# Técnicas de machine learning aplicadas al diagnóstico y tratamiento oncológico de precisión mediante el análisis de datos ómicos

**José Liñares Blanco**

Tesis Doctoral UDC / 2021

**Directores:**
Dr. Carlos Fernández Lozano
Dr. José Antonio Seoane Fernández

Programa oficial de doctorado en Tecnologías de la Información y las Comunicaciones

UNIVERSIDADE DA CORUÑA

*Escola Internacional de Doutoramento*
*Vicerreitoría de Investigación e Transferencia*
UNIVERSIDADE DA CORUÑA

**Dr. Carlos Fernández Lozano**, Profesor Ayudante Doctor del área de Ciencias de la Computación e inteligencia Artificial de la Universidade da Coruña.

**Dr. José Antonio Seoane Fernández**, Junior Group Leader en el Vall d'Hebrón Instituto de Oncología.

**HACEN CONSTAR QUE:**

La tesis titulada **Técnicas de machine learning aplicadas al diagnóstico y tratamiento oncológico de precisión mediante el análisis de datos ómicos** realizada por D. José Liñares Blanco con DNI 47404120-Q, bajo nuestra dirección en el Departamento de Ciencias de la Computación y Tecnologías de la Información cumple los requisitos para optar al grado de Doctor y obtener la mención internacional.

Y para que así conste, firma esta autorización en A Coruña, a 19 de enero de 2022

Los directores de la tesis

Fdo. Carlos Fernández Lozano

Fdo. José Antonio Seoane Fernández

A mis padres,
por dármelo todo
a cambio de nada.

# Agradecimientos

Gracias por formarme, guiarme y aconsejarme a lo largo de estos años. Gracias por haber sido siempre honestos y sinceros. Gracias por hacer vuestro trabajo de la mejor manera posible. Por enseñarme. Por darme la oportunidad de crecer en lo que me gusta. Seguiré este camino con la ilusión y el respecto que vosotros me enseñasteis. Gracias Carlos. Gracias Jose.

Gracias también a ti, Alejandro. Por haber confiado en mí y darme la oportunidad en un primer momento. Gracias al grupo RNASA-IMEDIR. A la gente del laboratorio, que habéis compartido conmigo muchos momentos. Por el apoyo y el ambiente durante estos cuatro años.

A vosotras, por hacer que este camino haya sido realmente fácil y divertido a vuestro lado. Por los innumerables cafés y las *turras* que los acompañaron. Por las risas y los agobios. Por crecer juntos. Gracias por estar ahí siempre, por empezar siendo mis compañeras y acabar siendo mi familia. Gracias Sara. Gracias Nere.

Gracias por creer en mí, por confiar en todo momento, por dejarme elegir. Por apostar por algo que ni siquiera entendíais. Gracias por el apoyo constante, por el cariño y por el amor. Por enseñarme vuestros valores. Gran parte de mi motivación y mi esfuerzo ha sido por vosotros. Gracias Papá. Gracias Mamá.

Gracias a mi entorno. A los que han estado a mi lado en todo o parte de este camino. Habéis sido mi apoyo y mi desahogo. Gracias por hacerme ver las cosas importantes de esta vida. Sois los mejores.

A todos vosotros, muchas gracias.

Esta tesis, en parte, es vuestra.

*Nature is always more subtle, more intricate,*
*more elegant than what we are able to imagine.*

**— Carl Sagan**

# Resumen

Gracias al abaratamiento en los costes de secuenciación, cada día se genera una mayor cantidad de datos ómicos capaces de caracterizar el cáncer molecularmente. Grandes consorcios generan gran cantidad de estos datos, poniéndolos a disposición pública. Además, los modelos de Machine Learning (ML) ofrecen una ventaja significativa para extraer patrones complejos de datos biomédicos. Se requiere un estudio de su aplicación en este campo para poder obtener resultados más robustos y generalizados.

Esta tesis estudia la aplicación de modelos de ML para el análisis de datos ómicos. Gracias a una revisión de trabajos previos, se identificaron ciertas limitaciones en cuanto reproducibilidad y validación en las metodologías. A partir de este estudio se establecieron las directrices para llevar a cabo un análisis de ML robusto y reproducible con datos ómicos. Se identificaron biomarcadores y *pathways* alterados en pacientes con cáncer de colon, se predijeron condiciones clínicas relevantes para el desarrollo del tumor y se desarrolló un modelo de *screening* automático de fármacos antitumorales. Los resultados se presentan en un compendio de tres publicaciones científicas.

En conclusión, esta tesis ofrece diferentes aproximaciones computacionales que ayudan al diagnóstico y al tratamiento oncológico de precisión.

# Abstract

As sequencing costs have been dramatically reduced, an increasing amount of omics data have been generated to molecularly characterise cancer. Large consortiums are generating large amount of this data and making them publicly available. In addition, Machine Learning (ML) models offer a significant advantage extracting complex patterns from biomedical data. A study of their application in this field is necessary in order to obtain more robust and generalised results.

This thesis studies the application of ML models to omics data analysis. Performing a review of previous work, certain limitations in terms of reproducibility and validation of the methodologies were identified. From this study, a set of guidelines for robust and reproducible ML analysis of omics data have been established, allowing to identify altered biomarkers and pathways in colon cancer patients, predict clinical conditions relevant to tumour development, and develop an automatic anti-tumour drug screening model. These results are presented as a compendium of three scientific manuscripts.

In conclusion, this thesis provides a variety of computational approaches to improve diagnosis and precision oncological treatment.

# Resumo

Grazas aos menores custos de secuenciación, cada día xéranse máis datos ómicos capaces de caracterizar molecularmente o cancro. Grandes consorcios están a xerar gran cantidade destes datos de forma pública. Ademais, os modelos de Machine Learning (ML) ofrecen unha vantaxe significativa para extraer complexos patróns de datos biomédicos. É necesario un estudo da súa aplicación neste campo para obter resultados máis robustos e xeneralizados.

Esta tese estuda a aplicación de modelos de ML para a análise de datos ómicos. Grazas a unha revisión de traballos anteriores, identificáronse certas limitacións en termos de reprodutibilidade e validación nas metodoloxías. A partir deste estudo, establecéronse pautas para realizar unha análise de ML robusta e reproducible con datos ómicos. Identificáronse biomarcadores e vías alteradas en pacientes con cancro de colon, predixéronse condicións clínicas relevantes para o desenvolvemento tumoral e desenvolveuse un modelo de detección automática de medicamentos antitumorais. Os resultados preséntanse nun compendio de tres publicacións científicas.

En conclusión, esta tese ofrece diferentes enfoques computacionais que axudan ao diagnóstico e tratamento preciso do cancro.

# Estructura de la Tesis Doctoral

Esta tesis se compone de siete capítulos. El **Capítulo 1** presenta una introducción del trabajo, dividida en cuatro secciones donde se describen los pilares fundamentales de la investigación: el cáncer, la caracterización del cáncer a través de datos ómicos, y su análisis mediante técnicas de Machine Learning y otras técnicas computacionales. El **Capítulo 2** enumera los objetivos propuestos que motivaron el desarrollo de la tesis.

A continuación, el **Capítulo 3** se divide en tres secciones. Cada sección corresponde a uno de los artículos del compendio de la tesis. Tras una breve explicación del contexto y contenido del artículo, se adjunta el documento publicado. En conjunto, los tres artículos presentan los resultados de esta tesis doctoral, alcanzando los objetivos propuestos en el capítulo anterior. Las referencias de los artículos, por orden de aparición, son las siguientes:

- **Liñares-Blanco, J.**, Pazos, A. and Fernandez-Lozano, C.. *Machine learning analysis of TCGA cancer data.* PeerJ Computer Science 7 (2021). Q1, 3.091 IF. https://doi.org/10.7717/peerj-cs.584
- **Liñares-Blanco, J.**, Munteanu, C.R., Pazos, A. et al. *Molecular docking and machine learning analysis of Abemaciclib in colon cancer.* BMC Mol and Cell Biol 21, 52 (2020). Q3, 3.066 IF. 3 citas (Google Scholar). https://doi.org/10.1186/s12860-020-00295-w
- **Liñares-Blanco, J.**, Porto-Pazos, A.B., Pazos, A. et al. *Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection.* Sci Rep 8, 15688 (2018). Q1, 4.011 IF. 21 citas (Google Scholar). https://doi.org/10.1038/s41598-018-33911-z

En el **Capítulo 4** se presenta una discusión transversal a los tres artículos, dándole coherencia, cohesión y justificación al trabajo de investigación desarrollado.

Los **Capítulos 5** y **6** enumeran las conclusiones en castellano e inglés, respectivamente. El **Capítulo 7**, muestra los futuros desarrollos propuestos tras la investigación realizada a lo largo de la tesis.

# Prefacio

El ser humano es una simbiosis de billones de células eucariotas y procariotas. Cada célula eucariota humana tiene una secuencia ADN de alrededor de 3.000 millones de pares de bases distribuida en 23 pares de cromosomas, donde se encuentran un total de 20.000 genes, aproximadamente. Estos genes, que se convertirán en proteínas, producen de forma intermediaria entre $10^5$ y $10^6$ moléculas de mRNA, las cuales posteriormente darán lugar alrededor de 100.000 proteínas diferentes, expresadas de forma específica en cada célula y tejido. Las posibles alteraciones que pueden darse en este proceso de flujo de información genética son incalculables. Además, la interacción que tiene el metabolismo celular con el ambiente (microorganismos, exposiciones ambientales, nutrición, etc.) y toda la maquinaria de regulación molecular que existe en regiones no codificantes del genoma genera una inmensa complejidad.

Cuando nos encontramos enfrente de patologías complejas y heterogéneas, como es el cáncer, son muchos y muchas veces desconocidos los genes que están involucradas en la aparición del fenotipo patológico. Es por ello, que uno de los retos más importantes en el campo de la biomedicina reside en el diagnóstico molecular de cada paciente, en un tiempo lo suficientemente temprano como para poder realizar su tratamiento de forma precisa. En este contexto, la búsqueda de un diagnóstico genético y un posible tratamiento personalizado se convierte en una tarea inabarcable para un humano.

En los últimos años, técnicas dentro de la rama de la Inteligencia Artificial, como son los algoritmos de Machine Learning, han revolucionado el campo de la biomedicina, al ofrecer herramientas capaces de analizar e interpretar toda la cantidad de datos que se generan en este campo.

La presente tesis desarrolla un trabajo de investigación sobre la aplicabilidad de las técnicas de Machine Learning en el análisis de datos ómicos, con el fin de identificar nuevas formas de diagnóstico y tratamiento de precisión.

En primer lugar, se realiza un estudio de trabajos previos que hayan utilizado técnicas de ML para el análisis de datos ómicos [1]. Se utiliza el repositorio de datos públicos multi-ómicos del The Cancer Genome Atlas (TCGA) para realizar el estudio comparativo. Se identifican y analizan más de 100 artículos donde se utiliza una aproximación de ML para el análisis de los datos albergados en el TCGA. Se estudia la metodología, los modelos y los tipos de datos utilizados por cada uno de los trabajos. Se identifican las principales debilidades y fortalezas en las metodologías utilizadas, para posteriormente utilizar de forma precisa modelos de ML sobre datos ómicos [2, 3].

Gracias a las conclusiones obtenidas en el artículo de revisión, se identifican firmas de expresión genética con valor prognóstico en cáncer de colon. Siguiendo una rigurosa metodología de ML, se estudia el potencial predictivo de las firmas en diferentes condiciones clínicas. Un análisis del modelo, de la expresión diferencial específica de tejido, de supervivencia y una revisión de la literatura, propone al gen FABP6 como potencial candidato a biomarcador en cáncer de colon. Posteriormente se realiza un análisis de *docking* molecular para estudiar las interacciones entre fármacos anti-tumorales ya aprobados y diferentes dianas moleculares en cáncer de colon. Las isoformas del gen FABP6 mostraron una significativa interacción con el fármaco Abemaciclib. Por lo tanto, también se identifica un posible reposicionamiento del fármaco Abemaciclib (utilizado en cáncer de mama) para la inhibición de las isoformas del producto génico de FABP6 [2]. Estos resultados son propuestos para validación experimental.

Motivado por la búsqueda de nuevas formas de tratamiento, se estudia el proceso de angiogénesis en cáncer. La angiogénesis es un proceso por el que las células inducen la creación de nuevos vasos sanguíneos a partir de otros ya existentes. En condiciones normales está altamente regulado. Por el contrario, en situaciones cancerígenas, las mismas células del tumor inducen el proceso metabólico para desarrollar nuevos vasos sanguíneos dentro del microambiente tumoral. De esta manera, las células del tumor reciben constantemente oxígeno y nutrientes para su proliferación. La inhibición de este proceso provoca el cese del crecimiento del tumor. Aunque es una diana terapéutica para el desarrollo de nuevos fármacos, muy pocos han sido aprobados debido al tiempo y dinero necesarios. Con el fin de abordar este problema, se desarrolla un modelo basado en ML capaz de predecir la actividad anti-angiogénica de péptidos de una forma rápida, sencilla y barata [3]. Este modelo se desarrolla a partir de descriptores moleculares, superando en rendimiento a los modelos del estado del arte. Un estudio en profundidad del modelo muestra secuencias específicas del péptido tienen un rol principal en la actividad anti-angiogénica de péptidos.

# Tabla de contenidos

**1 Introducción** **1**

1.1 Fundamentos del cáncer . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.1.1 Definición de cáncer . . . . . . . . . . . . . . . . . . . . . . . . 1

1.1.2 Perfil genético del cáncer . . . . . . . . . . . . . . . . . . . . . . 2

1.1.3 El cáncer como una enfermedad multi-ómica . . . . . . . . . 4

1.2 Tecnologías de secuenciación masiva . . . . . . . . . . . . . . . . . . 7

1.2.1 Obtención de datos genómicos . . . . . . . . . . . . . . . . . . 8

1.2.2 Obtención de datos transcriptómicos . . . . . . . . . . . . . . . 9

1.2.3 Obtención de datos proteómicos . . . . . . . . . . . . . . . . . 10

1.3 The Cancer Genome Atlas . . . . . . . . . . . . . . . . . . . . . . . . . 11

1.4 Técnicas computacionales utilizadas en la tesis . . . . . . . . . . . 13

1.4.1 Machine Learning . . . . . . . . . . . . . . . . . . . . . . . . . . 13

1.4.1.1 Visión general del aprendizaje no supervisado . . . . 14

1.4.1.2 Visión general del aprendizaje supervisado . . . . . 14

1.4.1.3 Modelos de Machine Learning para aprendizaje supervisado . . . . . . . . . . . . . . . . . . . . . . . . 15

1.4.1.3.1 Modelos lineales . . . . . . . . . . . . . . . 17

1.4.1.3.2 Modelos basados en distancias . . . . . . . 19

1.4.1.3.3 Modelos basados en árboles . . . . . . . . . 21

1.4.1.3.4 Modelos basados en *kernels* . . . . . . . . . 24

1.4.1.4 Métricas de rendimiento . . . . . . . . . . . . . . . . 26

1.4.1.5 Preprocesado de los datos . . . . . . . . . . . . . . . 28

1.4.1.6 Validación del modelo . . . . . . . . . . . . . . . . . 30

1.4.1.7 Selección del mejor modelo . . . . . . . . . . . . . . 31

1.4.2 Técnicas de selección de características . . . . . . . . . . . . . 32

1.4.2.1 Técnicas filter . . . . . . . . . . . . . . . . . . . . . . . 33

1.4.2.2 Técnicas wrapper . . . . . . . . . . . . . . . . . . . . 34

1.4.2.3 Técnicas embedded . . . . . . . . . . . . . . . . . . . 35

1.4.3 Modelos QSAR . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

1.4.3.1 Descriptores de composición aminoacídica . . . . . . 37

1.4.4 Técnicas de *docking* molecular para el reposicionamiento de fármacos . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39

2 Objetivos 43

3 Resultados 45

3.1 Revisión y estado del arte del uso de técnicas de Machine Learning en oncología . . . 45

3.2 Búsqueda de nuevas dianas terapéuticas para el tratamiento del cáncer de colon . . . 94

3.3 Predicción de fármacos para el tratamiento del cáncer mediante técnicas de Machine Learning . . . 114

4 Discusión 127

5 Conclusiones 135

6 Conclusions 137

7 Futuros desarrollos 139

Referencias 143

Anexo A Producción científica 153

A.1 Compendio de tres artículos JCR (WOS) . . . 153

A.2 Otros artículos JCR (WOS) (2) . . . 153

A.3 Participación en congresos nacionales e internacionales (7) . . . 154

A.4 Capítulos de libro (1) . . . 155

# Índice de figuras

1.1 Los rasgos del cáncer . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Esquema del flujo de información genética . . . . . . . . . . . . . . . . . 5

1.3 Datos ómicos en el repositorio TCGA . . . . . . . . . . . . . . . . . . . . . 12

1.4 Estrategias para el ajuste de hiperparámetros . . . . . . . . . . . . . . . 16

1.5 Algoritmo k-NN . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

1.6 Algoritmo de árbol de decisión . . . . . . . . . . . . . . . . . . . . . . . . 21

1.7 Algoritmo Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

1.8 Hiperplano óptimo de separación de SVM . . . . . . . . . . . . . . . . . 24

1.9 Tipos de kernels en SVM . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

1.10 Matriz de confusión . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

1.11 *Cross-Validation* anidada . . . . . . . . . . . . . . . . . . . . . . . . . . 31

1.12 Técnicas de Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . 33

1.13 Metodología QSAR . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

1.14 Proceso de descubrimiento de fármacos . . . . . . . . . . . . . . . . . . 40

1.15 Docking molecular . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

# Índice de ecuaciones

1.1 Ecuación general de regresión lineal. . . . . . . . . . . . . . . . . . . . . 17

1.2 Función objetivo del algoritmo glmnet en regresión logística . . . . . . . 18

1.3 Distancia Euclidea . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

1.4 Distancia de Manhattan . . . . . . . . . . . . . . . . . . . . . . . . . 19

1.5 Distancia de Minkowski . . . . . . . . . . . . . . . . . . . . . . . . . 19

1.6 Ecuación para el ajuste del algoritmo k-NN . . . . . . . . . . . . . . . . 20

1.7 Cálculo de la importancia de una variable en el algoritmo Random Forest 23

1.8 Cálculo de la *accuracy* . . . . . . . . . . . . . . . . . . . . . . . . . 27

1.9 Cálculo del *recall* . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

1.10 Cálculo de la *precision* . . . . . . . . . . . . . . . . . . . . . . . . . 28

1.11 Cálculo del *F measure* . . . . . . . . . . . . . . . . . . . . . . . . . 28

1.12 Cálculo del AUC . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28

1.13 Amino Acid Composition . . . . . . . . . . . . . . . . . . . . . . . . 38

1.14 Di-Amino Acid Composition . . . . . . . . . . . . . . . . . . . . . . 38

1.15 Tri-Amino Acid Composition . . . . . . . . . . . . . . . . . . . . . . 38

# Abreviaturas

- **NGS**: Next Generation Sequencing
- **HUGO**: Human Genome Organization
- **DNA**: Desoxirribonucleic acid
- **cDNA**: Complementary DNA
- **RNA**: Ribonucleic acid
- **mRNA**: Messenger RNA
- **RNAi**: RNA interference
- **tRNA**: Transfer RNA
- **rRNA**: Ribosomal RNA
- **miRNA**: Micro-RNA
- **SNP**: Single-nucleotide Polymorphism
- **CNV**: Copy Number Variation
- **LOH**: Loss of heterozygosity
- **GC**: Gas Chromatography
- **LC**: Liquid Chromatography
- **NCI**: National Cancer Institute
- **NHGRI**: National Human Genome Research Institute Home
- **TCGA**: The Cancer Genome Atlas
- **GBM**: Glioblastoma multiforme
- **LUSC**:Lung squamous cell carcinoma
- **OV**: Ovarian serous cystadenocarcinoma
- **ACC**: Adrenocortical carcinoma
- **BLCA**: Bladder Urothelial Carcinoma
- **BRCA**: Breast invasive carcinoma
- **CESC**: Cervical squamous cell carcinoma and endocervical adenocarcinoma
- **CHOL**: Cholangiocarcinoma
- **COAD**: Colon adenocarcinoma
- **DLBC**: Lymphoid Neoplasm Diffuse Large B-cell Lymphoma

- **ESCA**: Esophageal carcinoma
- **FPPP**: FFPE Pilot Phase II
- **HNSC**: Head and Neck squamous cell carcinoma
- **KICH**: Kidney Chromophobe
- **KIRC**: Kidney renal clear cell carcinoma
- **KIRP**: Kidney renal papillary cell carcinoma
- **LAML**: Acute Myeloid Leukemia
- **LGG**: Brain Lower Grade Glioma
- **LIHC**: Liver hepatocellular carcinoma
- **LUAD**: Lung adenocarcinoma
- **MESO**: Mesothelioma
- **PAAD**: Pancreatic adenocarcinoma
- **PCPG**: Pheochromocytoma and Paraganglioma
- **PRAD**: Prostate adenocarcinoma
- **READ**: Rectum adenocarcinoma
- **SARC**: Sarcoma
- **SKCM**: Skin Cutaneous Melanoma
- **STAD**: Stomach adenocarcinoma
- **STES**: Stomach and Esophageal carcinoma
- **TGCT**: Testicular Germ Cell Tumors
- **THYM**: Thymoma
- **UCEC**: Uterine Corpus Endometrial Carcinoma
- **UCS**: Uterine Carcinosarcoma
- **UVM**: Uveal Melanoma
- **GDC**: Genomic Data Commons
- **ML**: Machine Learning
- **QSAR**: Quantitative strucuture-activity relationship
- **k-NN**: k-Nearest Neighbors
- **RF**: Random Forest
- **SVM**: Support Vector Machines
- **RBF**: Radial Basis Function
- **NZV**: Near Zero Variants
- **NA**: Not Available
- **AUC-ROC**: Area under the ROC curve
- **ROC**: Receiver Operating Characteristic
- **VP**: Verdadero Positivo
- **FN**: Falso Negativo
- **VN**: Verdadero Negativo

- **FP**: Falso Positivo
- **Acc**: Accuracy
- **CV**: Cross-Validation
- **MI**: Mutual Information
- **LASSO**: Least Absolute Shrinkage and Selection Operator
- **FDA**: Food and Drug Administration
- **HTS**: High-throughput screening

# 1 Introducción

Se elabora este capítulo con el fin de mostrar una justificación razonada de la unidad y coherencia temática y metodológica de la tesis. Se aborda y se profundiza en los pilares en los que se basa el trabajo: el cáncer, los datos ómicos y las técnicas de Machine Learning.

