was 0.9874 ± 0.0129 and the mean ACC was 0.9464 ± 0.0135 . By increasing the number of folds to 10, the statistics showed a mean AUROC of 0.9831 ± 0.0158 , and a mean ACC of 0.9401 ± 0.0226 . All the models are saved into folder *best_classifier*. Therefore, we can conclude that the performance of the best model slightly increases with increased SD values. If

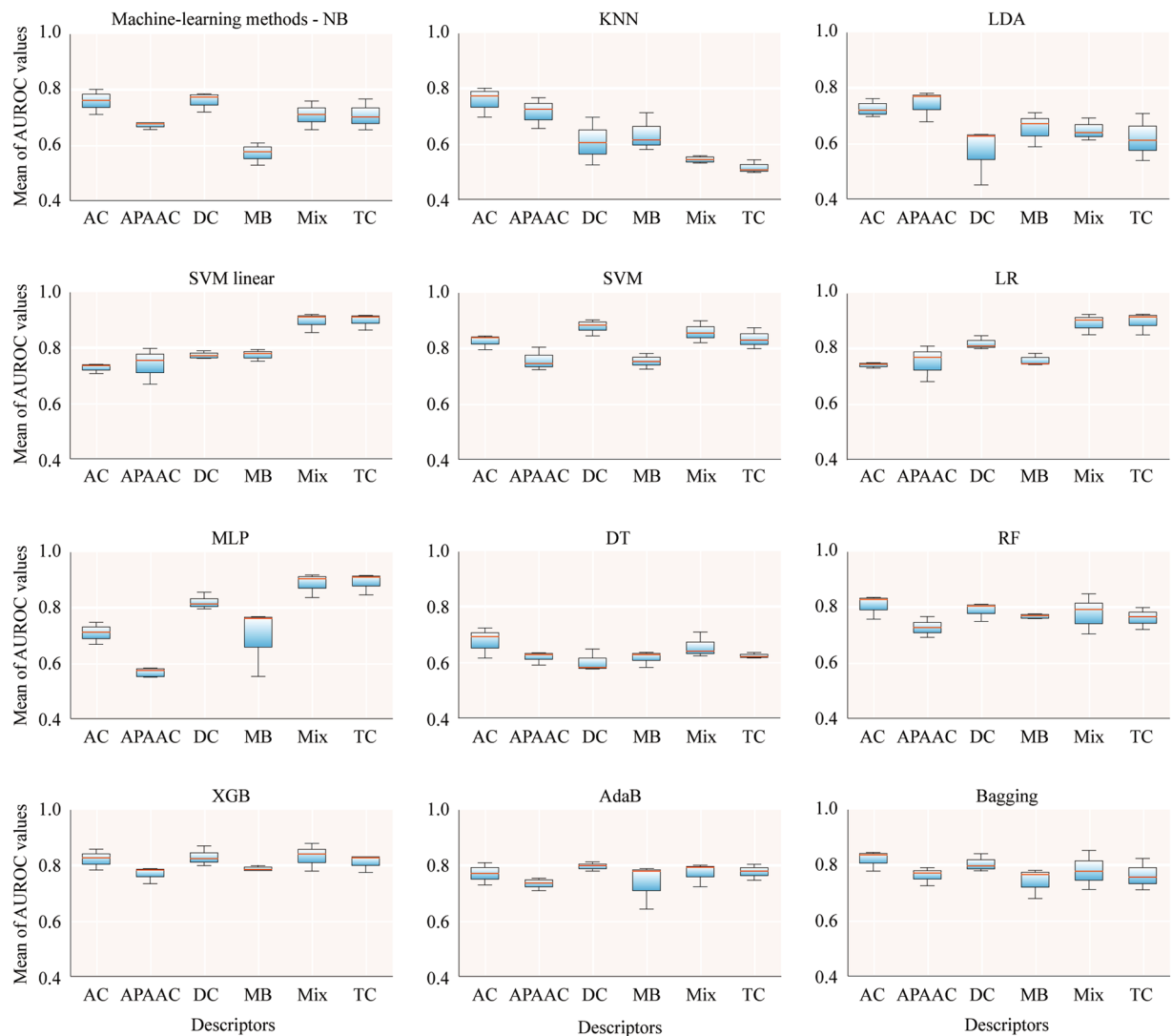


Figure 2. Mean AUROC of classifiers for breast cancer proteins using all features. NB, Gaussian Naive Bayes; KNN, k-nearest neighbors algorithm; LDA, linear discriminant analysis; SVM linear, support vector machine linear; LR, logistic regression; MLP, multilayer perceptron; DT, decision tree; RF, random forest; XGB, XGBoost; AdaB, AdaBoost classifier; Bagging, Bagging classifier; AC, amino acid composition; APAAC, amphiphilic pseudo-amino acid composition; DC, di-amino acid composition; MB, Moreau-Broto autocorrelation; Mix, total descriptors; TC, tri-amino acid composition.

these statistics are not fine for a specific application, it is possible to choose a different model based on 20 descriptors but with statistics greater than 0.80.

The 4,504 external proteins (1,903 without repetition) were transformed into the molecular descriptors of the best model and were used to predict the breast cancer activity (see 2-Predictions-BreastCancerPeptides.ipynb): 1,232 CIPs, 1,903 MDPs and 1,369 RBPs. Thus, all these proteins were transformed into 300 selected descriptors of a Mix-300 set and were used with the saved MLP classifier. As a result, 608 cancer immunotherapy proteins, 971 metastasis driver proteins and 757 RNA binding proteins were predicted to be related to breast cancer (Supplementary Tables 3 to 5).

Cancer immunotherapy proteins. These proteins have a promising projection in clinical oncology due to successful long-term durable responses in advanced stages and metastasis. Similarly, cancer immunotherapy sparked tremendous interest in clinical, basic and translational science⁷¹. The 10 cancer immunotherapy proteins best related to BC, according to our machine-learning predictions, were RPS27, SUPT4H1, CLPSL2, POLR2K, RPL38, AKT3, CDK3, RPS20, RASL11A, and UNTD1 (Supplementary Table 3). For instance, Atsuta *et al.* determined that RPS27 is a tumor associated antigen in BC patients⁷².