## 1.1 Fundamentos del cáncer

El cáncer es una emergencia sanitaria. A día de hoy, es la segunda causa de muerte más frecuente en el mundo, provocando la muerte de casi 10 millones de personas cada año y con una estimación de 19 millones de nuevos casos al año [4]. Se han llevado a cabo exhaustivos esfuerzos por la comunidad científica para entender la biología subyacente y proponer nuevos y eficaces métodos de diagnóstico y tratamiento. Aún con los avances experimentados en la última década, las estadísticas de incidencia y mortalidad dejan en evidencia las necesidades de nuevas aproximaciones que ayuden a los pacientes. La motivación para el desarrollo de esta tesis ha sido la aplicación y estudio de nuevas aproximaciones computacionales que apoyen el diagnóstico y ayuden a la búsqueda de nuevos tratamientos personalizados en cáncer.

Para contextualizar esta investigación, en primer lugar debemos entender qué es el cáncer, por qué aparece y cómo se desarrolla.

### 1.1.1 Definición de cáncer

El cáncer es una enfermedad compleja y multifactorial. Aunque existen muchos tipos de cáncer, todos ellos se caracterizan por un anormal y descontrolado crecimiento celular causado por una constante adquisición de alteraciones genéticas [5]. En conjunto, afectan al correcto funcionamiento de las células, provocando una invasión en otros tejidos y eventualmente la muerte del individuo.

Es importante diferenciar entre los términos tumor y cáncer. Hablamos de tumor cuando nos referimos a un conjunto de células que se han dividido descontroladamente, formando una masa celular que en condiciones normales no aparecería. Cuando estas células adquieren la capacidad de invasión en tejidos adyacentes y/o lejanos, hablamos de tumores malignos, o más comúnmente conocidos como cáncer.

Los tumores malignos pueden subclasificarse además en tumores primarios y tumores metastásicos. Los primeros crecen y se desarrollan en el tejido de origen, mientras que los segundos pueden desplazarse a localizaciones lejanas del cuerpo para crecer y desarrollarse.

A medida que las células cancerígenas se van multiplicando, la masa tumoral va adquiriendo un conjunto de características fenotípicas que facilita su continuo crecimiento. Hanahan y Weinberg denominaron estas características como rasgos del cáncer [6, 7]. La Figura 1.1 es un extracto del artículo publicado en 2011 por ambos autores donde se identifican los 10 rasgos que caracterizan un cáncer. Se observa además, que cada rasgo se identifica también como una diana o estrategia para los diferentes fármacos utilizados en terapias oncológicas. Algunas de estas terapias serán abordadas en posteriores capítulos de esta tesis.

### 1.1.2 Perfil genético del cáncer

Un tumor debe concebirse como una población de diferentes líneas celulares (o subpoblaciones) que están bajo una selección Darwiniana [8]. Cada subpoblación celular se caracteriza por una serie de alteraciones que son adquiridas a lo largo del tiempo. Las células individuales pertenecientes a cada subpoblación van adquiriendo variación genética mediante cambios aleatorios. Estas alteraciones se van heredando mediante la replicación de las células. Los linajes celulares con mejor adaptación, son los que sobreviven. Por lo tanto, los rasgos fenotípicos, anteriormente presentados, son consecuencias del efecto de la Selección Natural en las células tumorales.

Hablamos de alteración genética al referirnos, en general, a las posibles variaciones que puedan llegar a ocurrir en el proceso de información genética. Por el contrario, una mutación genética hace referencia a un cambio en la secuencia del DNA de una célula. Esta modificación puede ser debida a un cambio entre bases nucleotídicas, una deleción ó una inserción.

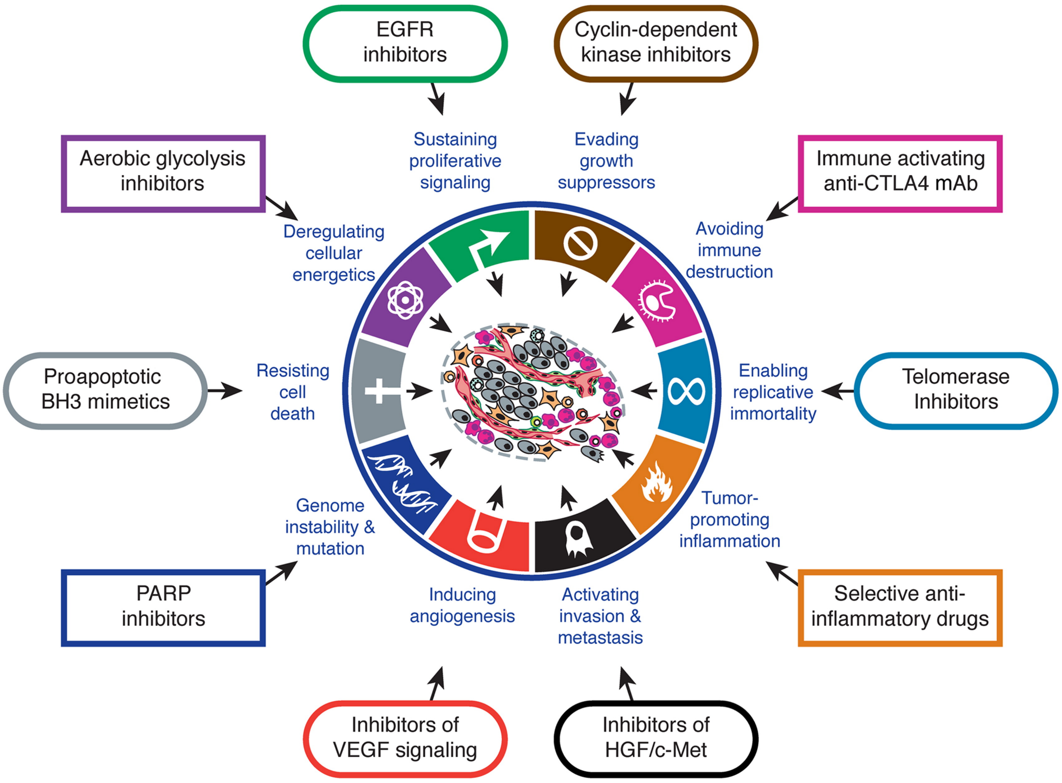

EGFR inhibitors
Cyclin-dependent kinase inhibitors
Aerobic glycolysis inhibitors
Sustaining proliferative signaling
Evading growth suppressors
Immune activating anti-CTLA4 mAb
Deregulating cellular energetics
Avoiding immune destruction
Proapoptotic BH3 mimetics
Resisting cell death
Enabling replicative immortality
Telomerase Inhibitors
Genome instability & mutation
Tumor-promoting inflammation
PARP inhibitors
Inducing angiogenesis
Activating invasion & metastasis
Selective anti-inflammatory drugs
Inhibitors of VEGF signaling
Inhibitors of HGF/c-Met

Fig. 1.1.: Los rasgos del cáncer comprenden diez capacidades biológicas adquiridas durante el desarrollo del tumor [7]. Cada una de ellas es una diana terapéutica de los diferentes fármacos anti-tumorales desarrollados, resaltados en la figura.

Se pueden diferenciar dos tipos de mutaciones: conductoras y pasajeras. Las primeras, confieren a la célula ciertas ventajas adaptativas con respecto a las células vecinas [8]. Por el contrario, las mutaciones pasajeras no están involucradas en la aparición del tumor, aunque pueden presentar un rol en la aparición y desarrollo del tumor [9]. Publicaciones previas han reportado que estas últimas pueden ser predictoras entre los diferentes tipos de tumores [10].

El hecho de que ciertas mutaciones confieren directamente algún tipo de ventaja adaptativa a las poblaciones celulares del tumor, es debido a la localización de la mutación en el genoma. Estas mutaciones afectan a genes que tienen un rol principal en ciertas tareas celulares. Dichos genes se denominan conductores y pueden clasificarse en oncogenes y genes supresores del tumor. A día de hoy se conoce un gran número de genes conductores en cada uno de los tumores [11, 12, 13].

Ser capaces de identificar los cambios completos que genera cada tipo de cáncer en el DNA (en su genoma) y conseguir entender cómo estos cambios interactúan para que se manifieste la enfermedad puede sentar las bases para lograr grandes avances en prevención, detección temprana, estratificación y éxito del tratamiento del cáncer.

### 1.1.3 El cáncer como una enfermedad multi-ómica

El término ***ómico*** proviene del sufijo de origen latino *-oma*, que significa *conjunto de*. En biología, se utiliza este sufijo para referirse a la totalidad o al conjunto de algo. La posibilidad de generación e interpretación de conjuntos de datos biológicos, ha abierto la puerta a una nueva mentalidad en la que se desarrolla una visión global de los procesos biológicos, y se ve reflejada en el desarrollo de lo que se ha denominado como **La era ómica**.

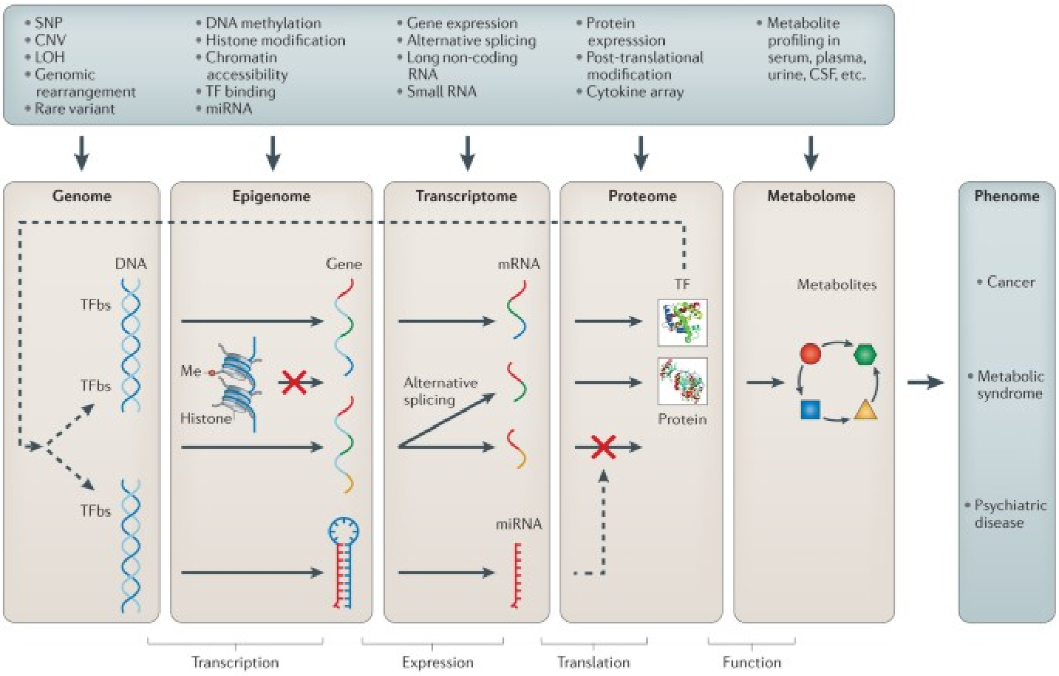
La investigación oncológica ha experimentado un crecimiento exponencial en la última década gracias a la caracterización *ómica* de los tumores. De esta forma, podemos tener información de cada uno de los procesos involucrados en el Dogma Central de la Biología Molecular [14].

En la sección anterior, se ha definido la causa del cáncer como una adquisición constante de alteraciones genéticas. La consecuencia del conjunto de estas alteraciones será una proteína, parcial o completamente defectuosa, que no hará correctamente su función, provocando un descontrol en la proliferación celular. Estas alteraciones, en genes conductores son lo que conocemos como los rasgos del cáncer.

Desgraciadamente, la identificación de estas alteraciones no es suficiente para generar un diagnóstico específico, y muchos pacientes no se ven beneficiados por tratamientos generales. Lo necesario en este caso son tratamientos oncológicos de precisión diseñados concretamente para un subgrupo de pacientes que compartan ciertas características moleculares. Debido a ello, la caracterización molecular del cáncer debe centrarse no sólo en alteraciones de la secuencia del DNA, si no en todos los niveles del Dogma Central de la Biología Molecular [14].

Por lo tanto, la detección de los rasgos del cáncer, actualmente, no se centra únicamente en la secuencia de DNA, si no también en modificaciones en el epigenoma, transcriptoma, proteoma ó metaboloma, entre otros.

En la Figura 1.2 se muestra una simplificación de como el flujo de información genética va desde el genotipo de un individuo (genoma) hasta la presencia de un determinado fenotipo (fenoma). Los datos extraídos en cada uno de estos niveles, a partir de plataformas biotecnológicas, ofrecen a los investigadores más datos para comprender la enfermedad.



Nature Reviews | Genetics

Fig. 1.2.: Esquema simplificado del flujo de información genética. En la parte superior se indican los diferentes tipos de datos que pueden ser extraídos en cada uno de los pasos. Por ejemplo, del genoma se pueden extraer datos de polimorfismos (SNP), número de copias (CNV), pérdida de heterocigosidad (LOH), reorganización genética o variantes raras. Las flechas indican la dirección de la información genética, desde el genotipo hasta finalmente expresarse el fenotipo. Las cruces rojas indican inactivación. Por ejemplo, modificaciones epigenéticas como la metilación de ciertos genes provoca una inactivación de la expresión de dicho gen. Imagen extraída de [15].

Cada tipo de dato ómico presenta diferencias moleculares asociadas al cáncer (en el genoma [8], en el epigenoma [16], en el transcriptoma [17], en el proteoma [18], en el metaboloma [19] e incluso en el microbioma [20]), revelando marcadores útiles del proceso de la enfermedad y aportando conocimientos sobre las vías biológicas. Pero los datos ómicos por separado no son suficientes para revelar la relación causal entre las firmas moleculares y la manifestación del cáncer [15].

Los enfoques que integran diferentes tipos de datos ómicos tienen el potencial de descubrir esos cambios causales y de mejorar la comprensión de la variabilidad en la respuesta al tratamiento, dos grandes retos en oncología.

Por lo tanto, en condiciones fenotípicas complejas como el cáncer, es preciso usar datos multi-ómicos como variables predictoras para obtener un modelo más comprensivo en estos casos [15]. A continuación se describen los tres principales tipos de datos ómicos utilizados en esta tesis: genoma, transcriptoma y proteoma.

- **Genoma**: Un genoma es el conjunto completo de DNA comprendido en una célula de un organismo. Es donde reside la información necesaria para las funciones del mismo. Contiene aproximadamente 3 billones de pares de bases, distribuidas en 23 pares de cromosomas dentro del núcleo de las células. Presenta alrededor de 20.000 genes codificantes de proteínas. La genómica es el campo de la genética que tiene como objetivo comprender el contenido, la organización, la función y la evolución de la información molecular del DNA.

- **Transcriptoma**: A diferencia de la genómica, que centra su atención en el DNA, la transcriptómica es el estudio del conjunto de moléculas de RNA que existen en una célula, tejido u órgano. Existen múltiples tipos de RNA (mRNA, miRNA, rRNA, tRNA, iRNA, etc.), la mayoría de ellos desempeñando un papel regulador dentro de la célula. El transcriptoma es altamente variable ya que muestra lo que se está expresando en un determinado momento. Aunque todas las células presenten la misma información genética, no todas las células expresan los mismos genes. Es esta expresión, la que aporta la diversidad entre células de un mismo organismo. Conocer el transcriptoma es muy útil en la clínica para observar los patrones de expresión genética en condiciones sanas y patológicas, como en el caso del cáncer.

- **Proteoma**: Las proteínas son las moléculas biológicas construidas mediante bloques denominados aminoácidos. Las proteínas son las encargadas de realizar las acciones en la célula (estructura, metabolismo, transporte, inmunidad, señalización o regulación, entre otros). El término de proteoma fue utilizado por primera vez por el Dr. Marc Wilkins en 1994 y se refiere al conjunto de proteínas que un organismo sintetiza a partir de los genes que contiene. Así como el transcriptoma, la expresión del proteoma está en un estado constante de flujo a lo largo del tiempo y a través del individuo. Podemos, por lo tanto, obtener información del proteoma mediante técnicas de secuenciación (secuenciando el DNA y convirtiéndolo a secuencia aminoacídica mediante el código genético) o cuantificando la expresión de proteínas de una muestra.

A continuación se verán las tecnologías y las estrategias utilizadas para generar estos tipos de datos.

## 1.2 Tecnologías de secuenciación masiva

La investigación oncológica ha crecido gracias al desarrollo de tecnologías de secuenciación masiva capaces de obtener datos multi-ómicos a gran escala, con bajo coste y en un tiempo reducido. Aún así, la caracterización de grandes cohortes de pacientes solo es asumible mediante grandes consorcios internacionales. Muchos de estos consorcios, adhiriéndose al concepto de *Open Science*, ponen a disposición de la comunidad científica los datos generados para posibilitar nuevos desarrollos y reproducibilidad de sus resultados.

La tecnología de secuenciación masiva, conocida también como *Next Generation Sequencing* (NGS, por sus siglas en inglés) ha revolucionado las ciencias biológicas. Gracias a su altísimo rendimiento, escalabilidad y velocidad, la NGS permite a los investigadores realizar una gran variedad de aplicaciones y estudiar sistemas biológicos a un nivel nunca antes posible.

Se considera a la técnica de Sanger como el método de primera generación [21]. Fue esta técnica la que impulsó a los integrantes del Human Genome Organization (HUGO) a desarrollar el Proyecto Genoma Humano [22]. Posteriormente, emergieron técnicas de secuenciación más económicas, rápidas y de alto rendimiento. Empresas como Roche, Ion torrent, Illumina (Solexa) y/o SOLiD desarrollaron y mejoraron diversas plataformas para la secuenciación del DNA.