The development of cutting-edge technologies focused on the analysis of genomic alterations in cancer patients has allowed finding novel driver genes and therapeutic targets⁷³. Hence, we performed an analysis to compare the amount of genomic alterations of the cancer immunotherapy proteins best related to breast cancer, according to

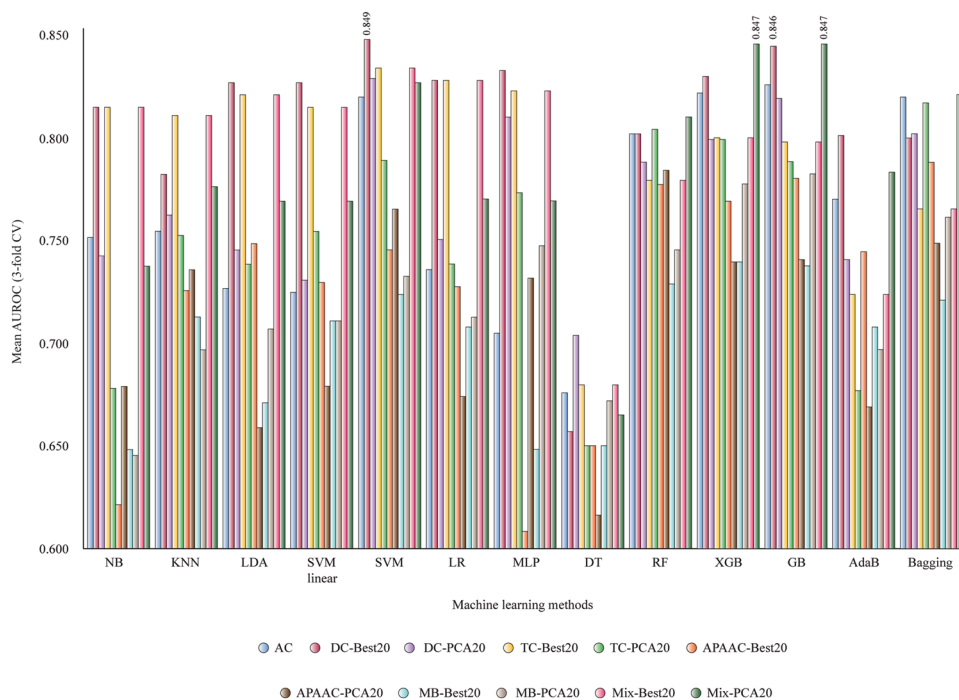


Figure 3. Mean AUROC values for classifiers obtained with 20 selected features (3-fold CV). NB, Gaussian Naive Bayes; KNN, k-nearest neighbors algorithm; LDA, linear discriminant analysis; SVM linear, support vector machine linear; LR, logistic regression; MLP, multilayer perceptron; DT, decision tree; RF, random forest; XGB, XGBoost; AdaB, AdaBoost classifier; Bagging, Bagging classifier; AC, amino acid composition; APAAC, amphiphilic pseudo-amino acid composition; DC, di-amino acid composition; MB, Moreau-Broto autocorrelation; Mix, total descriptors; TC, tri-amino acid composition.

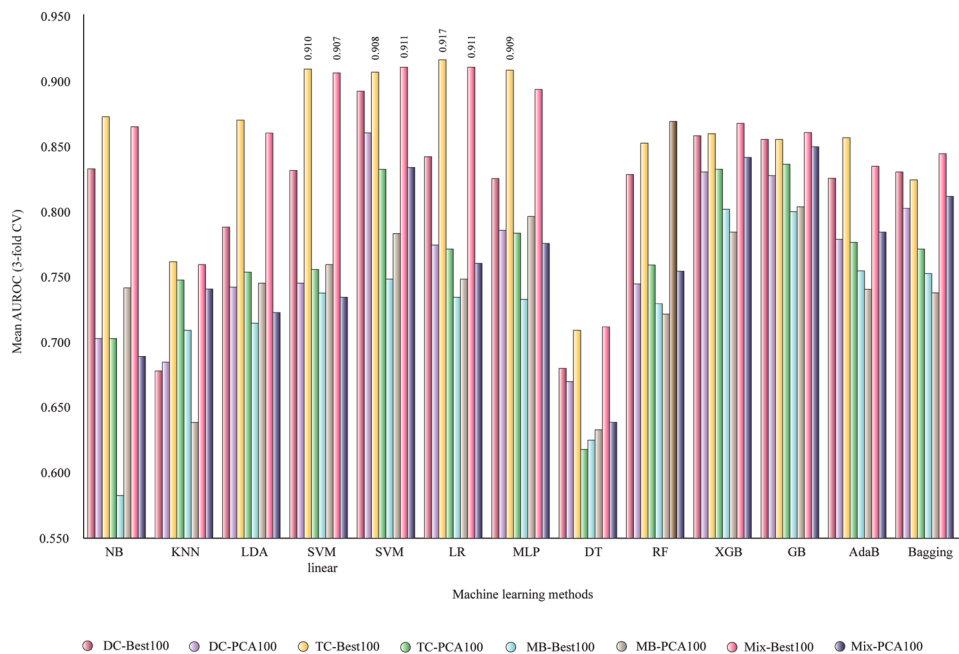


Figure 4. Mean AUROC for classifiers based on 100 selected features (3-fold CV). NB, Gaussian Naive Bayes; KNN, k-nearest neighbors algorithm; LDA, linear discriminant analysis; SVM linear, support vector machine linear; LR, logistic regression; MLP, multilayer perceptron; DT, decision tree; RF, random forest; XGB, XGBoost; AdaB, AdaBoost classifier; Bagging, Bagging classifier; DC, di-amino acid composition; MB, Moreau-Broto autocorrelation; Mix, total descriptors; TC, tri-amino acid composition.