Las aplicaciones de estas tecnologías presentan un importante avance en el campo de la biomedicina. Las tecnologías NGS se especializaron en la generación de diferentes tipos de datos, permitiendo a los investigadores centrarse en cuestiones relacionadas con el genoma, el transcriptoma o el epigenoma, entre otros.

Son muchas las tecnologías NGS que generan diferentes datos ómicos. Esta sección no pretende profundizar en cada una de ellas. Por lo tanto, se describirán, de manera general, las tecnologías NGS necesarias para obtener datos **genómicos**, **transcriptómicos** y **proteómicos**.

### 1.2.1 Obtención de datos genómicos

Los datos genómicos se obtienen mediante la secuenciación de cada uno de los nucleótidos que conforman el DNA de una célula. Debido a su dimensionalidad, existen diversas estrategias para extraer información del genoma, que se diferencias según las regiones del genoma secuenciadas. Las diferentes técnicas moleculares utilizadas dependen de la plataforma de secuenciación. Dos términos muy importantes en cuanto a la secuenciación del DNA son: la *cobertura* y la *profundidad*. La cobertura es el número de lecturas únicas que incluyen un nucleótido. La profundidad es la cantidad del número de lecturas de cada región de la secuencia. A continuación se detallan las diferentes estrategias de secuenciación.

- **Secuenciación de panel de genes** El objetivo es la secuenciación de un número específico de genes que estén relacionados con algún tipo de condición. La limitación de esta prueba es intrínseca al diseño del panel de genes, por lo que no se podrán observar modificaciones genéticas en regiones que no estén predefinidas por el panel de genes utilizado. La mayoría de las aplicaciones se centran en el diagnóstico de enfermedades donde se hayan definido previamente genes candidatos.

- **Secuenciación completa de exoma** En este caso, se secuencian todas las regiones codificantes del genoma, es decir, su exoma, lo que representa entre el uno y el dos por ciento de la secuencia genómica completa. Es una prueba muy útil en clínica, ya que es donde se encuentran la mayoría de las variantes conocidas asociadas a enfermedad. La limitación que presenta es que no se obtiene información acerca de las modificaciones genéticas que ocurren en las regiones no codificantes del genoma. Es utilizada principalmente en el diagnóstico de enfermedades heterogéneas, donde no se conocen los genes involucrados en su patogenicidad.

- **Secuenciación completa de genoma** Al contrario que las otras dos pruebas anteriores, esta prueba secuencia todas las bases del genoma de un individuo. Debido a su gran cobertura, es una prueba bastante costosa, pero a diferencia de las otras, ofrece información de todas las posibles variantes del genoma. La principal limitación es la interpretación de las variantes encontradas, ya que es posible que gran parte de las variantes no estén anotadas en bases de datos públicas, utilizadas como referencia, sobre todo en regiones no codificantes del genoma.

### 1.2.2 Obtención de datos transcriptómicos

A diferencia de los datos genómicos, los transcriptómicos son datos dinámicos. La medición de estos va a depender altamente tanto de las condiciones internas como externas del individuo. Por ello, son muy utilizados para el estudio de afecciones moleculares complejas. A lo largo de la historia se fueron desarrollando diferentes técnicas biotecnológicas capaces de ofrecer información del transcriptoma. En esta sección se comentarán las dos técnicas que han marcado el curso de la investigación y que son utilizadas a día de hoy.

- **Microarrays** La idea subyacente a esta técnica es la cuantificación de moléculas de mRNA de un tejido y/o célula en un determinado momento temporal. Ser capaces de cuantificar las moléculas de mRNA nos permite inferir el grado de expresión de un determinado gen. Para ello, el RNA de una muestra es convertido a cDNA y éste es marcado con fluorescencia ó radioactividad. De forma paralela, en múltiples pocillos, se añaden fragmentos de DNA previamente conocidos de las secuencias génicas que se quiere analizar (una por pocillo). Tras la mezcla de cDNA y DNA, a mayor hibridación, mayor señal lumínica/radioactiva. La intensidad de la señal es convertido posteriormente a una escala de la intensidad de expresión del gen.

  Esta técnica se denomina microarrays, porque la hibridación es llevada en placas que contienen numerosos micropocillos. En cada micropocillos se introduce un fragmento de DNA conocido que para que hibride con una región concreta del genoma.

  Durante décadas fue la técnica utilizada para el análisis del transcriptoma. Con la llegada de las técnicas de secuenciación masiva, se derribaron las principales limitaciones de las microarrays, como es la necesidad de conocimiento previo para la construcción de librerías, la hibridación cruzada y la reproducibilidad de resultados, debido principalmente a la conversión de la intensidad lumínica en conteos de moléculas de mRNA.

- **RNA-Seq** La técnica RNA-Seq fue introducida formalmente en el 2008, después de la publicación de dos artículos en la revista *Nature Methods* [23, 24], donde se describieron dos aplicaciones para la caracterización de mRNA en varios ratones con una profundidad y resolución nunca antes conseguida. Por una parte, Sean Grimmond y sus colaboradores utilizaron la plataforma SOLiD de

Apply Biosystems, mientras que Barbara Wold y colaboradores aplicaron la plataforma Solexa de Illumina GA. Ambas plataformas son conceptualmente parecidas, además de ofrecer longitudes de lectura y ratios de error similares.

Esta técnica reemplazó casi por completo el uso de microarrays. Se basa en técnicas de NGS para la cuantificación de mRNA. En primer lugar se extrae el RNA de las muestras a analizar. En este caso, la construcción de la librería se consigue tras la fragmentación del DNA, su conversión a cDNA y su secuenciación. La técnica utilizada en este paso es dependiente de la plataforma de secuenciación. Cuando la librería está construida, se alinea a un genoma de referencia y se identifica el número de secuencias que se alinean a cada región del genoma. De esta forma la conversión es directa, entre el conteo de moléculas de mRNA y la expresión del gen.

### 1.2.3 Obtención de datos proteómicos

Los datos proteómicos pueden ser dinámicos o estáticos. Los estáticos son aquellos que provienen a partir de la secuencia del gen. Gracias al código genético, si conocemos la secuencia exacta de una región codificante somos capaces de inferir la secuencia proteica. Por otra parte, los datos dinámicos hacen referencia a la expresión de las proteínas a un determinado tiempo de muestreo. Estas técnicas cuantifican el número de proteínas. Existen diferentes técnicas cuantitativas para la obtención de datos de expresión proteicas. En cuanto a los métodos de alto rendimiento, podemos diferenciar dos de ellos:

- **Microarrays de fase inversa** Los microarrays de proteínas aplican pequeñas cantidades de muestra a un *chip* para su análisis (a veces en forma de portaobjetos de vidrio con una superficie modificada químicamente). Los anticuerpos específicos pueden inmovilizarse en la superficie del chip y utilizarse para capturar las proteínas objetivo en una muestra compleja. Esto se denomina microarray analítico de proteínas, y estos tipos de microarray se utilizan para medir los niveles de expresión y las afinidades de unión de las proteínas en una muestra. Los microarrays de proteínas funcionales se utilizan para caracterizar las funciones de las proteínas, como las interacciones proteína-RNA y el recambio de enzimas por sustratos. En un microarray de proteínas de fase inversa, las proteínas de, por ejemplo, tejidos sanos y enfermos o células no tratadas y tratadas se unen al *chip*, y éste se sondea con anticuerpos contra las proteínas objetivo.

- **Espectrometría de masas** Existen varios métodos para separar las proteínas. Estos enfoques permiten tanto la cuantificación como la proteómica comparativa/diferencial. También existen otras técnicas menos cuantitativas que ofrecen las ventajas de ser más rápidas y sencillas. Otras técnicas cromatográficas para la separación de proteínas son la cromatografía de gases (GC) y la cromatografía de líquidos (LC).

## 1.3 The Cancer Genome Atlas

Como se ha comentado anteriormente, el estudio del cáncer se ha convertido a día de hoy en un estudio sistémico que alberga numerosas fuentes de datos heterogéneas. Para la caracterización precisa de los diferentes tumores, es necesario un estudio que integre los diferentes tipos de datos ómicos.

En el año 2006, el NCI y el NHGRI crearon el proyecto The Cancer Genome Atlas (TCGA), con el objetivo de obtener mapas genómicos multidimensionales completos de todos los cambios clave en varios tipos y subtipos de tumores [25].

En primer lugar se centraron en la caracterización de tres tipos de tumores humanos: el glioblastoma multiforme (GBM) [26], el carcinoma de células escamosas de pulmón (LUSC) [27] y el cáncer de ovario (OV) [28].

Posteriormente, el TCGA ha recogido tejidos de más de 11.000 individuos, un esfuerzo que permitió estudiar más de 33 tipos y subtipos de tumores, incluidos 10 cánceres raros. Lo más interesante de esta iniciativa es que toda la información es gratuita y accesible para cualquier investigador que quiera centrar sus esfuerzos en la enfermedad. Los datos se ofrecen en acceso abierto a la comunidad, un factor que facilita la generación de nuevos análisis y aproximaciones sin necesidad de una inversión económica inicial para obtener los datos.

Además, la caracterización se realizó a diferentes niveles ómicos, generando datos genómicos, transcriptómicos, epigenómicos y proteómicos. Todos estos datos están acompañados de rigurosas anotaciones clínicas de cada una de las muestras. En la Figura 1.3 se muestran los diferentes tipos de datos ómicos que alberga el repositorio del TCGA. Se detalla también el porcentaje según el tipo de cohorte. Para una mayor profundidad de las estadísticas que presenta el repositorio, se recomiendan los siguientes portales de acceso a los datos [25, 29].

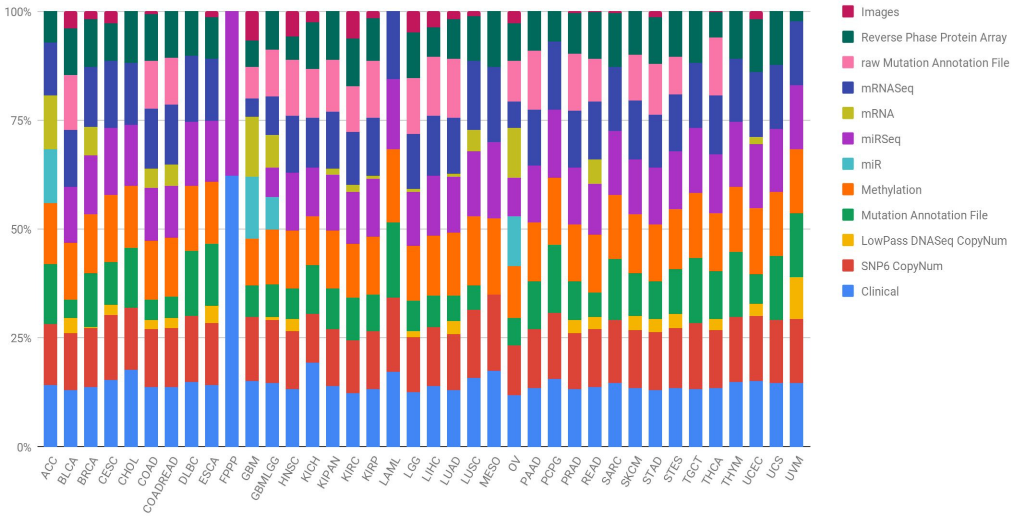


**Fig. 1.3.:** Proporción del número de muestras según tipo de dato y tipo de tumor en el repositorio TCGA. Imagen extraída de [2].

Actualmente el TCGA alberga en el repositorio de datos GDC Data Portal [29] un total de 33 cohortes de diferentes tejidos y/o tumores. Hasta la fecha, el TCGA ha caracterizado y publicado 33 tipos diferentes de tumores en las principales revistas internacionales. En [30] se puede encontrar un *snapshot* publicado en la revista Cell donde se resumen los principales hallazgos realizados por el consorcio TCGA en cada uno de los tumores analizados.

La generación de tal cantidad de datos, y su disponibilidad pública, generó un contexto perfecto para la aplicación de diferentes aproximaciones computacionales para el análisis de este tipo de datos. Entre ellos, las técnicas de Machine Learning (descritas en la sección 1.4.1) han sido utilizadas en gran medida sobre estos tipos de datos. Una de las publicaciones que compone esta tesis [1] (Capítulo 3.1) presenta un profundo análisis de los trabajos y resultados reportados por diferentes laboratorios a partir del uso de técnicas de Machine Learning con los datos del TCGA.

## 1.4 Técnicas computacionales utilizadas en la tesis

En este apartado se hará una descripción de las técnicas computacionales utilizadas para el desarrollo de esta tesis. La sección se ha dividido en cuatro subsecciones: Machine Learning, técnicas de selección de características, modelos QSAR y técnicas de *docking* molecular para el reposicionamiento de fármacos.

### 1.4.1 Machine Learning

De manera general, las técnicas de ML pueden definirse como un subconjunto de tareas asociadas con la inteligencia artificial, las cuales, a través de algún tipo de algoritmo o modelo extraen información relevante de un conjunto de datos.

En biomedicina se utilizan principalmente dos tipos de aprendizaje dependiendo del problema a resolver. El **aprendizaje no supervisado** se utiliza principalmente para la estratificación de pacientes en subgrupos. El **aprendizaje supervisado** se utiliza para la predicción de condiciones biológicas, tanto de muestras de pacientes como de moléculas.

Los modelos predictivos se obtienen a partir del aprendizaje de un conjunto de ejemplos, y se esperan que sean generalizables a toda la población. En biomedicina, las observaciones se refieren a las muestras de los pacientes, que estarán definidas por diferentes variables ómicas. La heterogeneidad tanto de plataformas de secuenciación como de las metodologías empleadas, y el gran coste económico que supone, convierte la tarea de predicción en un gran reto dentro del campo de la biomedicina.

Gracias a grandes proyectos multi-ómicos como el TCGA es posible el uso de algoritmos de ML en grandes cohortes de pacientes. Esta tesis tiene como hilo argumental la aplicación de algoritmos de ML sobre datos biomédicos. Para ello se hace uso tanto de cohortes de pacientes como la del TCGA, como de otros datos públicos necesarios. Dentro del compendio de artículos de la tesis, en [1] (ver sección 3.1) se presenta una revisión de los trabajos que han utilizado modelos de aprendizaje supervisado y no supervisado con los datos del TCGA. Las limitaciones encontradas en otras metodologías, tanto en la reproducibilidad como en la validación, sirvió de punto de partida para aplicar una metodología robusta en diferentes problemas biológicos. En [2] (ver sección 3.2) y en [3] (ver sección 3.3) se aplica una metodología de ML siguiendo las directrices recogidas en [1], para identificar biomarcadores y/o *pathways* alterados en cáncer de colon y para desarrollar un modelo de *screening* automático de fármacos anti-cancerígenos, respectivamente.

A continuación se describen los dos tipos de aprendizaje más utilizados en el campo de la biomedicina. Se prestará mayor atención al aprendizaje supervisado, ya que fue utilizado en [2] y en [3]. Posteriormente, se describen los modelos y la metodología utilizada para alcanzar los objetivos propuestos.

#### 1.4.1.1 Visión general del aprendizaje no supervisado

En los problemas de aprendizaje no supervisado los algoritmos reciben una serie de entradas sin presentar ningún tipo de etiqueta. En este contexto, los algoritmos agrupan los datos de entrada basándose en similaridades [31].

Un ejemplo clásico de aprendizaje no supervisado son los métodos de *clustering*, ampliamente utilizados en biomedicina para la estratificación de pacientes en subgrupos [32]. El objetivo principal es agrupar las observaciones en *clusters* de tal forma que aquellas observaciones dentro de un mismo *cluster* están más cercanas que las pertenecientes a otros *clusters*.

Además, a veces el objetivo es ordenar los *clusters* en una jerarquía natural [33]. Para ello, se agrupan sucesivamente los *clusters* de manera que, en cada nivel de la jerarquía, los *clusters* del mismo grupo sean más similares entre sí que los de grupos diferentes. El análisis de *clusters* también se utiliza para formar estadísticas descriptivas que permitan determinar si los datos están formados por un conjunto de subgrupos distintos, cada uno de los cuales representa observaciones con propiedades sustancialmente diferentes. Este último objetivo requiere una evaluación del grado de diferencia entre las observaciones asignados a los respectivos *clusters*. La noción de grado de similitud (o disimilitud) entre las observaciones individuales agrupados es fundamental para todos los objetivos del análisis de *clusters*.

#### 1.4.1.2 Visión general del aprendizaje supervisado

Todo problema de aprendizaje supervisado presenta dos tipos de variables en común: las variables de entrada y las variables de salida. Las entradas, también llamadas características, predictores y/o variables independientes, son medidas y tienen algún tipo de influencia sobre una o varias variables de salida. El objetivo, por lo tanto, es para cada ejemplo, predecir los valores de salida utilizando los valores de entrada [31].

Por su parte, también existen dos tipos de variables de salida. Puede ser una medida cuantitativa, es decir, algunas medidas son mayores que otras, o puede ser una medida cualitativa, asumiendo un valor en un conjunto finito.

Según la distinción entre la variable de salida, nos encontramos con dos tipos de tareas de predicción: **regresión** cuando se predicen variables cuantitativas, y **clasificación** cuando se predicen variables cualitativas.

Se necesita un conjunto de datos para construir las reglas de las predicciones, denominado conjunto de entrenamiento. Dichas reglas son intrínsecas a los algoritmos que se utilizan para realizar las predicciones.

En las siguientes secciones se presentan los modelos y la metodología de aprendizaje supervisado utilizados para el desarrollo de la presente tesis.

#### 1.4.1.3 Modelos de Machine Learning para aprendizaje supervisado

En el proceso de diseño de un modelo de ML, no se conoce la arquitectura óptima para la resolución del problema. Por lo tanto, será necesario explorar una gama de posibilidades. Se denomina arquitectura al conjunto de hiperparámetros que conforman un modelo de ML. La búsqueda de los valores óptimos se denomina, por lo tanto, ajuste de hiperparámetros (ó *hyperparameter tuning* en inglés).

Cada algoritmo de ML tiene sus propios hiperparámetros. Es importante diferenciar hiperparámetros de parámetros. Los parámetros son aquellos que son entrenados directamente a partir de los datos, y el propio modelo optimiza mediante una función de rendimiento. Mientras que los parámetros del modelo especifican cómo transformar los datos de entrada en la salida deseada, los hiperparámetros definen cómo se estructura realmente el modelo de ML.

El ajuste de hiperparámetros comienza con la definición de un espacio de búsqueda. Será un espacio de búsqueda finito, delimitado por rangos de valores definidos de cada hiperparámetro. La selección de las estrategias se basará principalmente en el tamaño del espacio de búsqueda y la capacidad computacional disponible. Existen tres tipos de búsqueda: *grid*, aleatoria y heurística. A continuación se realiza una explicación de cada una de ellas. La Figura 1.4 muestra la diferencia entre los tres tipos de estrategias.

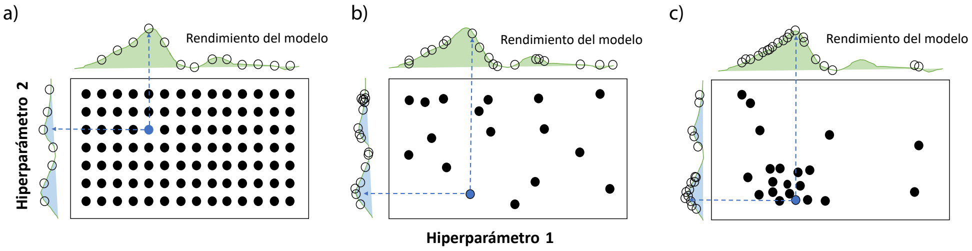


Fig. 1.4.: Técnicas de búsqueda para el ajuste de hiperparámetros. a) búsqueda *grid*; b) búsqueda aleatoria; c) búsqueda heurística. Los diagramas de densidad representados en los ejes de abscisas corresponden al rendimiento del modelo. En a) y b) se observa una distribución aleatoria. En c) la búsqueda va enfocada hacía la maximización del rendimiento. Figura adaptada de [34].

- **Búsqueda *grid***: esta estrategia (ver Figura 1.4a) implica especificar una lista de posibles valores de hiperparámetros para que el algoritmo sea entrenado con cada combinación posible. Se elegirá la mejor combinación mediante una métrica de rendimiento especificada.