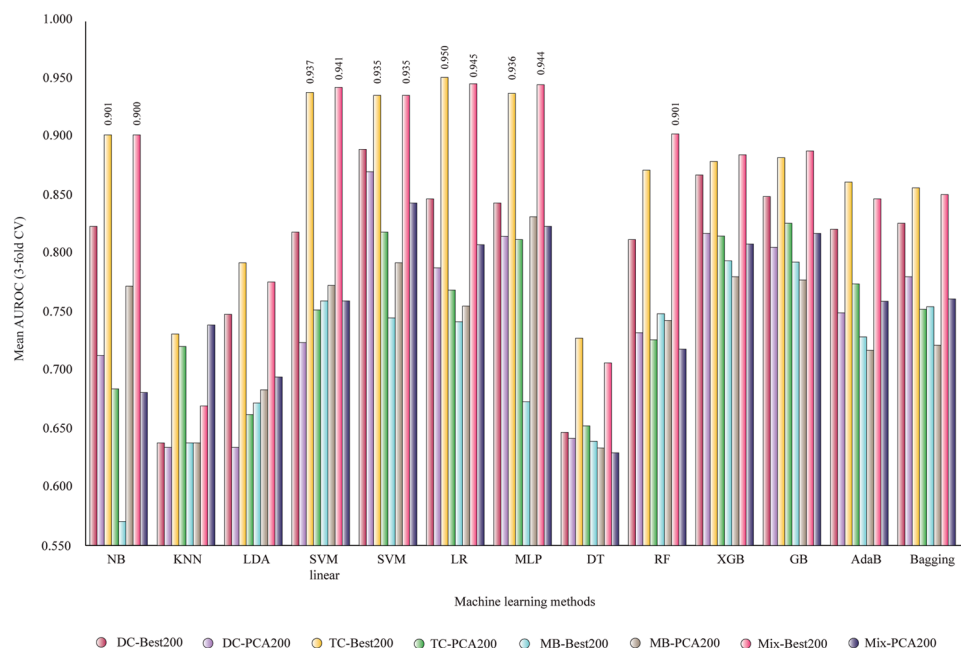


Figure 5. Mean AUROC of classifiers based on 200 selected features (3-fold CV). NB, Gaussian Naive Bayes; KNN, k-nearest neighbors algorithm; LDA, linear discriminant analysis; SVM linear, support vector machine linear; LR, logistic regression; MLP, multilayer perceptron; DT, decision tree; RF, random forest; XGB, XGBoost; AdaB, AdaBoost classifier; Bagging, Bagging classifier; DC, di-amino acid composition; MB, Moreau-Broto autocorrelation; Mix, total descriptors; TC, tri-amino acid composition.

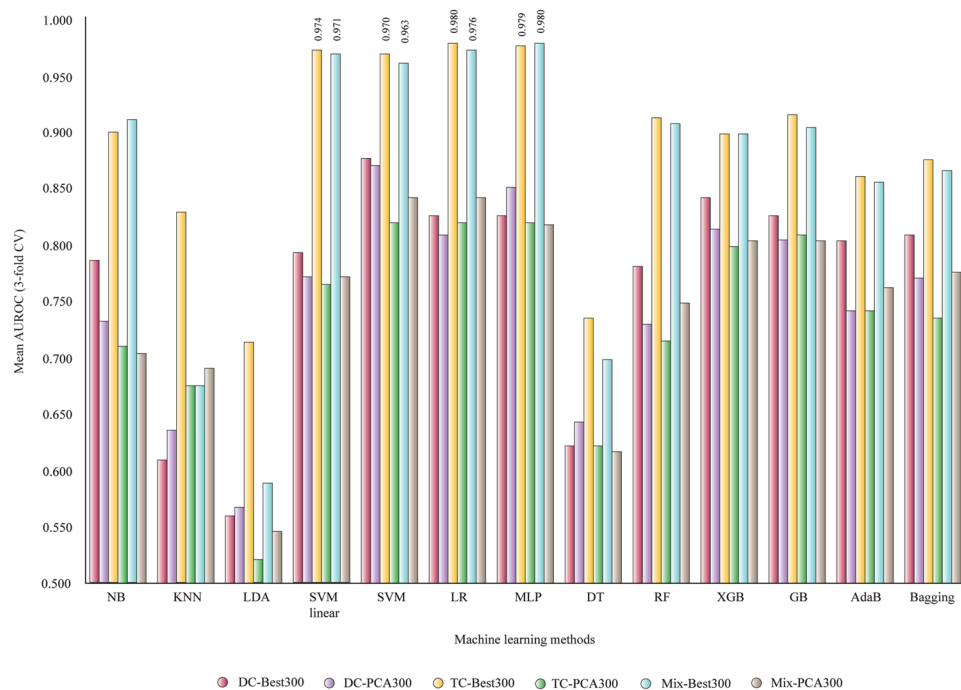


Figure 6. Mean AUROC of classifiers based on 300 selected features (3-fold CV). NB, Gaussian Naive Bayes; KNN, k-nearest neighbors algorithm; LDA, linear discriminant analysis; SVM linear, support vector machine linear; LR, logistic regression; MLP, multilayer perceptron; DT, decision tree; RF, random forest; XGB, XGBoost; AdaB, AdaBoost classifier; Bagging, Bagging classifier; DC, di-amino acid composition; MB, Moreau-Broto autocorrelation; Mix, total descriptors; TC, tri-amino acid composition.