- **Búsqueda aleatoria**: En este caso, en lugar de dar una lista de valores de hiperparámetros para que el algoritmo de optimización los pruebe, se proporcionan distribuciones estadísticas de valores. De esta forma, se muestrea aleatoriamente a partir de las distribuciones definidas para luego probarlas generando un modelo (ver Figura 1.4b). Al igual que la anterior estrategia, esta utiliza una métrica de rendimiento para seleccionar la mejor combinación de valores. La búsqueda aleatoria se utiliza en situaciones donde el espacio de búsqueda es muy grande.

- **Búsqueda heurística**: A diferencia de las dos estrategias anteriores, este tipo de estrategia utiliza información de los experimentos anteriores para mejorar los siguientes experimentos. Se engloba en esta categoría métodos como el descenso del gradiente, optimización bayesiana o los algoritmos genéticos. Cómo se observa en la Figura 1.4c, la búsqueda se realiza de forma guiada, hasta converger en un punto óptimo. Estas estrategias, con un gran gasto computacional, son utilizadas en situaciones donde el espacio de búsqueda es muy grande. A medida que el espacio es más grande, estos algoritmos pueden converger hacia mínimos locales, sin ser capaces de encontrar la solución óptima.

El ajuste de hiperparámetros de cada algoritmo es crucial para poder obtener modelos que generalicen y sean capaces de evitar el sobreentrenamiento. A continuación se describen brevemente los fundamentos de aquellos modelos que mejores resultados han obtenido en esta tesis, sus características principales y cómo ha sido el ajuste para cada uno de ellos.

#### 1.4.1.3.1 Modelos lineales

Se puede considerar que un modelo lineal es el algoritmo de ML más sencillo que existe. Relaciona las variables de entrada con la variable de salida mediante la inferencia de valores $\beta$ que multiplican a cada una de las variables del sistema. La ecuación se define como:

$$f(y) = \beta_0 + \beta_1 x_1 i + ... + \beta_k x_k, i \tag{1.1}$$

Este modelo reporta un valor continuo entre $(\infty, -\infty)$. Para problemas de clasificación, este rango de respuesta es acotado entre 0 y 1 mediante la función *logit*, convirtiendo el modelo en una regresión logística [35]. Se establece un umbral para la clasificación de las observaciones. Suponiendo una clasificación binaria, con el umbral definido en 0.5, las predicciones mayores a 0.5 serán categorizadas en una clase, mientras que las inferiores a 0.5 pertenecerán a la otra. Este umbral puede ser modificado para ajustar las predicciones.

En problemas de alta dimensionalidad (siendo $p >>> n$), como es el caso del análisis de datos ómicos, no existe una solución única, habiendo realmente un conjunto infinito de soluciones en estos sistemas. Por lo tanto, es imposible hallar la solución correcta sin alguna información adicional o restricción que acote el número de variables en el modelo. El supuesto de dispersión considera que solo un reducido número de variables tiene valor distinto de cero en los coeficientes de regresión óptimos.

Por lo tanto, un modelo estadístico disperso es aquel en el que solo un número relativamente pequeño de variables juegan un papel importante. Las técnicas de regularización (o *shrinkage*) [31] son las que permiten la obtención de modelos que ofrezcan un equilibrio entre simplicidad y rendimiento del modelo, realizando estimaciones sesgadas de los coeficientes. Estos estimadores consideran todas las variables predictoras pero fuerzan a que algunos coeficientes estén muy próximos a cero, o directamente establecerlos a cero.

El estimador *Ridge* constituye la primera aplicación de las técnicas de regularización mediante aplicación de penalización [36]. Posteriormente, se desarrolla el método *LASSO* [37] que presenta, a diferencia del anterior, la capacidad de selección de variables. Posteriormente se desarrolló *elastic net*, que combina ambos métodos.

En esta tesis se aplica esta última técnica mediante el algoritmo *glmnet*. *Glmnet* es un algoritmo que ajusta modelos lineales generalizados mediante *penalized maximum likelihood* [38].

Para la regresión logística, la función objetivo de *glmnet* se define como:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -[\frac{1}{N}\sum_{i=1}^{N} y_i * (\beta_0 + x_i^T\beta) - \log(1+e^{(\beta_0+x_i^T\beta)})] + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1] \tag{1.2}$$

*Glmnet* presenta dos parámetros importantes para ser ajustados. Por una parte, el parámetro $\lambda$ regula la fuerza del *shrinkage*. Por otra parte, el parámetro $\alpha$ controla el tipo de *shrinkage*. $\alpha = 1$ se corresponde con la técnica *Lasso*, la cual tiene una penalización $l_1$ sobre los parámetros y realiza tanto la contracción como la selección de variables. $\alpha = 0$ corresponde con la técnica *ridge regression*, con una penalización $l_2$ sobre los parámetros y no presenta selección de variables.

En problemas de biomedicina es importante establecer un *score* que indique la importancia de cada variable en el modelo, para así tener una explicabilidad biológica de los resultados. Por ejemplo, cuando se analizan datos genómicos, es importante inferir cuál de los genes es más importante en las predicciones que hace el modelo. En este tipo de modelos lineales, se puede inferir esta importancia a partir de los valores de los coeficientes. Como se verá más adelante, el entrenamiento y la validación de los modelos se realiza mediante procesos de *cross-validation*. *Grosso modo*, estas técnicas generan $n$ modelos con configuraciones diferentes. Una aproximación para inferir la importancia de cada variable es la suma de los coeficientes asignados a cada variable en todos los modelos generados. Por lo tanto, a mayor valor absoluto del coeficiente, mayor peso tendrá la variable en el modelo. Además, debido a las técnicas *shrinkage*, una variable con un coeficiente alto, se entiende que presenta una gran importancia en el modelo.

#### 1.4.1.3.2 Modelos basados en distancias

Ciertos modelos basan sus predicciones en el cálculo de distancias. Para ello, sitúan las observaciones del conjunto de entrenamiento en un espacio multi-dimensional, definido por las variables disponibles. Cuando se posiciona una nueva muestra en dicho espacio, la distancia entre las demás observaciones será lo que dicte la predicción del modelo. Este tipo de aprendizaje se denomina *lazy learning*, debido a que el dataset es almacenado en memoria, y se usa para realizar las predicciones sobre el nuevo conjunto de datos.

Existen múltiples tipos de distancias. A continuación se formulan las más utilizadas:

- **Distancia Euclídea**:

$$D_E = \sqrt{\sum_{i}^{N}(x_{1i} - x_{2i})^2} \tag{1.3}$$

- **Distancia de Manhattan**:

$$D_{Ma} = \sum_{i=1}^{n} |x_{1i} - x_{2i}| \tag{1.4}$$

- **Distancia de Minkowski**:

$$D_{Mi} = (\sum_{i=1}^{n} |x_{1i} - x_{2i}|^p)^{\frac{1}{p}} \tag{1.5}$$

El algoritmo k-NN es uno de los algoritmos más conocidos y utilizados en problemas de clasificación [39]. Debido a su simplicidad y rapidez, ha sido ampliamente utilizado en problemas de biomedicina.

Para la predicción de una nueva observación se basa en la etiqueta de los vecinos cercanos (*nearest neighbors*, en inglés). De esta forma, calcula la distancia entre el punto nuevo y sus vecinos. Los vecinos con distancias más cortas se seleccionan. El nuevo punto es asignado a la clase con mayor número de vecinos cercanos (ver Figura 1.5).

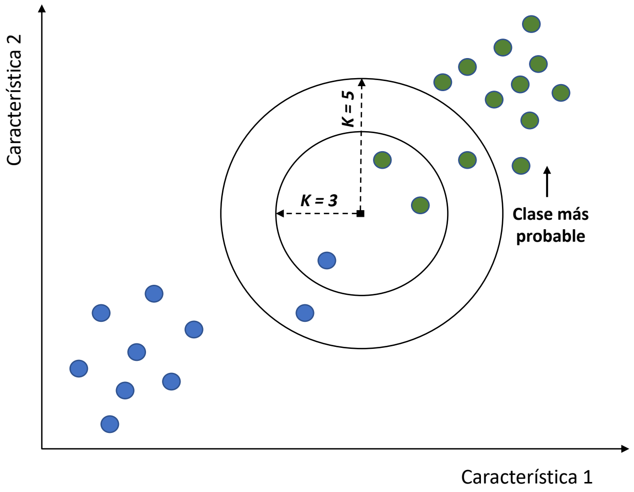


Fig. 1.5.: Representación del algoritmo k-NN. Los círculos muestran el espacio vectorial comprendido según el valor definido por $k$. Los colores representan la clase perteneciente de cada muestra. El cuadrado negro indica la nueva muestra a predecir.

De formas más específica, k-NN usa aquellas observaciones en el conjunto de entrenamiento $T$ cercanas en el espacio de entrada a $X$ para formar $\hat{Y}$. En concreto, el ajuste de k-NN para $\hat{Y}$ se define como sigue:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{1.6}$$

donde $N_k(x)$ es el vecino de $x$ definido por el punto *k* más cercano $x_i$ en la muestra de entrenamiento.

En los resultados que presenta esta tesis, se ajusta únicamente el valor de $k$. La distancia que se utiliza es la de Minkowski, considerándose una medida más robusta, al ser una combinación de las otras dos mencionadas. Además, se utiliza la variante *weighted* de k-NN, la cual no solo se fija en la distancia, si no que establece un peso según el valor de la distancia. Es decir, si una observación está particularmente cerca se le asigna un peso mayor que una que esté relativamente lejos.

#### 1.4.1.3.3 Modelos basados en árboles

Ciertos modelos están formados por reglas logísticas, generando una estructura de árbol, y dividiendo las observaciones del conjunto de entrenamiento en función de sus variables hasta predecir el valor de la variable de salida.

Uno de los algoritmos más conocidos son los árboles de decisión. Un árbol de decisión consta de tres componentes como muestra la Figura 1.6: los nodos de decisión, los nodos hoja y un nodo raíz. El algoritmo divide un conjunto de datos de entrenamiento en ramas, que a su vez se segregan en otras ramas. Esta secuencia continúa hasta que se alcanza un nodo hoja. El nodo hoja no puede segregarse más. La división de las observaciones se realiza en los nodos, que albergan una regla logística en función de las variables.

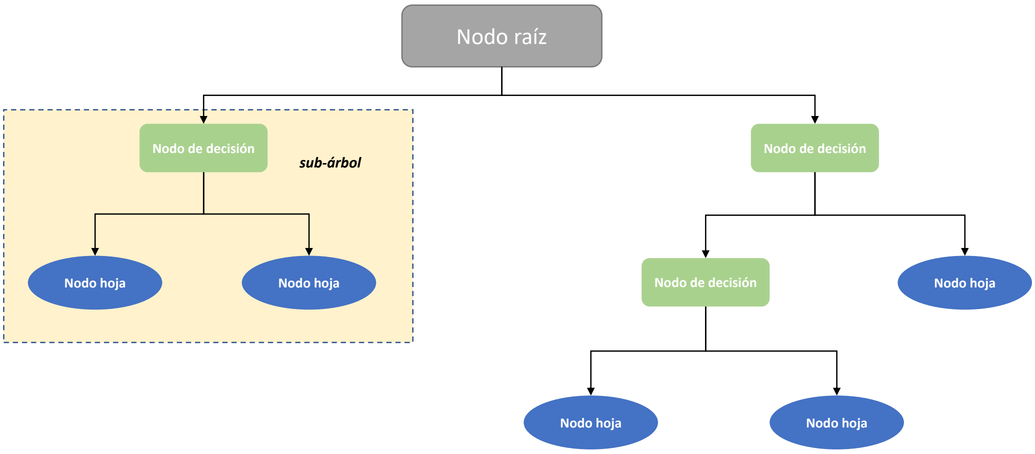


Fig. 1.6.: Representación de un árbol de decisión. El árbol se compone de tres componentes: nodo raíz, nodo de decisión y nodo hoja. Las muestras se van subdividiendo a lo largo del árbol según las reglas de división ajustadas. El nodo hoja es el último nodo dentro un árbol y comprende las muestras pertenecientes a una determinada clase.

Estos algoritmos presentan grandes ventajas. Por ejemplo, son altamente interpretables, no se ven influenciados por *outliers*, son capaces de seleccionar variables de forma automática y manejan tanto variables categóricas como numéricas.

Una de sus desventajas es que el rendimiento en la predicción de un único árbol normalmente es bajo y tienen una tendencia al sobreentrenamiento.

Para paliar estas desventajas, se desarrollaron métodos más robustos como Random Forest (RF) [40]. Este algoritmo utiliza una estrategia de *ensemble learning*, es decir, combina muchos clasificadores para ofrecer una solución a un problema complejo. Los clasificadores son árboles de decisión. La predicción la realiza tomando la media de los resultados de todos los árboles. Por lo tanto, al aumentar el número de árboles se incrementa la precisión del resultado. La estructura y funcionamiento del algoritmo RF se representa en la Figura 1.7.

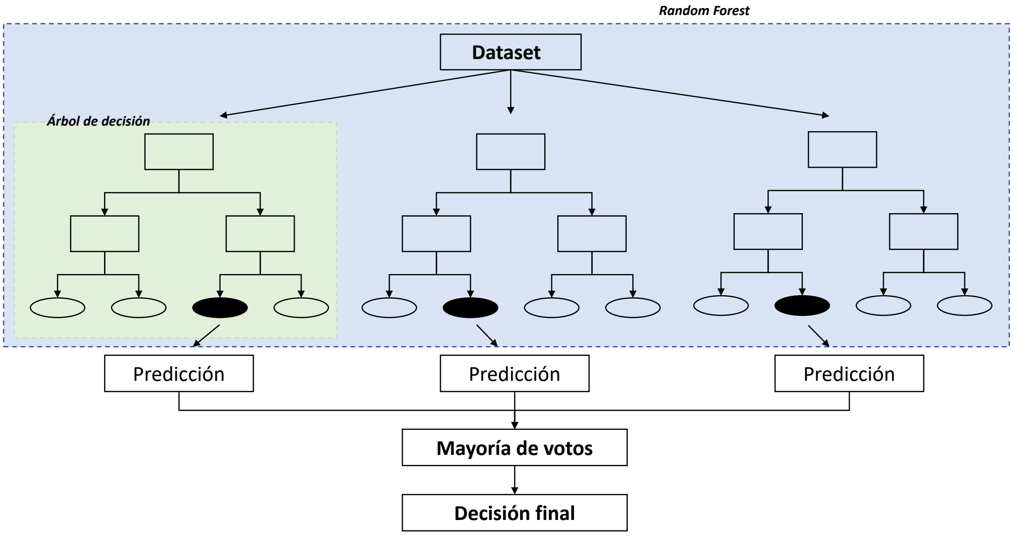


Fig. 1.7.: Representación del algoritmo Random Forest. La composición de cada árbol es la misma que en la Figura 1.6. El algoritmo toma un número de variables aleatorias del conjunto de datos inicial para construir cada árbol. El número de árboles es definido por el usuario. Cada muestra es clasificada al menos por un árbol. La predicción final se lleva a cabo tras una ponderación de votos.

Una característica importante en el algoritmo RF es que emplea un método denominado *bagging* para generar las predicciones. El concepto *bagging* implica el uso de diferentes muestras de datos (datos de entrenamiento) en lugar de una sola muestra. Un conjunto de datos de entrenamiento comprende observaciones y características que se utilizan para hacer predicciones. Cada uno de los árboles de decisión que forman un RF producen diferentes resultados, en función de los datos de entrenamiento introducidos en el algoritmo. El promedio más alto de predicciones es definido como resultado final.

Durante la fase de entrenamientos que se realiza en los experimentos de esta tesis, se ajustan tres de los hiperparámetros de RF: el número de variables escogidas aleatoriamente en cada división de los datos, el tamaño mínimo de los nodos hoja, y el número de árboles. Se conoce que un valor estándar en el número de variables escogidas en problemas de clasificación es la raíz cuadrada del número total de variables. Para la búsqueda del valor óptimo, normalmente se establece un rango donde dicho valor es la mediana. Por otro lado, el valor del número de muestras en los nodos hoja indica la profundidad del árbol. Hay que tener en cuenta que un valor bajo conlleva una mayor profundidad del árbol, mejorando el rendimiento del modelo. Finalmente, un elevado número de árboles asegura que cada observación se predice al menos varias veces, evitando el sobreentrenamiento.

A la hora de tener una interpretación del modelo, RF permite también calcular la importancia de cada variable. Una forma sencilla es simplemente contar el número de veces que cada variable es seleccionada por todos los árboles del conjunto. Aunque esta sería una aproximación válida, existen otras formas más elaboradas y exactas de analizar la importancia de las variables.

Otras medidas incorporan una media ponderada de la mejora de cada árbol de forma individual. Esta medida es calculada en el criterio de división producida por cada variable. Un ejemplo de ella es la denominada *Gini importance* que es la que se utiliza en este trabajo.

Breiman [40] propuso evaluar la importancia de una variable $X_m$ para predecir $Y$ sumando las disminuciones de impureza ponderadas $p(t)\Delta_i(s_t, t)$ para todos los nodos $t$ en los que se utiliza $X_m$, promediados en todos los árboles $N_T$ del bosque:

$$Imp(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t) = X_m} p(t)\Delta_i(s_t, t) \tag{1.7}$$

siendo $p(t)$ la proporción $N_t/N$ de muestras que llegan a $t$ y $v(st)$ es la variable utilizada en la división $s_t$.

Debido a que se realiza un proceso de *cross-validation*, la importancia de cada variable es resultado del sumatorio a lo largo de los $n$ modelos generados.

#### 1.4.1.3.4 Modelos basados en *kernels*

Un *kernel* básicamente, es una función que mapea un espacio de entrada en otro de una dimensionalidad mayor, dónde se espera que los datos puedan ser linealmente separables. En ML se utilizan conjuntos de puntos en un determinado espacio para aprender una manera de tratar nuevas observaciones. Los métodos basados en *kernel* utilizan esos puntos para inferir cómo son de similares las nuevas observaciones y tomar una decisión. Es decir, los *kernels* codifican y miden la semejanza entre objetos [41, 42].

Una de los modelos basados en *kernels* más conocido y ampliamente utilizado son las *Support Vector Machines* (SVM), introducidas por Vapnik [43]. La base del algoritmo SVM es la separación de las diferentes clases mediante un hiperplano definido por un número de vectores soporte. Se denomina hiperplano a un subespacio de dimensiones $n-1$ que divide el espacio en dos mitades que se corresponde con las entradas de las dos clases. Por lo tanto, el objetivo del método SVM es encontrar el hiperplano que separe correctamente ambas clases. Se muestra en la Figura 1.8 una ejemplificación de la construcción del hiperplano óptimo de separación.

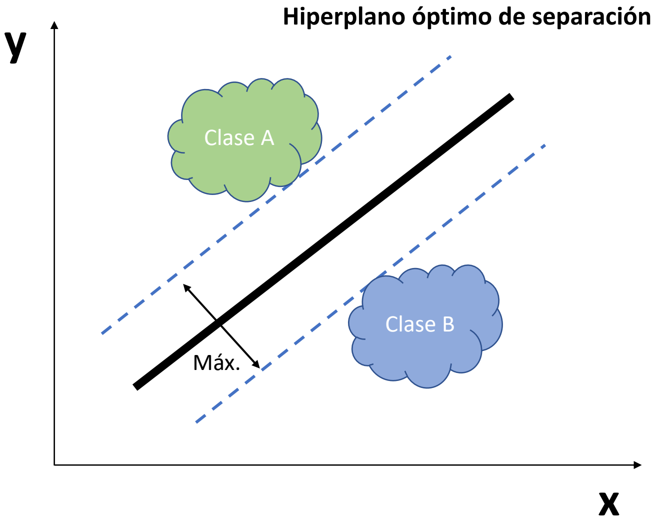


Fig. 1.8.: Representación del hiperplano óptimo de separación del algoritmo SVM. La línea negra continua representa el hiperplano de https://www.overleaf.com/project/60b4c8f84b816a3444c4b26bseparación. A cada lado, las líneas discontinuas representan los vectores de soporte. Estos vectores son calculados mediante distancias a los puntos en el espacio vectorial. Las nubes corresponden hipotéticamente al conjunto de muestras de cada clase.

La definición matemática de un hiperplano en dos dimensiones corresponde a la ecuación de una recta, pudiendo generalizarse para *p*-dimensiones.

En casos linealmente separables existen infinitos hiperplanos posibles. El objetivo es encontrar el hiperplano óptimo de separación, que se define como aquel que consigue un mayor margen, es decir, que la distancia mínima entre el hiperplano y las observaciones sea lo más grande posible. Debido a la infinidad de posibles hiperplanos, este problema se resuelve mediante métodos de optimización.

En casos no separables linealmente las técnicas de SVM utilizan lo que se denomina el *truco del kernel*. Esta idea es muy sencilla y trata de mapear el espacio de entrada inicial en un espacio de mayor dimensionalidad [44] (denominado espacio de características). Si el *kernel* cumple las condiciones de Mercer [45], éste representa un producto escalar en el nuevo espacio de características. La búsqueda del hiperplano, gracias a la serie de características que cumple, será directo y más eficiente que el cálculo y el reescalado en el espacio multidimensional.

Son muchos los *kernels* utilizados en la práctica. Una posible clasificación es separarlos en *kernels* locales y globales [46]. Los locales utilizan solamente aquellos datos que están cerca de otros, mientras que los globales tienen en cuenta todos los puntos en el conjunto de entrenamiento. Los *kernels* más conocidos se representan en la Figura 1.9.

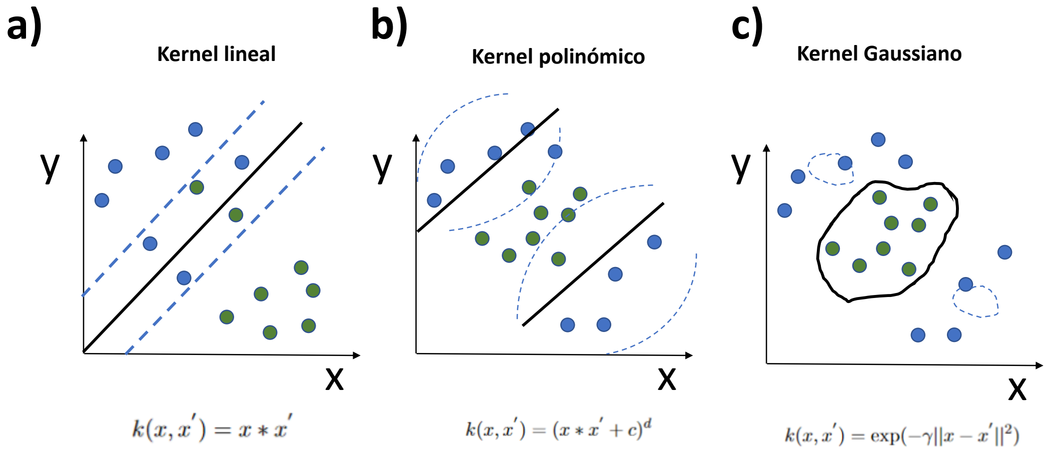

a)
Kernel lineal
y
x
$k(x, x^{'}) = x * x^{'}$
b)
Kernel polinómico
y
x
$k(x, x^{'}) = (x * x^{'} + c)^d$
c)
Kernel Gaussiano
y
x
$k(x, x^{'}) = \exp(-\gamma||x - x^{'}||^2)$

Fig. 1.9.: Representación de los diferentes tipos de kernels en un algoritmo SVM. a) kernel lineal; b) kernel polinómico; c) kernel Gaussiano.

Los *kernels* mostrados son solo unos pocos de los muchos que existen. Cada uno tiene una serie de hiperparámetros cuyo valor óptimo puede ajustarse. No puede decirse que exista uno que supere al resto, depende en gran medida de la naturaleza del problema que se esté tratando. Aunque sí es cierto que es recomendable siempre hacer una búsqueda a partir del *kernel* RBF, como indican los autores de [47]. Este kernel tiene dos ventajas: que solo tiene dos hiperparámetros que optimizar ($\gamma$ y la penalización C común a todos los SVM) y que su flexibilidad puede ir desde un clasificador lineal a uno muy complejo.

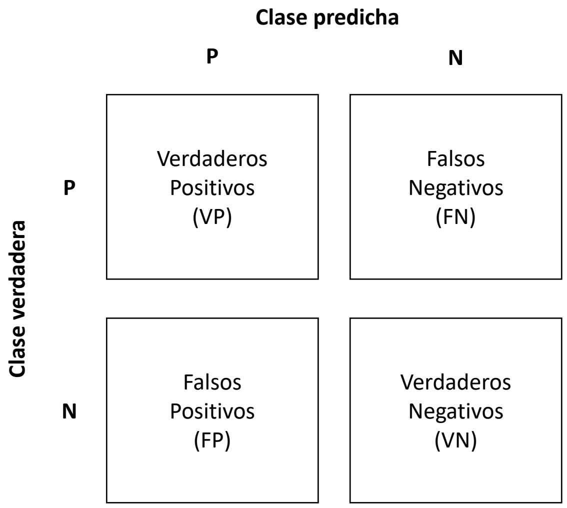
La modificación del valor C implica un ajuste de la penalización de la clasificación. En muchas ocasiones la búsqueda de hiperplano de separación se realiza asumiendo ciertos fallos en las predicciones. De esta manera, se asegura una mayor generalización del algoritmo, y se evita un sobreentrenamiento del mismo. El hiperparámetro C controla este efecto. Por otra parte, el hiperparámetro $\sigma$ representa la desviación estándar de la distribución Gaussiana. Con un valor alto de $\gamma$, el límite de decisión del SVM dependerá simplemente de los puntos más cercanos, ignorando los puntos más lejanos. En consecuencia, los valores altos de $\gamma$ suelen producir límites de decisión muy flexibles, y los valores bajos de $\gamma$ suelen dar lugar a un límite de decisión más lineal.

#### 1.4.1.4 Métricas de rendimiento

Cuando evaluamos el rendimiento de un algoritmo de clasificación debemos guiarnos por algún tipo de métrica que nos indique cuál ha sido el rendimiento del modelo. Existen muchas métricas en el literatura que se utilizan para evaluar el rendimiento de los algoritmos de ML.

La mayoría de estas métricas se calculan a partir de las matrices de confusión. Por ejemplo, un modelo de clasificación para problemas binarios identifica cada ejemplo con su clase de pertenencia. Dichos modelos obtienen vectores de salidas continuos en los que se puede trazar un umbral para diferenciar las clases de pertenencia. Existen, por lo tanto, cuatro tipos de respuestas del clasificador, considerando que sea binaria la variable de salida, como se observa en la Figura 1.10.

A continuación se describen algunas de las métricas utilizadas a lo largo de la presente tesis, como se describen en [48]. Se utilizará la siguiente notación. Dado un conjunto de datos de test, $m$ representa el número de ejemplos, y $c$ el número de clases. $f(i,j)$ representa la probabilidad de que el ejemplo $i$ pertenezca a la clase $j$. Asumimos que $f(i,j)$ siempre toma valores entre $0, 1$. Con $m_j = \sum_{i=1}^{m} f(i,j)$, denotamos el número



|  |  | Clase predicha |  |
|---|---|---|---|
|  |  | P | N |
| Clase verdadera | P | Verdaderos<br>Positivos<br>(VP) | Falsos<br>Negativos<br>(FN) |
|  | N | Falsos<br>Positivos<br>(FP) | Verdaderos<br>Negativos<br>(VN) |

Fig. 1.10.: Ejemplo de matriz de confusión para una variable de salida binaria.

de ejemplos de la clase $j$. Por su parte, $p(j)$ denota la probabilidad a priori de la clase $j$.

Dado un clasificador, $p(i,j)$ representa la probabilidad estimada del ejemplo $i$ para ser de la clase $j$ tomando valores en $[0,1]$. $C_\theta(i,j)$ es 1 si $j$ es la clase predicha para $i$ obtenida de $p(i,j)$ utilizando un umbral $\theta$. De otra manera, $C_\theta(i,j)$ es 0.

- ***Accuracy***: esta es la medida más común y simple para evaluar un clasificador. Se define como el grado de predicciones correctas realizadas por el modelo.

$$Acc = \frac{VP+VN}{VP+VN+FP+FN} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{c} f(i,j)C(i,j)}{m} \tag{1.8}$$

- ***Recall***: probabilidad de que se obtenga un resultados positivo para un caso positivo.

$$Recall = \frac{VP}{VP+FN} = \sum_{i=1}^{m} \frac{f(i,j)C(i,j)}{m_j} \tag{1.9}$$

- ***Precision***: proporción de positivos clasificados correctamente.

$$Precision = \frac{VP}{VP + FP} = \frac{\sum_{i=1}^{m} f(i,j)C(i,j)}{\sum_{i=1}^{m_j} C(i,j)} \tag{1.10}$$

- ***F Measure***: medida combinada entre la *precision* y el *recall*, definida por la siguiente fórmula.

$$F = \frac{2.recall.precision}{recall + precision} \tag{1.11}$$

- ***AUC***: El AUC (*Area Under the ROC Curve*) [49] de un clasificador binario es equivalente a la probabilidad de que el clasificador clasificará mejor una instancia positiva elegida al azar que una instancia negativa elegida al azar.

$$AUC = \frac{\sum_{i=1}^{m} f(i,j) \sum_{t=1}^{m} f(t,k) I(p(i,j), p(t,j))}{m_j.m_k} \tag{1.12}$$

  Donde $I(.)$ es una función de comparación que satisface $I(a,b) = 1$ iff $a > b$, $I(a,b) = 0$ iff $a < b$ y $I(a,b) = 0{,}5$ iff $a = b$

Entre estas medidas, la métrica *AUC* es la que se tomó como referencia en la mayoría de los resultados obtenidos a lo largo de esta tesis, como se verá en el capítulo 3. En [48] reportan, tras un análisis comparativo de varias métricas, que el AUC es la medida más segura a la hora de evaluar un clasificador. Además, la curva ROC también es una forma muy útil para visualizar el rendimiento de un clasificador.

#### 1.4.1.5 Preprocesado de los datos

Para todo análisis basado en la aplicación de modelos de ML, es crucial un paso previo de preprocesado de los datos. Este proceso involucra una serie de acciones que se realizan para introducir en el modelo los datos de la mejor manera posible.

Generalmente, es común encontrarse con valores no disponibles. Por ejemplo, al trabajar con datos ómicos son muchas las ocasiones donde valores de una variable no están presentes, ello es debido al tipo de secuenciación que se realiza. Por otra parte, un hecho también bastante común es la presencia de *outliers* o valores atípicos. Estos valores deben de tratarse adecuadamente y decidir si se añaden en el modelo. Técnicas tales como Principal Component Analysis (PCA) [50], t-SNE [51], o incluso configuraciones de RF [52] son capaces de identificar estos valores.

En conjuntos de datos ómicos, es común la presencia de variables constantes a lo largo de todo el conjunto de datos. En estos casos, estas variables deben de ser eliminadas ya que no aportan ninguna información al modelo. Por otra parte, existen variables que están muy correlacionadas entre ellas. Por ejemplo, la expresión de ciertos genes va a estar altamente correlacionada con la expresión de otros. En este caso, los métodos de selección de características juegan un papel muy importante, al ser capaces de seleccionar variables importantes dentro del conjunto de datos.

Otro aspecto importante es el balanceo de las muestras. Se dice que un conjunto de datos está balanceado cuando cada clase de salida está representada aproximadamente por el mismo número de observaciones. Cuando esto no se cumple, los modelos de clasificación pueden presentar un sesgo en sus rendimientos. Por ejemplo, en datos biomédicos es común encontrarse con una clase de salida infrarrepresentada, debido a la dificultad de la obtención de datos o a los pocos casos que existen. En esta situación, el modelo tenderá a categorizar hacia la clase mayoritaria, obteniendo, sin embargo, un rendimiento en *accuracy* alto. Por lo tanto, es importante atender a diferentes métricas de rendimiento para detectar este tipo de comportamiento. Por ejemplo, rendimientos altos en *accuracy* con valores bajos en AUC, son indicativos de tal comportamiento erróneo de un modelo. Existen técnicas capaces de resolver el problema de desbalanceo añadiendo muestras de la clase minoritaria [53] o eliminando muestras de la clase mayoritaria [54]. En biomedicina, por ejemplo, es más idóneo el submuestreo, ya que añadir muestras sintéticas podría generar un sesgo importante en el modelo. El sobremuestreo es utilizado mayoritariamente en el análisis de imágenes al poder añadirles diferentes filtros [55, 56, 57, 58].

Por otra parte, las variables del conjunto de datos suelen estar en diferentes escalas y ser de distintos tamaños. Por ello, es difícil hacer la comparación y que el modelo pueda extraer información útil. En datos transcriptómicos, es necesario una normalización de las pruebas de secuenciación. Existen numerosas formas de normalización [59, 60, 61], ninguna de ellas mejor que la anterior, ya que su utilización depende en gran medida del problema y el modelo utilizado. Posteriormente, en muchas ocasiones es idóneo la conversión de los datos a una distribución normal, generando variables con media igual a cero y desviación típica igual a uno. Esto se consigue aplicando transformaciones logarítmicas u otras técnicas específicas a datos ómicos, como se describe en [62], para poder observar patrones más claramente. Es importante tener en cuenta que si se normalizan valores *ouliers* el proceso de normalización escalará los valores normales a valores muy pequeños, lo cual es un comportamiento no deseado. Finalmente, los datos deben ser transformados para presentar la misma escala. Este proceso facilita el aprendizaje y generalización de los modelos de ML.

#### 1.4.1.6 Validación del modelo

Para evitar el sobreentrenamiento de los modelos en la fase de aprendizaje, es necesario realizar procesos de validación. En esta fase lo datos se separan en dos conjuntos: entrenamiento y test. En circunstancias donde el conjunto de datos no es lo suficientemente grande como para tener un conjunto de test muy amplio y evitar el sesgo de selección, se recurre a técnicas de CV [63]. La idea del *k-fold CV* consiste también en separar los datos en los dos conjuntos de entrenamiento y test. La particularidad de esta técnica, es que el conjunto total de datos se separa en $k$ conjuntos que no se solapan. $k-1$ conjuntos son utilizados para entrenar el modelo, y el conjunto restante es utilizado como test. El proceso es repetido $k$ veces, utilizando cada subgrupo como test una vez. El rendimiento promedio de todos los conjuntos de test es reportado como el rendimiento final del modelo. Este proceso puede repetirse $n$ veces para una mayor validación.

Para obtener estimaciones de rendimiento precisas de un modelo, todas las partes de construcción del mismo deben incluirse en la validación. En el caso de pasos como el ajuste de parámetros de un algoritmo, que requiere un proceso de validación, esto da lugar a dos bucles de validación anidados (denominado también *Nested Resampling*). Este proceso se utiliza en la validación de los resultados obtenidos en el Capítulo 3. La característica de este proceso es la presencia de un CV interno para la selección de los mejores hiperparámetros del modelo y un CV externo para evaluar el modelo de forma general. La Figura 1.11 muestra el proceso seguido en este tipo de validación.

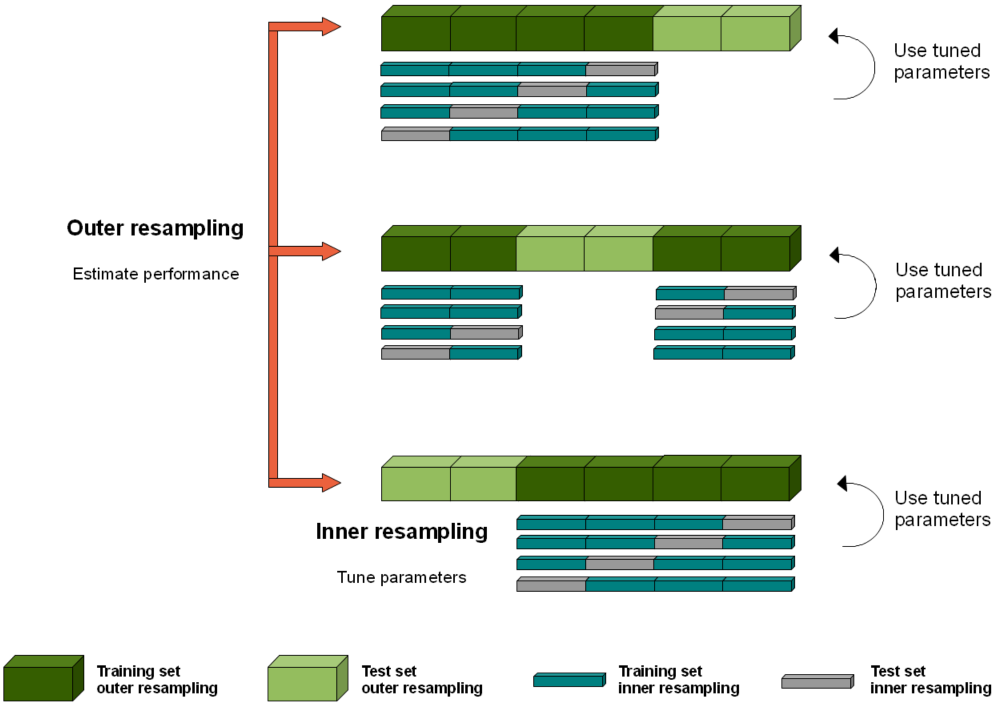


Fig. 1.11.: Método *Cross-Validation* anidada. Existen dos tipos de validación, interna y externa. La interna se utiliza para el ajuste de los hiperparámetros. La externa es utilizada para la validación general del modelo. Imagen extraída de [64].

### 1.4.1.7 Selección del mejor modelo

Este tipo de experimentos computacionales permiten probar diferentes combinaciones de datos y algoritmos con el fin de obtener los mejores rendimientos posibles. En muchas ocasiones es complejo seleccionar el mejor modelo. Para ello, es necesario atender a diferentes factores, como puede ser el rendimiento, la complejidad y el coste computacional de los modelos.

El valor en rendimiento es el primer dato que salta a la vista cuando se identifica el mejor modelo. En este caso, quedarse con el valor alto no siempre es la regla. Anteriormente se habló del proceso de *cross-validation*, el cual genera $k * n$ modelos diferentes (siendo $k$ el número de *folds* y $n$ el número de repeticiones). De esta manera se obtiene un vector (por cada par dato-modelo) de longitud $k * n$. Con el fin de determinar si el rendimiento de un particular algoritmo es estadísticamente mejor que los otros, se realiza un test de hipótesis nula. Además, es necesario comprobar las condiciones para saber qué tipo de test debe utilizarse. Se utilizará un test paramétrico en el caso de cumplirse tres hipótesis concretas: independencia de las variables, normalidad y heterocedasticidad [65].

Hay que tener en cuenta que estos supuestos no se refieren al conjunto de datos utilizados como entrada de las técnicas, sino a la distribución del rendimiento de las mismas. En estadística, un suceso es independiente de otros si el hecho de que uno ocurra no modifica la probabilidad de los demás. Es obvio que diferentes versiones de un conjunto de algoritmos distintos, en los que se utilizan semillas iniciales de forma aleatoria para la separación de los datos en el entrenamiento y la prueba, cumplen la condición de independencia. La normalidad se considera el comportamiento de una observación que sigue una distribución normal o Gaussiana; para comprobar esta condición, existen diferentes pruebas, como la de Shapiro-Wilk [66]. Por último, la violación de la hipótesis de igualdad de varianzas, o heterocedasticidad, debe comprobarse utilizando, por ejemplo, una prueba de Bartlett [67]. En caso de cumplir los tres requisitos mencionados, se debe utilizar un test paramétrico, en caso contrario, un test no paramétrico. En [68] se discute que tipo de test utilizar en cada uno de los casos.

Cuando no existe una diferencia significativa entre el rendimiento de ningún algoritmo, y mientras que los rendimientos sean satisfactorios, debe elegirse el modelo más simple. En este caso, el modelo más simple es aquel que presenta un menor número de variables para ser entrenado, o aquel que presente el menor número de hiperparámetros. De esta forma, se espera que dicho modelo sea más generalizable en datos externos, que otros más complejos.