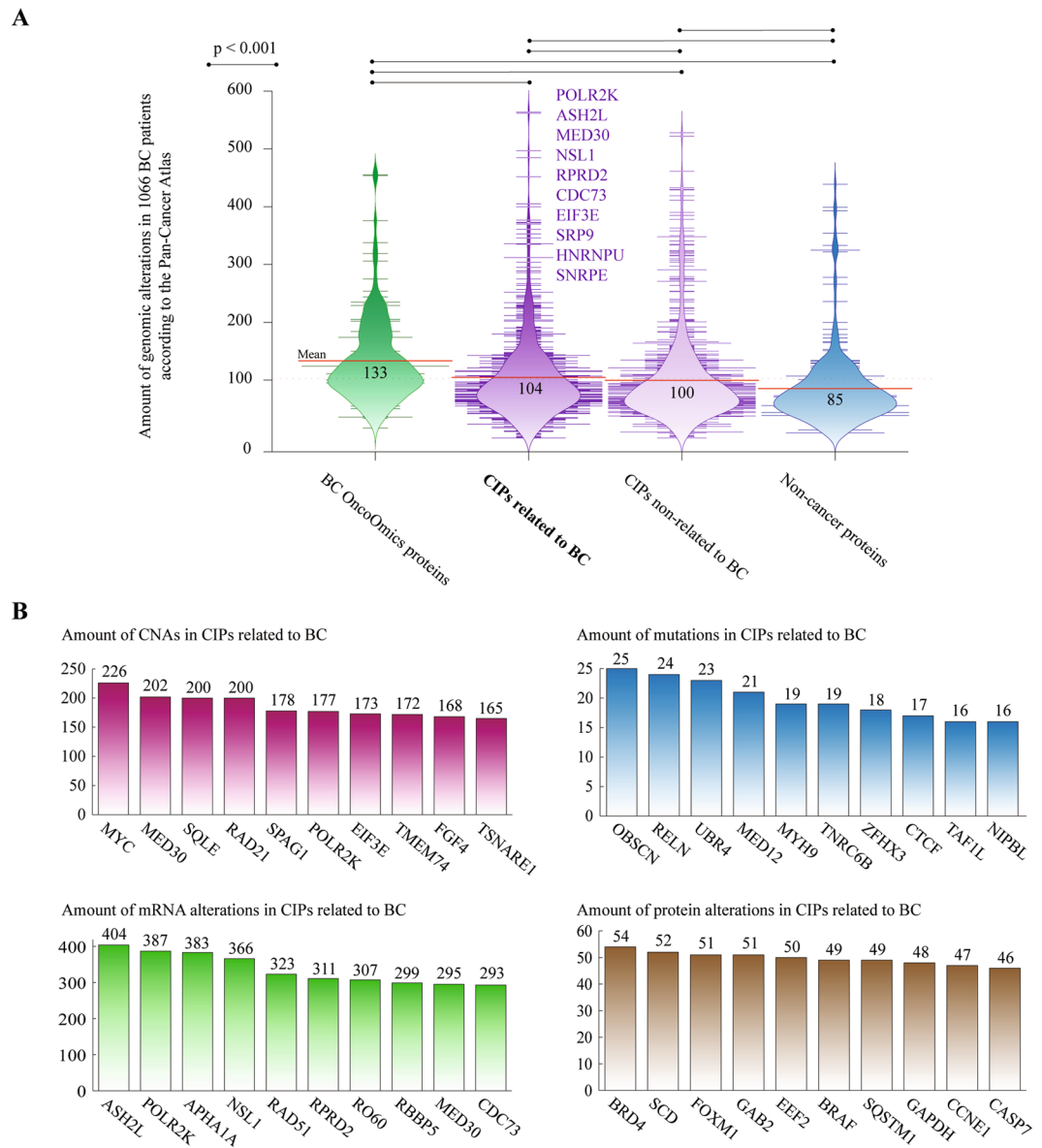


Figure 7. Cancer immunotherapy proteins (CIPs). **(A)** Bean plots comparing the amount (mean) of genomic alterations in 1066 patients between OncoOmics BC essential proteins, CIPs related to breast cancer, CIPs non-related to breast cancer, and non-cancer proteins according to the Pan-Cancer Atlas. **(B)** Ranking of the CIPs with the highest number of copy number alterations (CNAs), mutations, mRNA alterations, and protein alterations.

the Pan-Cancer Atlas^{3,11,12,22}, Figure 7A compares the amount of genomic alterations in a cohort of 1,066 patients between the OncoOmics BC essential proteins (mean of 133), CIPs related to BC (104), CIPs non-related to BC (100), and non-cancer proteins (85). As we can see, there was a significant difference ($p < 0.001$) of genomic alterations between CIPs related and non-related to BC after the Mann-Whitney U test. The top 10 CIPs related to BC and with the highest amount of genomic alterations were POLR2K, ASH2L, MED30, NSL1, RPRD2, CDC73, EIF3E, SRP9, HNRNPU and SNRPE (Supplementary Table 8). Additionally, Fig. 7B shows the most altered cancer immunotherapy proteins per genomic alteration type. MYC, OBSCN, ASH2L and BRD4 carried the highest number of CNAs, mutations, mRNA alterations and protein alterations, respectively.

Metastasis driver proteins. Metastasis, often preceded or accompanied by therapeutic resistance, is the most lethal and insidious aspect of cancer. Due to treatment pressure, tumor evolution or mitochondria dysfunction, genomic alterations of metastatic tumors can differ substantially from primary tumors^{74–76}. To date, the molecular and microenvironmental determinants of metastasis are largely unknown, as is the timing of systemic spread, hindering effective treatment and prevention efforts^{66,77}. Integrated analysis of ‘omics’ data improves our understanding of BC metastasis. Moreover, these data would help us identify gene expression signature associated with metastasis in order to choose appropriate treatment strategies^{78,79}. The 10 MDPs best related to BC,

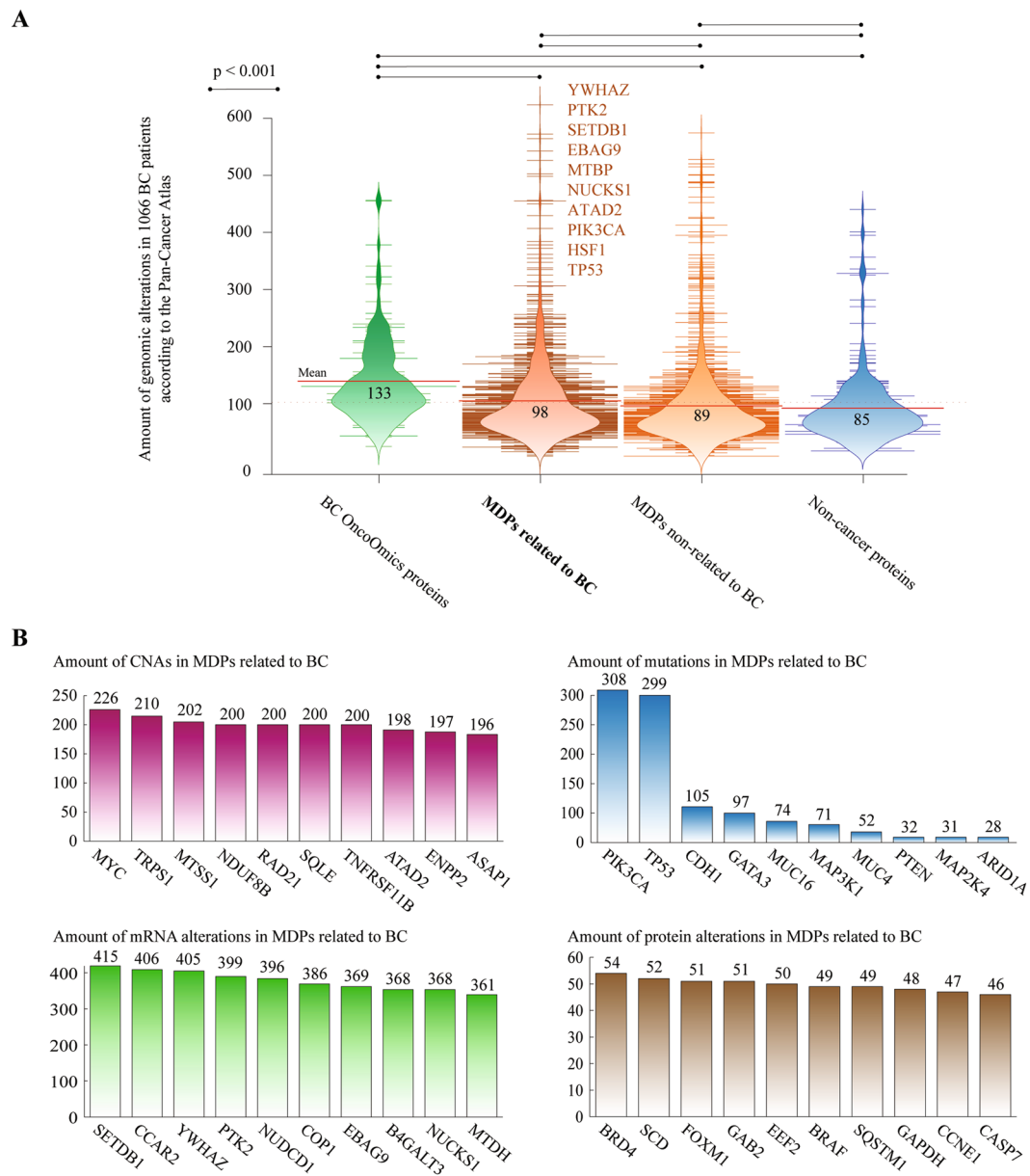


Figure 8. Metastasis driver proteins (MDPs). **(A)** Bean plots comparing the amount (mean) of genomic alterations in 1066 patients between OncoOmics BC essential proteins, MDPs related to breast cancer, MDPs non-related to breast cancer, and non-cancer proteins according to the Pan-Cancer Atlas. **(B)** Ranking of the MDPs with the highest number of copy number alterations (CNAs), mutations, mRNA alterations, and protein alterations.

according to our machine-learning predictions, were S100A9, DDA1, TXN, PRNP, RPS27, S100A14, S100A7, MAPK1, AGR3 and NDUFA13 (Supplementary Table 4). For instance, Bergenfelz *et al.* suggested that S100A9 expressed in negative estrogen receptor and negative progesterone receptor breast cancers induces inflammatory cytokines and it is associated with an impaired overall survival⁸⁰.

Figure 8A shows bean plots comparing the amount of genomic alterations between the OncoOmics BC essential proteins (mean of 133), MDPs related to BC (98), MDPs non-related to BC (89) and non-cancer proteins (85). There was a significant difference ($p < 0.001$) of genomic alterations between MDPs related and non-related to BC after the Mann-Whitney U test. The top 10 MDPs related to BC and with the highest amount of genomic alterations were YWHAZ, PTK2, SETDB1, EBAG9, MTBP, NUCKS1, ATAD2, PIK3CA, HSF1 and TP53 (Supplementary Table 8). In addition, Fig. 8B shows the most altered metastasis driver proteins per genomic alteration type. MYC, PIK3CA, SETDB1 and BRD4 carried the highest number of CNAs, mutations, mRNA alterations and protein alterations, respectively.

RNA-binding proteins. RNA biology is an under-investigated field of cancer even though pleiotropic changes in the transcriptome are key feature of cancer cell⁸¹. RBPs are able to control every aspect of RNA