### 1.4.2 Técnicas de selección de características

Como se ha visto en la sección anterior, los algoritmos de ML realizan una predicción de los valores de salida a partir de los valores de entrada. Dichos valores de entrada deben contener la información sin reiterar y evitar entradas correlacionadas para que los modelos sean capaces de entrenar y predecir con un bajo error. En la práctica, uno de los problemas más comunes que aparecen en un análisis de datos, en especial, en el campo de la bioinformática, es su alta dimensionalidad. De esta forma, se tiene a disposición un conjunto enorme de variables que describen a un paciente y/o muestra. Por ejemplo, en un análisis transcriptómico, se generan alrededor de 50.000 variables. Gran parte de estas variables pueden estar muy correlacionadas y contener, por tanto, información irrelevante.

Las técnicas de selección de características son utilizadas en el análisis de datos moleculares para evitar dichos problemas. Debido a que un modelo de aprendizaje supervisado puede verse afectado por el número y la relevancia de las variables de entrada, el objetivo de estas técnicas es encontrar el subconjunto de variables de entrada que mejor describa la estructura de los datos, tan bien o mejor que con el conjunto total de variables.

Como se describe en Saeys et al. [69], se pueden diferenciar tres tipo de técnicas de selección de características: *filter*, *wrapper* y *embedded*. A continuación se describen estas tres técnicas de forma detallada. La Figura 1.12 ilustra de forma esquemática el funcionamiento de los tres tipos de técnicas.

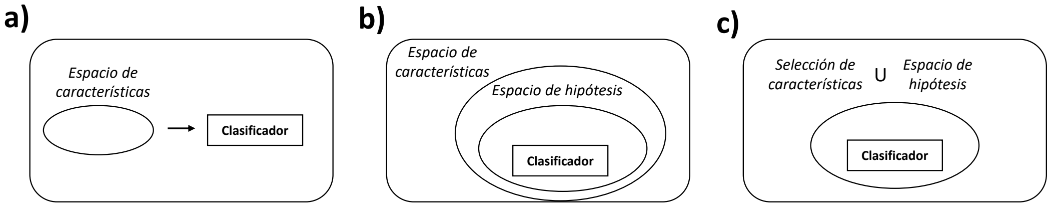


Fig. 1.12.: Tipos de técnicas de *Feature Selection*. a) técnica *filter*; b) técnica *wrapper*; c) técnica *embedded*. En cada imágen se representa el espacio de características, el clasificador y el espacio de la hipótesis. La relación entre ellas define la estrategia de búsqueda de características. Figura modificada de [69]

### 1.4.2.1 Técnicas filter

Las técnicas *filter* (ver Figura 1.12a) evalúan la relevancia de las características teniendo en cuenta únicamente las propiedades intrínsecas de los datos, estableciendo un *ranking* por orden de relevancia. Las variables con menor puntación son eliminadas del conjunto total creando un nuevo espacio de variables de entrada menor para el clasificador. Con estas técnicas, la selección de características se realiza una única vez. Además, diferentes algoritmos pueden ser evaluados. Son consideradas técnicas simples y rápidas computacionalmente y son independientes del clasificador.

El *ranking* se establece tras el cálculo de un *score*. Las variables que queden por debajo de un umbral límite son eliminadas del conjunto total de los datos. Una propiedad básica de las características que se busca con el cálculo del *score* es que contenga información útil sobre las diferentes clases de salida. Esta propiedad puede ser definida como la relevancia de la característica [70, 71, 72, 73, 74]. En esencia, esta definición afirma que si una característica ha de ser relevante puede ser independiente de los datos de entrada pero no puede ser independiente de las etiquetas de clase, es decir, la característica que no tiene influencia en las etiquetas de la clase puede ser descartada.

La correlación entre características desempeña un papel importante en la determinación de características únicas. En las aplicaciones prácticas, la distribución subyacente de los datos a menudo no es conocida y se mide por la precisión del clasificador. Debido a esto, un subconjunto de características óptimo puede no ser único porque puede ser posible alcanzar la misma precisión del clasificador utilizando diferentes conjuntos de características.

Existen dos tipos de técnicas *filter*: univariadas y multivariadas. Las primeras consideran cada variable independiente de las demás. Por lo tanto, ignoran las posibles dependencias existentes entre variables. Por otra parte, las técnicas multivariadas pretenden incorporar en cierto nivel, posibles dependencias entre las variables, utilizando métricas basadas en correlación. Ejemplos de técnicas univariadas son la distancia Euclídea, los *i*-test [75, 76] o *scores* basados en el *Information Gain*, tales como el *Gain Ratio* [77]. Por otro lado, ejemplos de técnicas multivariadas serían *Correlation-based feature selection* [78], *Markov blanket filter* [79] ó *Fast correlation-based feature selection* [80].

Las técnicas *filter* son las más utilizadas durante el desarrollo de la tesis, principalmente por la interpretabilidad de sus resultados y por la reducción del tiempo computacional necesario. En problemas donde la prioridad es interpretar los mecanismos biológicos subyacentes, no es conveniente el uso de técnicas donde sea dependiente el tipo de algoritmo utilizado para la justificación de resultados. En este contexto, las técnicas *filter*, al ser independientes del modelo, permiten el entrenamiento de diferentes modelos con los subconjuntos de variables obtenidos y su consecuente interpretación.

#### 1.4.2.2 Técnicas wrapper

A diferencia de las técnicas *filter*, las técnicas *wrapper* encapsulan el modelo dentro de la búsqueda de características. De esta forma, se generan posibles subconjuntos de características que son evaluados. La evaluación es llevada a cabo por un clasificador específico, a través de los conjuntos de entrenamiento y test. Para buscar en el espacio de todos los subconjuntos de características, se encapsula (*wrapper*, en inglés) un algoritmo de búsqueda alrededor del modelo de clasificación (ver Figura 1.12b). Sin embargo, debido a que el espacio de los subconjuntos de características aumenta exponencialmente con el número de características, métodos de búsqueda heurísticas son utilizados para guiar la búsqueda de un subconjunto óptimo.

Los métodos de búsqueda en las técnicas *wrapper* pueden clasificarse según dos tipos: algoritmos de búsqueda deterministas y aleatorios. Ejemplos de algoritmos deterministas son *Sequential forward selection* [81] o *Sequential backward elimination* [81]. Ejemplos de algoritmos aleatorios son los algoritmos genéticos [82] o *Estimation of distribution algorithms* [83].

#### 1.4.2.3 Técnicas embedded

Estos métodos combinan a la vez el proceso de entrenamiento con la búsqueda en el espacio de variables (ver Figura 1.12c). Estas técnicas tienen la ventaja que incluyen la interacción con el clasificador, aunque al mismo tiempo son más costosos computacionalmente que los otros dos tipos de técnicas.

Ejemplos de técnicas *embedded* son los árboles de decisión, *Weighted naive Bayes* [84] o LASSO [37].

### 1.4.3 Modelos QSAR

Bajo las premisas *la estructura de una molécula define su actividad biológica* y *moléculas estructuralmente similares tienen una actividad biológica similar*, los modelos de relación cuantitativa entre estructura y actividad (QSAR, por sus siglas en inglés) son capaces de relacionar numéricamente las estructuras químicas de las moléculas con su actividad biológica.

Para ello, los modelos QSAR integran técnicas informáticas y estadísticas para realizar una predicción teórica de la actividad biológica, reduciendo el proceso de ensayo y error en el desarrollo de fármacos. Al ser una ciencia que existe sólo en un entorno virtual, permite prescindir de ciertos recursos como equipos, instrumentos, materiales y personal de laboratorio, ofreciendo metodologías mucho más baratas y rápidas para el *screening* de fármacos.

Aunque se han realizado numerosos y diversos experimentos con la metodología QSAR [85, 86, 87, 88, 89], existe una metodología general común en todos ellos (ver Figura 1.13). La obtención de datos de secuencias y/o estructuras moleculares son extraídos de grandes repositorios tales como PubChem [90], DrugBank [91], ZINC [92] ó ChEMBL [93]. La ventaja de estos repositorios es que las moléculas están debidamente etiquetadas con formatos estándares. En cuanto a su formato, normalmente se trabaja con datos *fasta* (para proteínas) ó *smiles* (para moléculas pequeñas).

A continuación, es necesario la conversión de los datos de ambos formatos en matrices numéricas. Como se observa en la Figura 1.13, las moléculas son convertidas en una matriz de valores binarios. Este es un caso entre otros que veremos a continuación. Estos descriptores son conocidos como *fingerprints*, que son vectores binarios donde el valor 1 representa la presencia de una subestructura y el valor 0 su ausencia. Existen numerosos *fingerprints* ya definidos, tales como MACCS [94] ó ECFP [95]. Además de estos, existen otros tipos de descriptores que presentan valores continuos, como los utilizados en [3].

Una vez se ha obtenido una matriz numérica, los datos están listos para entrenar diferentes algoritmos predictivos. En este punto, la metodología se convierte en una metodología de Machine Learning (ver sección **Machine Learning**).

En cuanto a los descriptores moleculares, estos pueden dividirse en dos categorías principales. Los experimentales (refractividad molar, momento dipolar, etc.) y los descriptores moleculares teóricos, que se derivan de una representación simbólica de la molécula. Estos últimos, en los que se basan los modelos QSAR, que pueden clasificarse en:

- **Constitucionales**: reflejan propiedades generales de la naturaleza molecular

- **Topológicas**: su cálculo se realiza a través de la teoría de grafos

- **Geométricas**: se derivan de esquemas empíricos y codifican la capacidad de la molécula para participar en diferentes tipos de interacciones.

- **Electrónicas**: se refieren a las propiedades electrónicas

- **Físico-químicas**: definen el comportamiento de la molécula frente a reacciones externas

Estos tipos de descriptores también pueden ofrecer información acerca de las dimensiones de la molécula. Es decir, existen descriptores que se calculan a partir del grafo molecular de un compuesto químico, mientras que otro ofrecen información de su composición 3D. El uso y la complejidad de los diferentes descriptores depende de

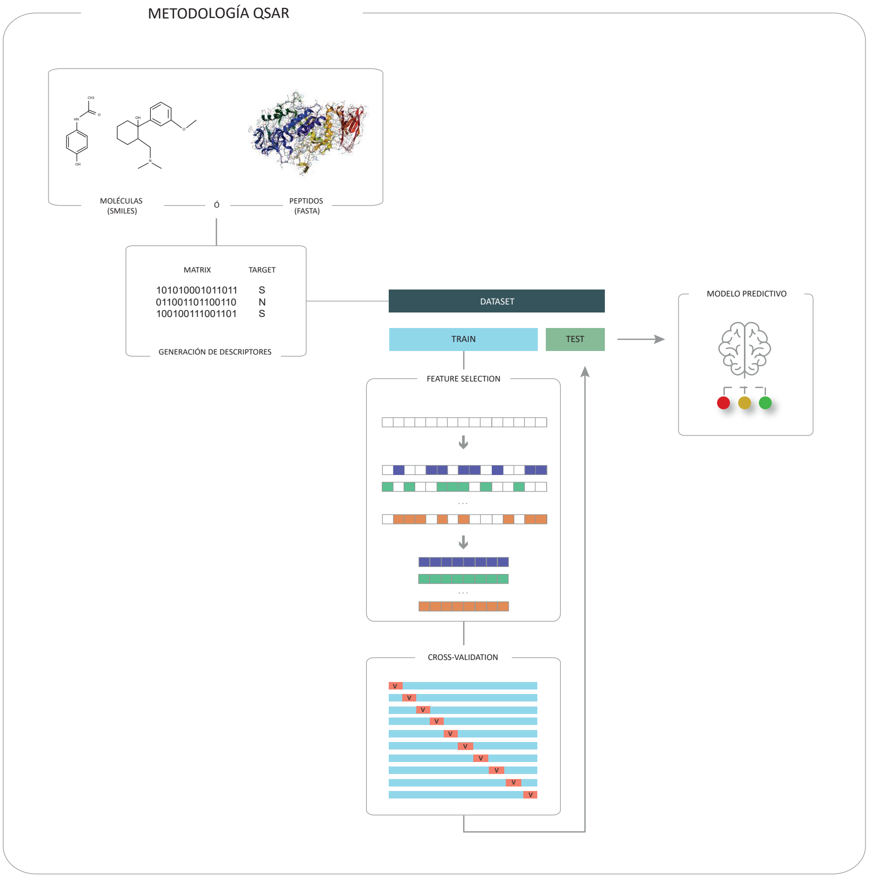


Fig. 1.13.: Metodología QSAR. Las secuencias tanto de moléculas (SMILES) como de proteínas (fasta) son convertidos en matrices numéricas. Cada fila corresponde con una molécula y es etiquetada según cierta característica biológica. Posteriormente se realiza una metodología de búsqueda de características, entrenamiento del modelo y validación mediante *cross-validation*. Adaptado de [96]

las características del problema a resolver. Una revisión extensa de diferentes trabajos que han utilizado técnicas de ML y descriptores moleculares para el descubrimiento de fármacos puede encontrarse en [96]

### 1.4.3.1 Descriptores de composición aminoacídica

Como se ha explicado anteriormente, los descriptores moleculares pueden calcularse a partir de moléculas pequeñas o a partir de péptidos.

Para la comparación y clasificación de los péptidos según su actividad, es necesario extraer información adicional de su secuencia.

Una secuencia de proteínas o péptidos con $N$ residuos de aminoácidos puede representarse generalmente como $R1, R2, ..., Rn$, donde $Ri$ representa el residuo en la posición $i - ésima$ de la secuencia. Las etiquetas $i$ y $j$ se utilizan para indexar la posición del aminoácido en una secuencia, y $r$, $s$, $t$ se utilizan para representar el tipo de aminoácido. A continuación se muestra el cálculo de tres tipos de descriptores:

- **Amino Acid Composition (AAC)**: describe la fracción de cada tipo de aminoácido dentro de una secuencia proteica. Se calcula la fracción de los 20 aminoácido de la siguiente forma:

$$f(r) = \frac{N_r}{N} \tag{1.13}$$

  donde $N_r$ es el número de aminoácido de tipo $r$ $(r = 1, 2, ..., 20)$ y $N$ es la longitud de la secuencia.

- **Di-Aminoacid Composition (DC)**: este descriptor genera una salida de 400 dimensiones, calculadas de la siguiente forma:

$$f(r, s) = \frac{N_r s}{N - 1} \tag{1.14}$$

  donde $N_r s$ es el número del dipéptido representado por el aminoácido de tipo $r$ y del tipo $s$ $(r, s = 1, 2, ..., 20)$.

- **Tri-Aminoacid Composition (TC)**: en este caso, TC ofrece una salida de 800 dimensionales, definidas como:

$$f(r, s, t) = \frac{N_r st}{N - 2} \tag{1.15}$$

  donde $N_r st$ es el número del tripéptido representado por el aminoácido de tipo $r$, del tipo $s$ y del tipo $t$ $(r, s, t = 1, 2, ..., 20)$.

De esta forma, estos descriptores ofrecen información de la secuencia proteica. Teóricamente, la longitud de búsqueda de secuencias podría incrementarse hasta infinito. El factor limitante sería la generación de un número de dimensiones inabarcables computacionalmente. Por ejemplo, si buscamos secuencias aminoacídicas de longitud igual a ocho, el número de variables generadas serían $20^8 = 25{,}600{,}000{,}000$. Estos descriptores deben integrarse y seleccionarse para adaptarse mejor al problema a resolver. En [2] se han utilizado únicamente estos tres tipos de descriptores, con resultados muy satisfactorios.

### 1.4.4 Técnicas de *docking* molecular para el reposicionamiento de fármacos

Los métodos de cribado de alto rendimiento (o *High-throughput screening* (HTS) en inglés) son procesos que permiten probar de forma automatizada un gran número de compuestos para un objetivo biológico específico. Los HTS se utilizan ampliamente en la industria farmacéutica, aprovechando la robótica y la automatización para probar rápidamente bibliotecas de compuestos a gran escala de forma rentable. Entorno a $10^3 - 10^6$ pequeñas moléculas, de estructura conocida, son cribadas de forma paralela en estos experimentos. Aun así, el número de posibilidades es mucho más grande. Por ejemplo, la base de datos de ChEMBL [93] alberga un total de 2.1M compuestos, el repositorio DrugBank [91] consta de 14K compuestos, PubChem [90] presenta 110M y la base de datos de ZINC [92] tiene 750M compuestos. Cribar todos estos compuestos con diferentes dianas moleculares es un trabajo experimentalmente inabarcable.

En el campo de la investigación, para reducir tiempo y costes, se utilizan modelos *in silico* para el estudio de todas las posibles interacciones. Estos modelos, mediante técnicas de computación de alto rendimiento, permiten el cribado virtual de millones de compuestos en un tiempo asequible e identifican potenciales candidatos para testar posteriormente de forma experimental.

En lo que respecta a las dianas terapéuticas, se está realizando una inmensa inversión en términos de tiempo, personal, recursos y dinero, para validar experimentalmente nuevas dianas biológicas y nuevos fármacos que actúen eficazmente sobre ellas. Una vez identificado el fármaco a validar, se realizan experimentos para comprobar y verificar si el fármaco tiene el efecto esperado en modelos celulares y animales. Posteriormente, es necesario realizar un ensayo clínico, en el que hay que reclutar

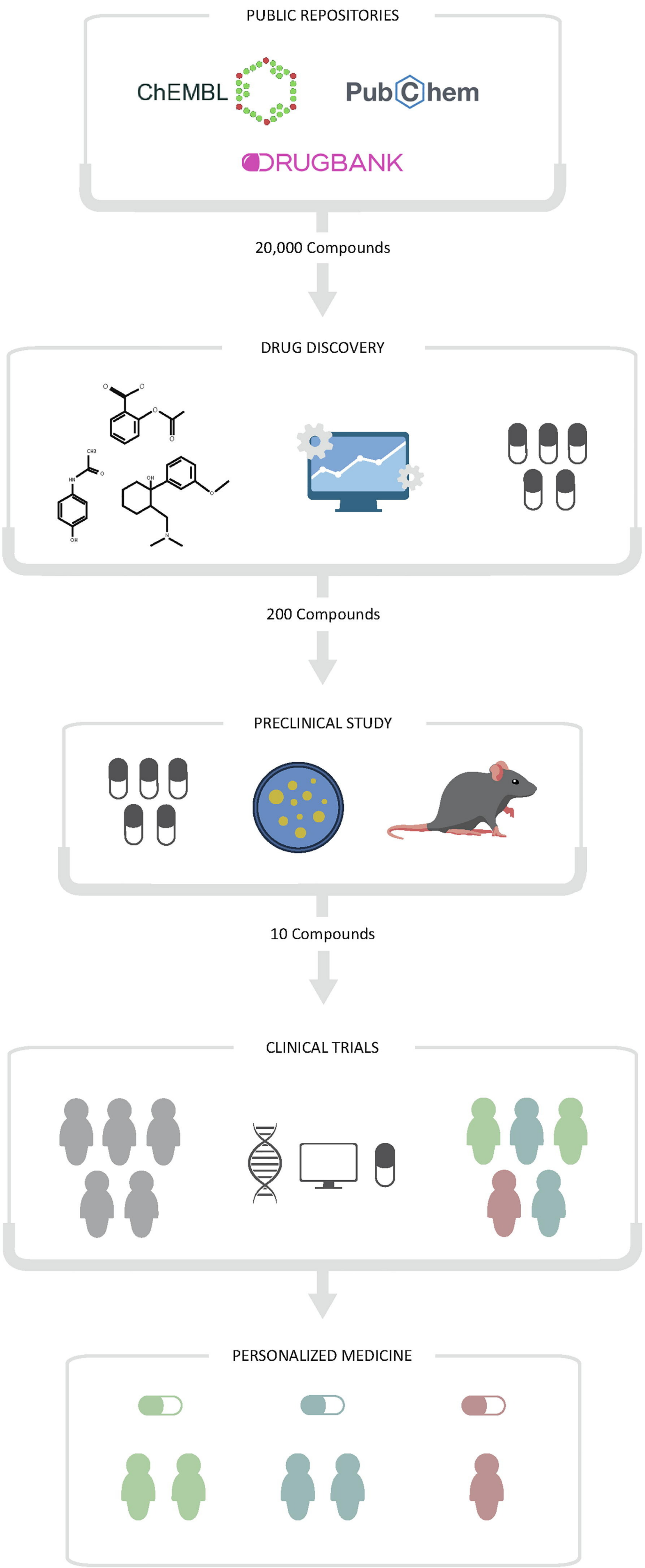


Fig. 1.14.: Representación del proceso de descubrimiento de fármacos. Figura extraído de [96]

un número importante de pacientes, analizar todos los aspectos adversos y superar todos los controles de calidad. Es aquí, en la fase de ensayos clínicos, donde los presupuestos se disparan. Por último, los pacientes deben de ser estratificados según características moleculares y ofrecerles un medicamento personalizado acorde a sus características. Una simplificación de este proceso se observa en la Figura 1.14.