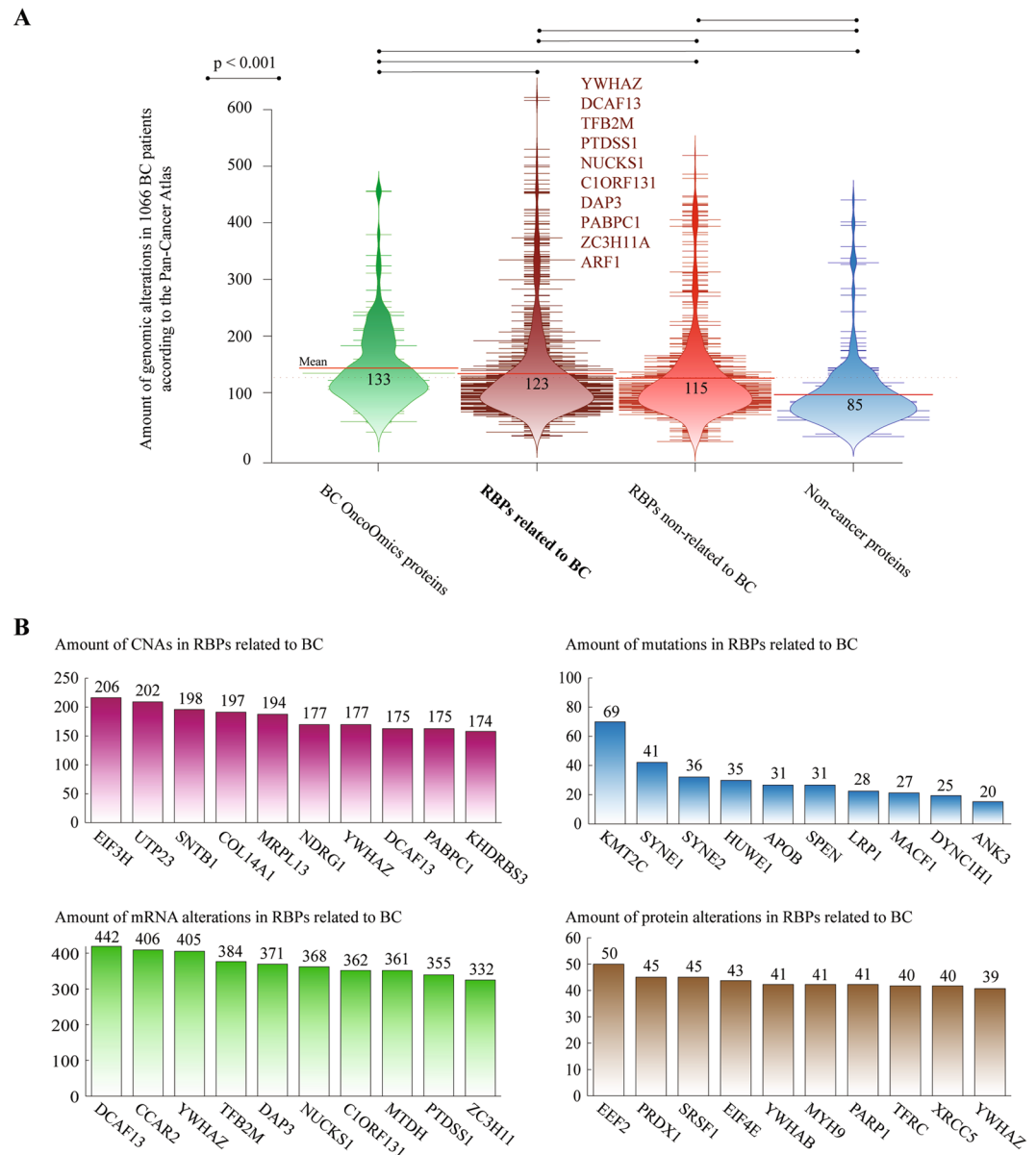


Figure 9. RNA-binding proteins (RBPs). **(A)** Bean plots comparing the amount (mean) of genomic alterations in 1066 patients between OncoOmics BC essential proteins, RBPs related to breast cancer, RBPs non-related to breast cancer, and non-cancer proteins according to the Pan-Cancer Atlas. **(B)** Ranking of the RBPs with the highest number of copy number alterations (CNAs), mutations, mRNA alterations, and protein alterations.

metabolism such as translation, splicing, stability, degradation of mRNA, nucleocytoplasmic transport, capping, and polyadenylation^{81–85}. RBPs are emerging as critical modulators of BC and the prediction of relation with this complex disease through machine-learning methods provides a better understanding of new genomic targets and biomarkers. The 10 RBPs best related to BC, according to our machine-learning predictions were S100A9, TXN, RPS27L, RPS27, RPS27A, RPL38, MRPL54, PPAN, RPS20 and CSRP1 (Supplementary Table 5). For instance, Rodrigues *et al.* suggested that TXN is overexpressed in BC, and it is related to tumor grade, being a key element in redox homeostasis⁸⁶.

Figure 9A shows bean plots comparing the amount of genomic alterations between the OncoOmics BC essential proteins (mean of 133), RBPs related to BC (123), MDPs non-related to BC (115) and non-cancer proteins (85). There was a significant difference ($p < 0.001$) of genomic alterations between RBPs related and non-related to BC after the Mann-Whitney U test. The top 10 MDPs related to BC and with the highest amount of genomic alterations were YWHAZ, DCAF13, TFB2M, PTDSS1, NUCKS1, C1ORF131, DAP3, PABPC1, ZC3H11A and ARF1 (Supplementary Table 8). Additionally, Fig. 9B shows the most altered RNA-binding proteins per genomic alteration type. EIF3H, KMT2C, DCAF13 and EIF2 carried the highest number of CNAs, mutations, mRNA alterations and protein alterations, respectively.

Finally, the prediction of breast cancer proteins related to immunotherapy, metastasis and RNA-binding proteins is a key step to find novel therapeutic targets. For which we suggest multi-omics analyses of these predicted proteins using several databases focused on genomics, transcriptomics and proteomics in human tissues. Additionally, a future study will include the implementation of a web tool that will integrate the entire process predicting proteins with our saved model.

Conclusions

The current study proposed better prediction models for breast cancer proteins using, as inputs, six sets of protein sequence descriptors from Rcp1 and 13 machine-learning classifiers (with or without feature selection/dimension reduction of features). We choose, as the best classifier, the MLP classifier. As inputs, a mixture of 300 selected molecular descriptors has been used: DC, TC and APAAC. The model has a mean AUROC of 0.980 ± 0.0037 and a mean accuracy of 0.936 ± 0.0056 (3-fold cross-validation). 4,504 sequences of proteins related to cancer have been screened for breast cancer relation. Best predicted cancer immunotherapy proteins with BC were RPS27, SUPT4H1, CLPSL2, POLR2K and RPL38, and the most altered ones were POLR2K, ASH2L, MED30, NSL1 and RPRD2. Best predicted metastasis driver proteins with BC were S100A9, DDA1, TXN, PRNP and RPS27, and the most altered ones were YWHAZ, PTK2, SETDB1, EBAG9 and MTBP. Best predicted RNA-binding proteins with BC were S100A9, TXN, RPS27L, RPS27 and RPS27A, and the most altered ones were YWHAZ, DCAF13, TFB2M, PTDSS1 and NUCKS1. Finally, the association between the best-predicted BC proteins using powerful machine-learning methods and the amount of pathogenic genomic alterations in cancer immunotherapy proteins, metastasis driver proteins and RNA-binding proteins gives us candidate proteins that should be deeply studied to find novel therapeutic targets.

Data availability

All data generated during this study are included in this published article including its Supplementary Information files, and the scripts are available as free repository at <https://github.com/muntisa/neural-networks-for-breast-cancer-proteins>.