Hay que tener en cuenta que la tasa de fracaso global en la desarrollo de medicamentos es del 96 % [97, 98, 99], lo que implica una inversión desproporcionada por parte de las empresas farmacéuticas. Por lo tanto, encontrar vías y atajos desde la investigación básica a la clínica a través de la investigación traslacional ofrece una ventaja significativa en este campo de investigación.

Existen varios métodos de cribado computacional basados en la estructura de las moléculas. Estos métodos se centran en la información 3D de una diana de interés para calcular la complementariedad estructural y electrónica de cada ligando con la diana. Dentro de estas técnicas, el *docking* molecular es el más popular y exitoso.

El *docking* molecular es un procedimiento computacional que intenta predecir la unión no covalente de macromoléculas o, más frecuentemente, de una macromolécula (receptor) y una molécula pequeña (ligando) de forma eficiente, partiendo de sus estructuras no unidas. El objetivo es predecir las conformaciones de unión y la afinidad de unión. En la Figura 1.15 se muestra una representación de un experimento de *docking* molecular mediante AutoDock Vina [2]. En este caso se muestra la conformación de la interacción más significativa entre la droga Abemaciclib y el producto del gen FABP6. La descripción del problema, y los resultados obtenidos se presentan en la sección 3.2. Para una mejor comprensión del funcionamiento del software, se recomiendan los siguientes trabajos [100, 101].

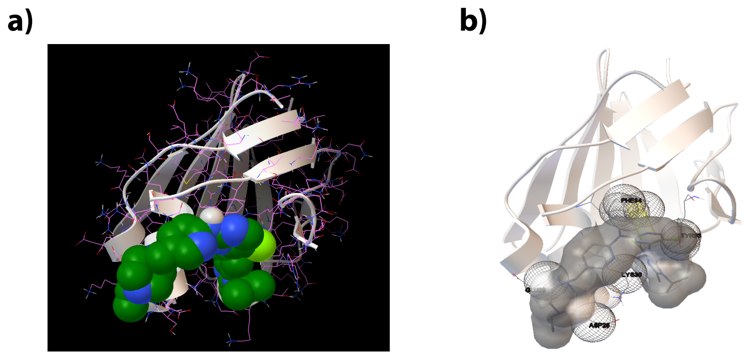


Fig. 1.15.: Representación del cálculo de *docking* molecular. a) sin interacción; b) muestra los átomos que interaccionan entre las dos moléculas. Imágen extraída de [102]

La asociación entre ligandos y proteínas dianas juega un papel principal en la transducción de señales. Si un par ligando-proteína presenta una afinidad de enlace significativa, teóricamente, la unión es posible para formar un complejo estable. Si se considera el ligando como un fármaco y la proteína como una diana terapéutica, la formación del complejo conlleva la inhibición de la proteína. Bajo esta hipótesis, el *docking* molecular facilita el descubrimiento de nuevos fármacos.

Desde hace algunos años, se han comunicado resultados prometedores sobre el efecto de varios fármacos en enfermedades para las que no fueron diseñados. Por ejemplo, el fármaco Zidovudina, que originalmente fue diseñado para el tratamiento del cáncer, se ha utilizado posteriormente para el VIH/SIDA. Otro ejemplo es el Rituximab, que originalmente estaba indicado para varios tipos de cáncer y que posteriormente ha sido aprobado para la artritis reumatoide, o el Raloxifeno, que pasó de utilizarse para la osteoporosis a emplearse para el cáncer de mama. Estos y otros muchos ejemplos puede verse en el artículo de revisión realizado por Pushpakom, S. *et al.* [103]. Probar experimentalmente todos los fármacos utilizados hoy en día en todas las enfermedades es inviable. Afortunadamente, con el aumento de la capacidad computacional, las técnicas de reutilización de fármacos pueden dar una aproximación realista de lo que podría ocurrir en la naturaleza.

En relación al cáncer existen numerosos fármacos anti-tumorales aprobados [104]. Trabajos previos han ofrecido un trabajo basado en el reposicionamiento de fármacos no anti-tumorales [105]. En esta tesis, se ofrece una aproximación basada en el reposicionamiento de fármacos anti-tumorales aprobados. El reposicionamiento de fármacos no tumorales (tales como estatinas, bloqueadores del receptor de la angiotensina, aspirina o vitamina D) carecen de actividad como agente único en el tratamiento del cáncer (considerado un requisito útil para el desarrollo de fármacos) [106]. Por el contrario, los fármacos anti-tumorales ya aprobados normalmente actúan como agentes únicos, por lo que son unos candidatos idóneos para su reposicionamiento. Por ejemplo, observar que un fármaco aprobado para cáncer de mama, tiene efecto también en cáncer de colon genera una oportunidades enormes en la industria farmacéutica. En la sección 3.2 se presentan unos resultados relacionados con esta aproximación.

# 2 Objetivos

En esta sección se enumeran los objetivos que han motivado el desarrollo de esta tesis doctoral.

El objetivo general del trabajo de investigación que se presenta es la **aplicación de técnicas de Machine Learning** para la búsqueda de **nuevas formas de diagnóstico y tratamientos de precisión** en pacientes con **cáncer** mediante el análisis de datos ómicos.

Los objetivos específicos que surgen del objetivo general son:

1. Explorar el uso de metodologías de Machine Learning utilizadas en investigación oncológica de precisión a partir de datos ómicos. Identificar posibles debilidades metodológicas con el fin de desarrollar posteriores modelos de Machine Learning, robustos y reproducibles, que ayuden a su aplicación en la práctica clínica habitual.

2. Desarrollar modelos basados en Machine Learning que permitan el descubrimiento y reposicionamiento de fármacos en el contexto de la Medicina de Precisión.

3. Utilización de algoritmos de Machine Learning para la identificación de biomarcadores y *pathways* alterados en pacientes de cáncer que ayuden a la estratificación de pacientes con el fin de buscar tratamientos específicos.

# Resultados 3

Este capítulo recoge los tres trabajos de investigación que componen la tesis doctoral. Cada sección corresponde a un artículo JCR publicado. En conjunto, dan respuesta a los objetivos enumerados en el capítulo anterior. Se presenta un estudio de revisión sobre el uso de las técnicas de Machine Learning para el análisis de datos ómicos en cáncer. Posteriormente, en base a las consideraciones extraídas de la revisión, se aplica una metodología de ML robusta y reproducible para el *screening* y reposicionamiento de fármacos anti-tumorales y para la identificación de nuevos marcadores y/o *pathways* alterados en pacientes con cáncer.

## 3.1 Revisión y estado del arte del uso de técnicas de Machine Learning en oncología

Este trabajo tiene como objetivo explorar el uso de metodologías de Machine Learning aplicadas en investigación oncológica de precisión a partir de datos ómicos.

Estudiando trabajos previos que utilizaron los datos del TCGA, se revisaron más de 150 artículos publicados que aplicaron técnicas de ML para analizar dichos datos. La mayoría de los trabajos revisados aplicaban soluciones de entrenamiento previamente mencionados en esta tesis doctoral: supervisado y no supervisado. Además, los trabajos están clasificados según el problema biológico, el algoritmo y el tipo de dato ómico utilizado.

Las características del repositorio del TCGA (grandes bases de datos de diferentes ómicas) propicia un ambiente idóneo para la comparación de metodologías. De esta forma, partiendo de los mismos datos, se puede comparar la aplicación de diferentes metodologías basadas en ML para la resolución de diversos problemas. Durante el desarrollo de la revisión se identifican las debilidades en la aplicación de las metodologías: principalmente, la falta de reproducibilidad de los resultados y el sobreentrenamiento de los modelos. La revisión de artículos ayudó a diseñar unas directrices para aplicar modelos de ML en el análisis de datos ómicos. Siguiendo estas pautas, se aplicaron modelos de ML para la resolución de dos problemas en biomedicina: 1) identificación de firmas y/o biomarcadores alterados en pacientes con cáncer de colon (ver sección 3.2); 2) *screening* automático de fármacos anti-tumorales (ver sección 3.3). Ambos problemas siguen los objetivos de una Medicina de Precisión.

En cuanto a las tendencias de los trabajos revisados, los modelos basados en *kernel* y en concreto, el modelo SVM y sus variantes es el más utilizado en este campo, seguido por modelos basados en árboles de decisión. Se ha detectado una tendencia creciente en el uso de redes neuronales, principalmente con topologías de Deep Learning, aunque presentan grandes dificultades para su aplicación en bases de datos ómicas, debido al bajo número de muestras que presentan estos estudios.

Por otra parte, los datos de expresión genética son los más frecuentemente utilizados. Su uso, combinado con otros datos ómicos, tales como miRNA y/o datos de metilación, obtienen mejores rendimientos de los modelos, como se indica en la revisión. El trabajo también discute las formas y los rendimientos alcanzados con las diferentes formas de integración de datos ómicos.

Finalmente, se realiza una clasificación de los problemas biológicos abordados por cada tipo de cáncer. Se observa, por ejemplo, como la cohorte de glioblastoma es la que más estudios de supervivencia presenta, principalmente por la alta mortalidad, que implica un mayor número de eventos e incrementa el poder de este tipo de análisis. Por otro lado, las cohortes de mama y riñón son las mas utilizadas en la búsqueda de nuevos subtipos, mediante técnicas de aprendizaje no supervisado.

Fui el autor principal de esta publicación. Concebí y diseñé la metodología utilizada, realicé la revisión de los artículos y preparé todas las figuras y las tablas. La información de la participación de cada autor está disponible en el artículo.

# Machine learning analysis of TCGA cancer data

Jose Liñares-Blanco[1,2], Alejandro Pazos[1,2,3] and
Carlos Fernandez-Lozano[1,2,3]

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruna, A Coruña, Spain

[2] Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruna, A Coruña, Spain

[3] Grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos. Imagen Médica y Diagnóstico Radiológico (RNASA-IMEDIR). Complexo Hospitalario Universitario de A Coruña (CHUAC), SERGAS, Universidade da Coruña, Instituto de Investigación Biomédica de A Coruña (INIBIC), A Coruña, Spain

## ABSTRACT

In recent years, machine learning (ML) researchers have changed their focus towards biological problems that are difficult to analyse with standard approaches. Large initiatives such as The Cancer Genome Atlas (TCGA) have allowed the use of omic data for the training of these algorithms. In order to study the state of the art, this review is provided to cover the main works that have used ML with TCGA data. Firstly, the principal discoveries made by the TCGA consortium are presented. Once these bases have been established, we begin with the main objective of this study, the identification and discussion of those works that have used the TCGA data for the training of different ML approaches. After a review of more than 100 different papers, it has been possible to make a classification according to following three pillars: the type of tumour, the type of algorithm and the predicted biological problem. One of the conclusions drawn in this work shows a high density of studies based on two major algorithms: Random Forest and Support Vector Machines. We also observe the rise in the use of deep artificial neural networks. It is worth emphasizing, the increase of integrative models of multi-omic data analysis. The different biological conditions are a consequence of molecular homeostasis, driven by both protein coding regions, regulatory elements and the surrounding environment. It is notable that a large number of works make use of genetic expression data, which has been found to be the preferred method by researchers when training the different models. The biological problems addressed have been classified into five types: prognosis prediction, tumour subtypes, microsatellite instability (MSI), immunological aspects and certain pathways of interest. A clear trend was detected in the prediction of these conditions according to the type of tumour. That is the reason for which a greater number of works have focused on the BRCA cohort, while specific works for survival, for example, were centred on the GBM cohort, due to its large number of events. Throughout this review, it will be possible to go in depth into the works and the methodologies used to study TCGA cancer data. Finally, it is intended that this work will serve as a basis for future research in this field of study.

Submitted 3 March 2021
Accepted 17 May 2021
Published 12 July 2021

Corresponding author
Carlos Fernandez-Lozano,
carlos.fernandez@udc.es

Academic editor
Li Zhang

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj-cs.584**

How to cite this article Liñares-Blanco J, Pazos A, Fernandez-Lozano C. 2021. Machine learning analysis of TCGA cancer data. **PeerJ Comput. Sci. 7:e584 DOI 10.7717/peerj-cs.584**

**Subjects** Bioinformatics, Computational Biology, Artificial Intelligence, Data Mining and Machine Learning
**Keywords** Cancer, Machine learning, TCGA, Multi-omics, Data integration, BRCA, Random Forest, Support Vector Machines

# INTRODUCTION

The appearance of the carcinogenic phenotype is the consequence of an alteration of one or more genes. In addition, the appearance of subtypes occurs in different ways in individuals of a population. Hence, a major problem that arises in cancer is the difficulty in its genetic diagnosis. Similar to Mendelian diseases, where the disease develops due to the alteration in the function of a single gene, the development of cancer is a consequence of epistatic behaviour of genes. There is already an extremely large search space in the identification of alterations in a single gene, including exonic and intronic mutations, single nucleotide polymorphisms (SNPs), copy number variants, indels, post-transcriptional alterations, post-translational alterations, three-dimensional assembly of the protein, epigenetic modifications, etc. Thus, the search space for alterations when we encounter a subgroup of 40 genes is immense. When we do not know exactly which genes are involved, we have to search among the more than 20,000 coding regions or even in whole genome sequence. In these cases the search space grows to incalculable levels. All this complexity is the result of intermolecular communications in and among cells, a phenomenon that constitutes an environment of molecular communication that is extremely complicated to understand and identify.

In order to lay the foundation and achieve great advances in the prevention, early detection, stratification and success in the treatment of cancer, it is necessary to identify the complete changes generated by each type of cancer in its genome. Further, researchers must understand how these changes interact with the cancer microenvironment, intra- and intercellularly, to manifest itself. Hence, the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) of the United States established The Cancer Genome Atlas (TCGA), with the aim of obtaining comprehensive multidimensional genomic maps of all key changes in several types and subtypes of cancer (*The Cancer Genome Atlas Research Network, 2008*). An initial pilot project in 2006 confirmed that an atlas of these changes could be specifically created for different types of cancer. Subsequently, TCGA has collected tissues from more than 11,000 cancer and healthy patients, an endeavour that allows the study of more than 33 types and subtypes of cancer, including 10 rare cancers. The most interesting aspect of this initiative is that all the information is free and accessible to any researcher who wants to focus their efforts on the disease. The different types of data presented by the TCGA project are summarised in Table 1 and Fig. 1 shows, for each cancer type, the percentage that each data type represents in the subtype's total. Data are provided open access to the community, a factor that facilitates the generation of novel models without requiring an initial financial investment to obtain the data. Therefore, there are increasingly specific models for the analysis of omics data. In particular, the rise and success of machine learning (ML) techniques to process a large amount of data is revolutionising bioinformatics and

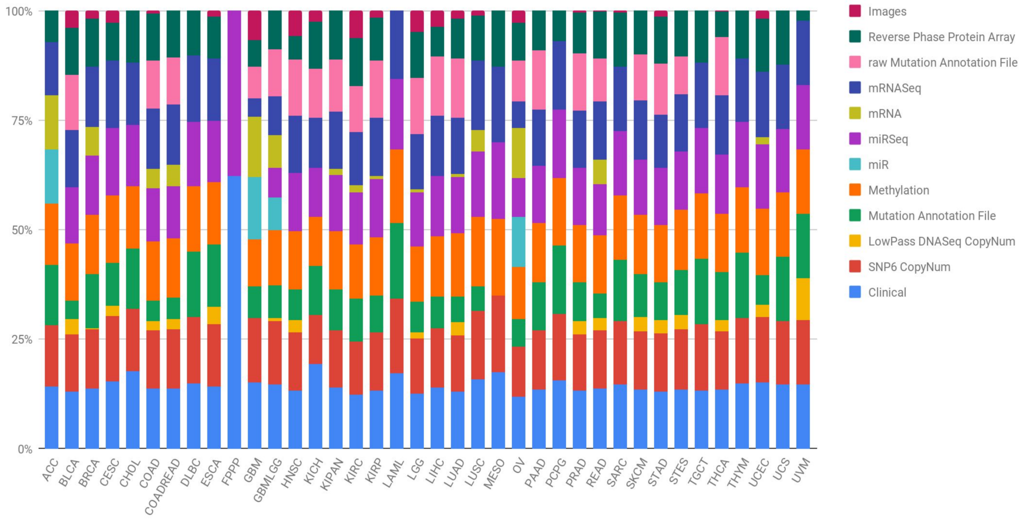


**Figure 1 Quantification of the number of samples in the TCGA repository, classified by type of tumour and type of biotechnological analysis.** Clin, Clinical; SNP6, SNP6 CopyNum; DNAseq, Low-Pass DNASeq CopyNum; Mutat, Mutation Annotation File; Met, Methylation; rawMut, rawMutation Annotation File; Prot, Reverse Phase Protein Array. Full-size DOI: 10.7717/peerj-cs.584/fig-1

conventional forms of genetic diagnosis. These methods have focused on making predictions by using general learning algorithms to find patterns in complex, larger and hard-to-handle problems. In addition, these ML methods work really well with very large datasets, even when the number of variables in each observation is much greater than the total number of observations ($n << p$).

This survey presents the state-of-the-art research on TCGA analysis using machine learning. Efforts have involved both supervised and unsupervised learning problems, as well as survival analysis, disease prognosis, cancer staging and pathways analysis to analyse different types of data ranging from multi-omics human cancer data to imaging. Therefore, review articles are needed to show an overview of machine learning-based analysis of TCGA data to highlight the findings and to discuss future research lines so that the obtained knowledge is useful and can be translated to clinical practice.

There are few published review articles on machine learning for biomedical genomic analysis (*Leung et al., 2015*; *Karczewski & Snyder, 2018*). These review articles are before 2018 and do not present a discussion on TCGA data nor a discussion on machine learning results neither present a multi-omic and imaging point of view for different biological questions. To the best of our knowledge, no survey has been conducted on Machine Learning analysis of TCGA using multi-level cancer data. Thus, this survey aims to present a comprehensive summary of the previous machine learning approaches applied to TCGA during the span of 2008-2020. The contributions of this review are:

- This review includes exhaustive review of the main results obtained by the TCGA consortium using conventional approaches in order to understand if machine learning is increasing the knowledge in the area.
- This review includes machine learning results by the TCGA consortium.

Table 1 **Different types of data present in the TCGA repository.**

| **DNA Sequencing** | **Whole genome sequences** |
|---|---|
| | Whole exome sequences |
| | Sequences traces |
| | Mutations, including coding, splice site, germline and noncoding somatic variants |
| **RNA sequencing** | mRNA sequences (calculated expression per gene, exon, splice junction and isoform) |
| | miRNA sequences (calculated expression per miNRA and isoform) |
| | Total RNA sequences (calculated expression per gene, exon, splice junction and isoform) |
| | Expression signals per gene, exon, splice junction, miRNA and isoform |
| **Copy number** | Arrays (raw, unnormalized, normalized) |
| | Low-pass DNA sequencing (whole genomes sequences, variants and coverage) |
| **Array-based expression** | Gene expression (raw, normalized and calls) |
| | Exon expression (raw, normalized and calls) |
| | miRNA expression (raw, normalized and calls) |
| **DNA methylation** | Array-based methylation (raw signal intensity, calculated beta values) |
| **Other** | Protein expression (high-resolution images of protein arrays, raw signals, normalized expression and mass spectrometry protein) |
| | Microsatelite instability (markers and classification) |
| | ATAC-seq (chromatine accesibility) |
| **Metadata** | Clinical information about patients (e.g., sex, race, ethnicity, drugs taken, metastasis status and response to treatment) |
| | Information about samples (e.g., the weight of a sample portion, days to collect and time of freezing) |
| | Images of the tumors |

- A classification of supervised, unsupervised and clustering methods that may point researchers to new approaches or new problems.
- Identification of data types mostly used in machine learning research of TCGA.
- A comprehensive discussion on biological questions solved by machine learning algorithms: prognosis, immunological phenotype, pathways, MSI status, and subtype prediction.
- A deeper examination of the most used TCGA cohort: Breast Cancer Adenocarcinoma (BRCA).
- We point data integration approaches as the future trend in TCGA analysis using machine learning.