Received: 2 November 2019; Accepted: 28 April 2020;

Published online: 22 May 2020

References

- López-Cortés, A. *et al.* Breast cancer risk associated with gene expression and genotype polymorphisms of the folate-metabolizing MTHFR gene: a case-control study in a high altitude Ecuadorian mestizo population. *Tumor Biol.* **36**, 6451–6461 (2015).
- López-Cortés, A. *et al.* Mutational Analysis of Oncogenic AKT1 Gene Associated with Breast Cancer Risk in the High Altitude Ecuadorian Mestizo Population. *Biomed Res. Int.* **2018**, 7463832 (2018).
- Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**(305–320), e10 (2018).
- Guerrero, S. *et al.* Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. *Sci. Rep.* **8**, 13978 (2018).
- López-Cortés, A., Guerrero, S., Redal, M. A., Alvarado, A. T. & Quiñones, L. A. State of art of cancer pharmacogenomics in Latin American populations. *Int. J. Mol. Sci.* **18**, 639 (2017).
- Quiñones, L. *et al.* Perception of the Usefulness of Drug/Gene Pairs and Barriers for Pharmacogenomics in Latin America. *Curr. Drug Metab.* **15**, 202–208 (2014).
- López-Cortés, A. *et al.* Pharmacogenomics, biomarker network, and allele frequencies in colorectal cancer. *Pharmacogenomics Journal.* **20**, 136–158 (2020).
- Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
- López-Cortés, A. *et al.* OncoOmics approaches to reveal essential genes in breast cancer: a panoramic view from pathogenesis to precision medicine. *Sci. Rep.* **10**, 5285 (2020).
- Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**(371–385), e18 (2018).
- Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**(321–337), e10 (2018).
- Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33**, 690–705 (2018).
- Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
- Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010).
- Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science.* **347**, 394–403 (2015).
- Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018).
- Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**(564–576), e16 (2017).
- Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
- McFarland, J. M. *et al.* Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* **9**, 1–13 (2018).
- Ivanov, A. A. *et al.* The OncoPPI Portal: An integrative resource to explore and prioritize protein-protein interactions for cancer target discovery. *Bioinformatics.* **34**, 1183–1191 (2018).
- López-Cortés, A. *et al.* Gene prioritization, communality analysis, networking and metabolic integrated pathway to better understand breast cancer pathogenesis. *Sci. Rep.* **8**, 16679 (2018).
- Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385 (2018).
- Thorn, C. F., Klein, T. E. & Altman, R. B. PharmGKB: The pharmacogenomics knowledge base. *Methods Mol. Biol.* **1015**, 311–320 (2013).
- Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **10**, e1417 (2018).
- Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
- Cabrera-Andrade, A. Gene Prioritization through Consensus Strategy, Enrichment Methodologies Analysis, and Networking for Osteosarcoma Pathogenesis. *Int. J. Mol. Sci.* **21**, 1–21 (2020).
- Tejera, E. *et al.* Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis. *BMC Med. Genomics* **10**, 50 (2017).

28. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**, 305–320 (2018).
29. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **23**, 227–238 (2018).
30. Huang, K. lin *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355–370 (2018).
31. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812–830 (2018).
32. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400–416 (2018).
33. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. G:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, 193–200 (2007).
34. Posey, J. E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N. Engl. J. Med.* **376**, 21–31 (2017).
35. Patel, S. J. *et al.* Identification of essential genes for cancer immunotherapy. *Nature* **548**, 537–542 (2017).
36. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
37. Bar-Joseph, Z. *et al.* Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl. Acad. Sci.* **105**, 955–960 (2008).
38. Knijnenburg, T. A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**(239–254), e6 (2018).
39. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Rev. Mol. Cell Biol.* **19**, 327–341 (2018).
40. Carvalho-Silva, D. *et al.* Open Targets Platform: New developments and updates two years on. *Nucleic Acids Res.* **47**, D1056–D1065 (2019).
41. Golbraikh, A., Wang, X. S., Zhu, H. & Tropsha, A. Predictive QSAR modeling: Methods and applications in drug discovery and chemical risk assessment. in *Handbook of Computational Chemistry*. https://doi.org/10.1007/978-3-319-27282-5_37 (2017).
42. Fernández-Blanco, E., Aguiar-Pulido, V., Robert Munteanu, C. & Dorado, J. Random Forest classification based on star graph topological indices for antioxidant proteins. *J. Theor. Biol.* **317**, 331–307 (2013).
43. Munteanu, C. R. *et al.* LECTINPred: Web server that uses complex networks of protein structure for prediction of lectins with potential use as cancer biomarkers or in parasite vaccine design. *Mol. Inform.* **33**, 276–285 (2014).
44. Fernandez-Lozano, C. *et al.* Classification of signaling proteins based on molecular star graph descriptors using Machine Learning models. *J. Theor. Biol.* **384**, 50–58 (2015).
45. Blanco, J. L., Porto-Pazos, A. B., Pazos, A. & Fernandez-Lozano, C. Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection. *Sci. Rep.* **8**, 15688 (2018).
46. Wei, L., Zhou, C., Chen, H., Song, J. & Su, R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **34**, 4007–4016 (2018).
47. Concu, R., Cordeiro, M. N. D. S., Munteanu, C. R. & González-Díaz, H. PTML Model of Enzyme Subclasses for Mining the Proteome of Biofuel Producing Microorganisms. *J. Proteome Res.* **18**, 2735–2746 (2019).
48. Vilar, S., González-Díaz, H., Santana, L. & Uriarte, E. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. *J. Comput. Chem.* **16**, 2613–2622 (2008).
49. Munteanu, C. R., Magalhães, A. L., Uriarte, E. & González-Díaz, H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *J. Theor. Biol.* **257**, 303–311 (2009).
50. Cao, D. S., Xiao, N., Xu, Q. S. & Chen, A. F. RcpI: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics* **31**, 279–281 (2015).
51. Hao, J. & Ho, T. K. Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics* **44**, 348–361 (2019).
52. Jolliffe, I. T. Principal Component Analysis, Second Edition. *Encycl. Stat. Behav. Sci.* (2002).
53. Russell, S. & Norvig, P. *Artificial Intelligence A Modern Approach Third Edition*. Pearson (2010).
54. Cover, T. M. & Hart, P. E. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **13**, 21–27 (1967).
55. Mika, S., Ratsch, G., Weston, J., Scholkopf, B. & Müller, K. R. Fisher discriminant analysis with kernels. in *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop* (1999).
56. Patle, A. & Chouhan, D. S. SVM kernel functions for classification. in *2013 International Conference on Advances in Technology and Engineering, ICATE 2013* (2013).
57. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**, 1373–1379 (1996).
58. White, B. W. & Rosenblatt, F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. *Am. J. Psychol.* (1963).
59. Swain, P. H. & Hauska, H. DECISION TREE CLASSIFIER: DESIGN AND POTENTIAL. *IEEE Trans Geosci Electron* (1977).
60. Breiman L. *Machine Learning*, 45(1), 5–32. Stat. Dep. Univ. California, Berkeley, CA 94720. (2001).
61. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System (2016).
62. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
63. Hughes, G. F. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Trans. Inf. Theory* **14**, 55–63 (1968).
64. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
65. Rocco, P. *et al.* OncoScore: A novel, Internet-based tool to assess the oncogenic potential of genes. *Sci. Rep.* **7**, 46290 (2017).
66. Zheng, G. *et al.* HCMDDB: The human cancer metastasis database. *Nucleic Acids Res.* **46**, 950–955 (2018).
67. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
68. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
69. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, 11 (2013).
70. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
71. Finotello, F., Rieder, D., Hackl, H. & Trajanoski, Z. Next-generation computational tools for interrogating cancer immunity. *Nat. Rev. Genet.* **20**, 724–746 (2019).
72. Atsuta, Y. *et al.* Identification of metalloproteinase-1 as a member of a tumor associated antigen in patients with breast cancer. *Cancer Lett.* **182**, 101–107 (2002).
73. Itamochi, H. *et al.* Whole-genome sequencing revealed novel prognostic biomarkers and promising targets for therapy of ovarian clear cell carcinoma. *Br. J. Cancer* **5**, 717–724 (2017).
74. Angus, L. *et al.* The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nat. Genet.* **51**, 1450–1458 (2019).
75. Caicedo, A. *et al.* MitoCeption as a new tool to assess the effects of mesenchymal stem/stromal cell mitochondria on cancer cell metabolism and function. *Sci. Rep.* **5**, 9073 (2015).
76. Aponte, P. M. & Caicedo, A. Stemness in cancer: Stem cells, cancer stem cells, and their microenvironment. *Stem Cells International* **2017**, 5619472 (2017).

77. Fokas, E., Engenhardt-Cabillic, R., Daniilidis, K., Rose, F. & An, H. X. Metastasis: The seed and soil theory gains identity. *Cancer and Metastasis Reviews* **26**, 3–4 (2007).
78. Schell, M. J. *et al.* A composite gene expression signature optimizes prediction of colorectal cancer metastasis and outcome. *Clin. Cancer Res.* **22**, 734–745 (2016).
79. Lee, J. Y. *et al.* Mutational profiling of brain metastasis from breast cancer: Matched pair analysis of targeted sequencing between brain metastasis and primary breast cancer. *Oncotarget* **6**, 43731–43742 (2015).
80. Bergenfelz, C. *et al.* S100A9 expressed in ER-PgR-breast cancers induces inflammatory cytokines and is associated with an impaired overall survival. *Br. J. Cancer* **113**, 1234–1243 (2015).
81. García-cárdenas, J. M. *et al.* Post-transcriptional Regulation of Colorectal Cancer: A Focus on RNA-Binding. *Proteins*. **6**, 1–18 (2019).
82. Burd, C. G. & Dreyfuss, G. Conserved structures and diversity of functions of RNA-binding proteins. *Science* **265**, 615–621 (1994).
83. Lukong, K. E. & Chang, K. wei, Khandjian, E. W. & Richard, S. RNA-binding proteins in human genetic disease. *Trends in Genetics* **24**, 416–425 (2008).
84. Kechavarzi, B. & Janga, S. C. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol.* **15**, R14 (2014).
85. Guerrero, S. *et al.* In silico analyses reveal new putative Breast Cancer RNA-binding proteins. *bioRxiv* (2020).
86. Rodrigues, P. *et al.* Oxidative stress in susceptibility to breast cancer: Study in Spanish population. *BMC Cancer* **14**, 861 (2014).

Acknowledgements

This work was supported by a) Universidad UTE (Ecuador), b) the Collaborative Project in Genomic Data Integration (CICLOGEN) PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013-2016 and the European Regional Development Funds (FEDER) - “A way to build Europe”; c) the General Directorate of Culture, Education and University Management of Xunta de Galicia ED431D 2017/16 and “Drug Discovery Galician Network” Ref. ED431G/01 and the “Galician Network for Colorectal Cancer Research” (Ref. ED431D 2017/23); d) the Spanish Ministry of Economy and Competitiveness for its support through the funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union; e) the Consolidation and Structuring of Competitive Research Units - Competitive Reference Groups (ED431C 2018/49), funded by the Ministry of Education, University and Vocational Training of the Xunta de Galicia endowed with EU FEDER funds; f) research grants from Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P), Basque government (IT1045-16), and kind support of Ikerbasque, Basque Foundation for Science; and, g) Sociedad Latinoamericana de Farmacogenómica y Medicina Personalizada (SOLFAGEM).

Author contributions

A.L.-C., A.C.-A. and C.R.M. conceived the subject, the conceptualization of the study and wrote the manuscript. A.L.-C., A.C.-A., J.M.V.-N. and C.R.M. did data curation and supplementary data. C.R.M. and J.M.V.-N. built the models using machine learning. A.P., H.G.-D., C.P.-y-M., S.G., Y.P.-C. and E.T. gave conceptual advice and valuable scientific input. A.P., H.G.-D., C.P.-y-M., S.G., Y.P.-C., E.T. and C.R.M. supervised the project. A.L.-C. and C.P.-y-M. did funding acquisition. Finally, all authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-65584-y>.

Correspondence and requests for materials should be addressed to A.L.-C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020