We believe that researchers in machine learning, bioinformatics, biology, computational biology and data integration would benefit from the findings of this exhaustive and comprehensive review.

This manuscript is organised as follows. “Survey Methodology” explains the methodology used in this survey. “TCGA Consortium” presents the main results obtained by the TCGA consortium. In “Machine Learning as a Source of New Knowledge”, we review the TCGA efforts with those algorithms as well as we present the most used algorithms on supervised, unsupervised and clustering approaches for external researchers. Special attention with a subsection on medical imaging analysis using deep learning approaches in recent years. “Biological Questions Solved by Machine Learning Algorithms” discusses the capability of those algorithms to solve the biological problem with the highest performance score and find that the predictions are biologically of relevance. To this aim we divide and study five biological problems: prognosis, immunological phenotype, pathways, MSI and subtypes prediction. We finish with special emphasis on the analysis of the BRCA cohort. Finally, we conclude the review in ‘Conclusions’.

## SURVEY METHODOLOGY

This work is based on a literature review in machine learning-based analysis of TCGA cancer data. We searched for the main findings of the TCGA consortium using classical statistical approaches and works using machine learning and classify them into supervised, unsupervised and clustering methods. Furthermore, we considered of relevance to answer to the intitial biological question with sense, not only with a higher performance score. The search keywords, data sources and on criteria are discussed.

### Search keywords

We initially reviewed the original TCGA consortim publication in order to carefully select the search keywords. The keywords used for the survey included the following terms to find the relevant papers: ‘machine learning’, ‘TCGA’. We used the ‘AND’ and ‘OR’ Boolean operators to combine terms. After the initial subset of papers we refined the search keywords according with the most used machine learning models, type of problems and biological question: ‘clustering’, ‘computer vision’, ‘deep learning’, ‘random forest’, ‘support vector machines’, ‘linear model’, ‘survival’, ‘MSI’, ‘prognosis’, ‘pathway’, ‘subtypes’ or ‘phenotype’.

### Data sources

The papers included in this survey were retrieved from prominent journals indexed in diverse quality databases: Pubmed and Scopus.

### Article inclusion/exclusion criteria

We decided which articles are eligible for the survey under the following inclusion/ exclusion criteria:

- Inclusion criteria:
  - manuscripts written in the English language and published by indexed journals in Pubmed to ensure the health science specialization and Scopus using TCGA as the main source of data
- Exclusion criteria:
  - manuscripts using machine learning marginally or without solid biological conclusions
  - manuscripts in preprint without peer review

### Article selection

The TCGA consortium papers were identified in the website and were included. Initially 345 papers were identified in Pubmed and Scopus using the search keywords. Of these, we filtered by the inclusion/exclusion criteria. In addition, duplicated papers retrieved from multiple sources were removed. Finally, more than 150 articles were included.

## TCGA CONSORTIUM

TCGA began as a pilot project for 3 years, with a focus on the characterisation of three types of human cancer: glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC) and ovarian cancer (OV). TCGA currently presents data from a total of 38 different cohorts. Four of them (COADREAD, GBMLGG, KIPAN and STES) are not original—they are combinations of other cohorts. Among the remaining 34 cancer cohorts are tumours of different tissue types, as can be seen in Table 2. To date, TCGA has characterised and published about 33 different types of tumours in leading international journals. Table 2 provides greater depth for each of the publications that TCGA has made in each recruited cohort.

In 2018, a series of works were published in Cell editorial, where they were exhaustively analysed the samples recruited throughout the project. These studies led to the identification and examination of mechanisms that underlie all types of tumours. These findings allow researchers to draw conclusions about tumour origins, molecular biology and subtyping. In this series of publications—and in order to understand the molecular biology underlying cancer—the TCGA consortium cross-checked general molecular aspects in all tumour types. To this end, they exhaustively studied, in the more than 10,000 samples stored in their repository, the process of alternative splicing (*Kahles et al., 2018*) and they identified the specific variants (*Huang et al., 2018*) and driver genes (*Bailey et al., 2018*) that generate greater predisposition to tumour development. They also analysed the effect of enhancer activation on different tumour types (*Chen et al., 2018a*) and the effect of aneuploidy (*Taylor et al., 2018*). They also catalogued the variants of the 10 pathways that are most frequently altered in most tumours (*Sanchez-Vega et al., 2018*), in addition to alterations in genes related to the ubiquitin (*Ge et al., 2018*), DNA damage repair (*Knijnenburg et al., 2018*) and the MYC pathways (*Schaub et al., 2018*).

The consortium also features a strong technology component; they published an integrated pancancerous clinical data resource from TCGA with the aim of driving the analysis of high-quality survival results (*Liu et al., 2018a*). In addition, they conducted studies where they used ML and deep learning algorithms to identify stemness features in tumour cells (*Malta et al., 2018*), the prediction of Ras pathway activation (*Way et al., 2018*) and the detection of tumour infiltrating lymphocytes using images (*Saltz et al., 2018*). In *Ellrott et al. (2018)* they described the Multi-Center Mutation Calling project, which aims to generate a complete encyclopaedia of somatic mutations from TCGA data that allows a robust analysis for different tumour types. They performed different studies that proposed new classifications among tumours. For example, they identified new immune tumour types across the 33 types of cancer that differ by somatic aberrations,

Table 2 **Enumeration of the different cohorts presented by the TCGA repository, classified according tothe tissue of origin of the tumour.** In addition, the original paper published by the TCGA consortium is cited.

| Cancer type | Acronym | Tissue | Citation |
|---|---|---|---|
| Breast Ductal/Lobular Carcinoma | BRCA | Breast | (*The Cancer Genome Atlas Network, 2012b*; *Ciriello et al., 2015*) |
| Glioblastoma Multiforme | GBM | Central Nervous System | (*The Cancer Genome Atlas Research Network, 2008*; *Verhaak et al., 2010*; *Noushmehr et al., 2010*; *Brennan et al., 2013*; *The Cancer Genome Atlas Research Network, 2015*, *Ceccarelli et al., 2016*) |
| Lower Grade Glioma | LGG | Central Nervous System | (*The Cancer Genome Atlas Research Network, 2015*) |
| Adrenocortical Carcinoma | ACC | Endocrine | (*Zheng et al., 2016*) |
| Papillary Thyroid Carcionma | THCA | Endocrine | (*Agrawal et al., 2014*) |
| Paraganglioma & Pheochromocytoma | PCPG | Endocrine | (*Fishbein et al., 2017*) |
| Cholangiocarcinoma | CHOL | Gastrointestinal | (*Farshidfar et al., 2017*) |
| Colon Adenocarcinoma | COAD | Gastrointestinal | (*The Cancer Genome Atlas Network, 2012a*) |
| Rectal Adenocarcinoma | READ | Gastrointestinal | (*The Cancer Genome Atlas Network, 2012a*) |
| Esophageal Cancer | ESCA | Gastrointestinal | (*The Cancer Genome Atlas Research Network, 2017c*) |
| Liver Hepatocellular Carcionoma | LIHC | Gastrointestinal | (*Ally et al., 2017*) |
| Pancreatic Ductal Adenocarcinoma | PAAD | Gastrointestinal | (*Raphael et al., 2017*) |
| Stomach Cancer | STAD | Gastrointestinal | (*The Cancer Genome Atlas Research Network, 2014a*) |
| Cervical Cancer | CESC | Gynecologic | (*The Cancer Genome Atlas Research Network, 2017b*) |
| Ovarian Serous Cystadenocarcinoma | OV | Gynecologic | (*The Cancer Genome Atlas Research Network, 2011*) |
| Uterine Carcinosarcoma | UCS | Gynecologic | (*Cherniack et al., 2017*) |
| Uterine Corpus Endometrial Carcinoma | UCEC | Gynecologic | (*Levine, 2013*) |
| Head and Neck Squamous Cell Carcinoma | HNSC | Head and Neck | (*The Cancer Genome Atlas Network, 2015*) |
| Uveal Melanoma | UVM | Head and Neck | (*Robertson et al., 2017b*) |
| Acute Myeloid Leukemia | AML | Hematologic | (*The Cancer Genome Atlas Research Network, 2013*) |
| Thymoma | THYM | Hematologic | (*Radovich et al., 2018*) |
| Cutaneous Melanoma | SKCM | Skin | (*Akbani et al., 2015*) |
| Sarcoma | SARC | Soft Tissue | (*The Cancer Genome Atlas Research Network, 2017a*) |
| Lung Adenocarcinoma | LUAD | Thoracic | (*The Cancer Genome Atlas Research Network, 2014c*; *Campbell et al., 2016*) |
| Lung Squamous Cell Carcinoma | LUSC | Thoracic | (*The Cancer Genome Atlas Research Network, 2012c*; *Campbell et al., 2016*) |
| Mesothelioma | MESO | Thoracic | (*Hmeljak et al., 2018*) |
| Chromophobe Renal Cell Carcinoma | KICH | Urologic | (*Davis et al., 2014*) |
| Clear Cell Kidney Carcinoma | KIRC | Urologic | (*The Cancer Genome Atlas Research Network, 2013*) |

(Continued)

Table 2 (continued)

| Cancer type | Acronym | Tissue | Citation |
|---|---|---|---|
| Papillary Kidney Carcinoma | KIRP | Urologic | (*The Cancer Genome Atlas Research Network, 2016*) |
| Prostate Adenocarcinoma | PRAD | Urologic | (*Abeshouse et al., 2015*) |
| Testicular Germ Cell Cancer | TGCT | Urologic | (*Shen et al., 2018*) |
| Urothelial Bladder Carcinoma | BLCA | Urologic | (*The Cancer Genome Atlas Research Network, 2014b*; *Robertson et al., 2017a*) |
| Diffuse Large B-cell Lymphoma | DLBC | Lymphatic tissue | |

microenvironment and survival (*Thorsson et al., 2018*). Furthermore, they classified tumours based on metabolic expression and subsequently proposed different subtypes that were not previously contemplated (*Peng et al., 2018*). In addition, they carried out exhaustive studies on groupings of tumours according to their origin in order to elucidate new therapeutic targets that might be useful for gastrointestinal adenocarcinomas (*Liu et al., 2018b*), gynaecological tumours and breast cancers (*Berger et al., 2018*) and squamous carcinomas (*Campbell et al., 2018*). In these papers, they performed clustering techniques to subtype patients into new groups for treatment or diagnosis. Finally, they studied tumours by cell (*Hoadley et al., 2018*) and tissue (*Hoadley et al., 2014*) of origins.

There are many results reported by TCGA that have had a very important impact on oncology. The results obtained by the consortium show a roadmap to follow and open countless avenues in this field where new research groups, until now unable to carry out their research globally, will be able to report important results in this field.

## MACHINE LEARNING AS A SOURCE OF NEW KNOWLEDGE

ML is the process by which machines acquire the ability to learn an action or behaviour. These processes are defined by different algorithms that enable the computer to learn a behaviour (classify, identify, etc.) and extract patterns from the data. These patterns are ultimately inherent knowledge of the problem to be analysed that the algorithms can extract and learn to identify. Subsequently, given a new case, these techniques can evaluate and predict to which group it is most likely to belong, always in accordance with prior knowledge. It is therefore critical that such techniques are applied with careful experimental design (*Fernandez-Lozano et al., 2016*) and that the data are as accurate as possible to define the problem. These techniques will learn and maximally exploit the intrinsic knowledge that underlies the data.

Depending on how this information extraction process is performed, we can speak of different approaches: supervised and unsupervised learning. Although in practice there are more types of learning, we will only focus on these two, mainly because these approaches have been the most widely used in biomedicine.

## The TCGA consortium and ML

The TCGA consortium has analysed cancer based on ML algorithms, sometimes with novel approaches specifically designed for the TCGA data. TCGA researchers recently presented a new ML that can predict the differentiation of certain tumour tissues (*Malta et al., 2018*). In this case, using data from non-differentiated stem cells and their differentiated progenitors (data obtained from public repositories), they constructed two classes of indicators that reflect epigenetic and genetic expression traits of the cells. Once they constructed these descriptors, they used a variant of one-class logistic regression to classify the different TCGA samples according to their degree of differentiation, a crucial characteristic for the development of the tumour and its invasive potential.

Another study (*Way et al., 2018*) used three types of omics platforms (expression, copy number and mutation) to predict the activation of the Ras pathway, which has been widely studied throughout oncological research. This model predicted whether this pathway was activated using RNAseq expression data. From the copy number and mutation data, the researchers were able to label the patients to design a supervised learning problem. Therefore, it was observed that certain omic patterns could be predicted from different omic data. This enables the prediction of a significant number of characteristics in tumours. This approach was also performed in another study by modifying the target in order to predict the activation of the TP53 pathway (*Knijnenburg et al., 2018*).

In other study, deep learning based on convolutional neural networks (CNN) mapped tumour infiltrating lymphocytes (TIL) based on haematoxylin and eosin (H&E) images. In this case, 13 types of TCGA tumours exhibited almost perfect performance when differentiating these cell types (*Saltz et al., 2018*). In this work, the TCGA consortium highlighted the importance of the images it stores and questions their relatively limited use by different researchers in comparison with omics platforms. The images in the TCGA repository will be discussed in the following sections.

## Popular ML models with TCGA data

The TCGA consortium has relied on both supervised and unsupervised ML techniques to extract new knowledge from its data. However, it is interesting to identify the work developed by other researchers who have used TCGA data. The approaches taken and the results obtained from the various published works will be discussed below.

Figure 2 shows the proportion of published papers according to the type of algorithm and the type of omic data used. We reviewed more than 100 papers that have used ML approaches with TCGA data. For each one, we identified: the algorithm and data type/s. Almost half of the identified works used variants of the support vector machine (SVM) or tree-based algorithms, followed by linear models as can be seen in Fig. 2A. On the other hand, Fig. 2B clearly shows that gene expression data is most abundant data type used in ML research. Other data types such as images, methylation, miRNA and copy number have been used, but majority in a combination with gene expression data.

The findings of this review highlight the low variability of reported research and analytical methods. It is true that the mostly used algorithms, Random Forest (RF) and SVM, as well as the types of omic data (expression) have reported promising results in

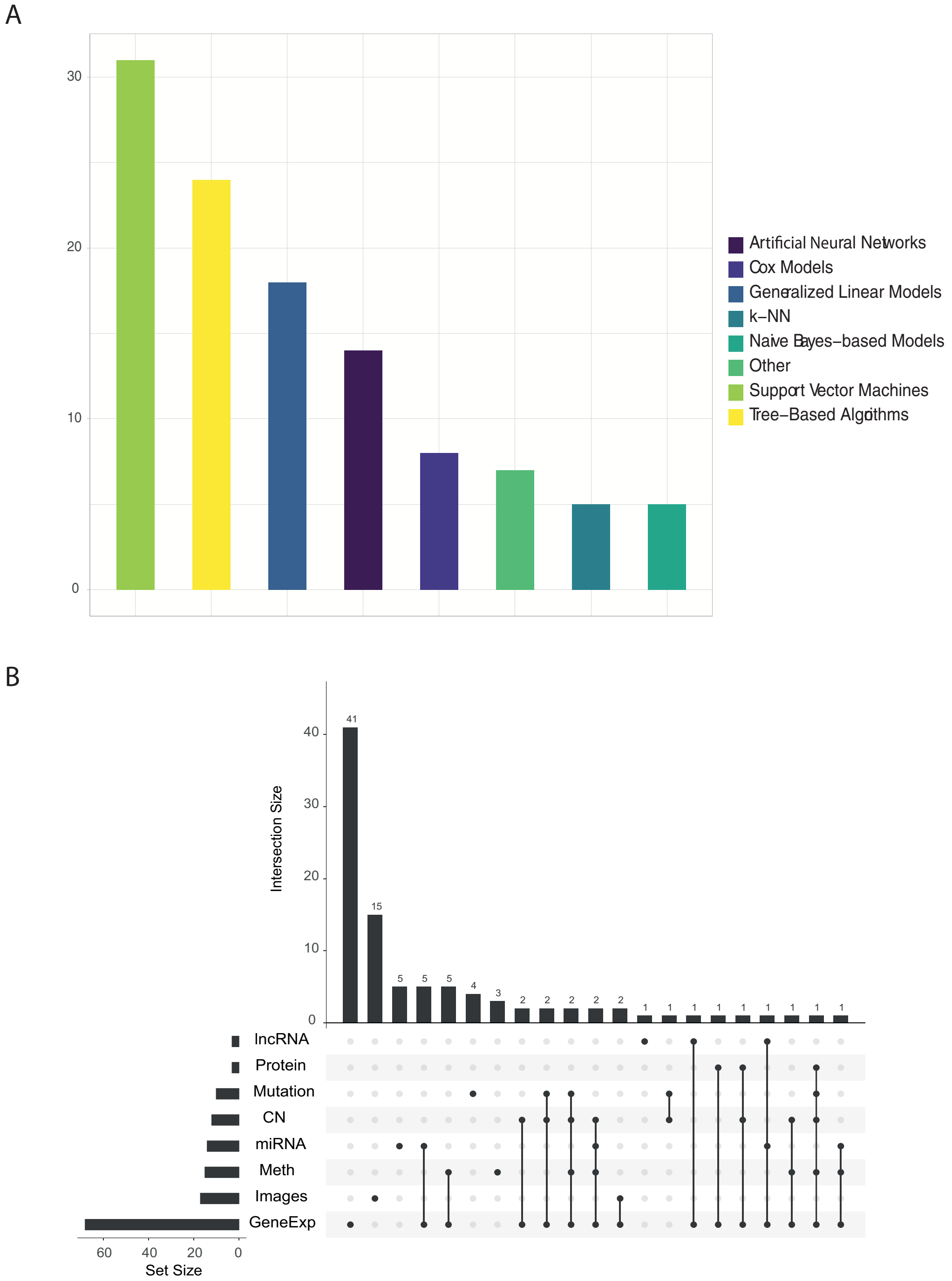


**Figure 2 (A) Number of papers that used each type of algorithm, and (B) relations between omics data used in each work.** Full-size DOI: 10.7717/peerj-cs.584/fig-2

the biomedical field during the last years. We believe that the low variability in the approaches established by researchers is mainly due to two reasons. First, the intrinsic characteristics of biomedical data, and specifically the omic data, present a much greater number of characteristics than observations. This fact is generally not idyllic for the training of ML algorithms. In this sense, the use of algorithms is mainly determined by the use of which type of omic data is being analyzed. In the context mentioned above, certain algorithms are able to handle some characteristics of the data better than others. For example, neural networks are more sensitive to the lack of observations than in this case RF, SVM or linear models. Given that the vast majority of works identified have used expression data, it is logical to observe a high density of works that have used RF or SVM type algorithms. On the other hand, those works that have used image data are more likely to use neuron networks. Secondly, there is no doubt that the possibilities in the exploitation of these data by ML algorithms are yet to be discovered. A break in the arrival of ML-based applications in the field of biomedicine has been detected. This is partly due to the complexity of the omic data, and the need for specialists in this field for its modelling and good practical use. Possible applications that could revolutionize the field of biomedicine could be the use of NLP (Natural Language Processing) algorithms for the analysis of Whole Genome Sequencing (WGS) data.

After all, if there is something to highlight in the results observed in the Fig. 2B is the trend towards more and more work integrating different omic data. Even so, this trend is not reflected in Fig. 2A, in which a variety of algorithms and/or new known and standardized methodologies that can solve this problem are not observed. This is the great challenge in the coming years presented by biomedicine, which could generate very useful predictions for tackling complex diseases, such as cancer.

### A general perspective of unsupervised learning with TCGA data

In oncology, clustering methods are extremely useful for subtyping or reclassification of patients in a particular cohort. Over the years, the classic clustering methods have been most widely used, including partitioning clustering or hierarchical clustering. Even today, they are widely used with their respective variations. For example, the TCGA consortium has used them to subtype different tumours (see publications in Table 2). The problem with these algorithms is that they can only model a single set of data and the concatenation of different types of data does not perform adequately. The complexity of the tumour is manifested at distinct biological levels; hence, methods that can accept different types of data are preferable. Thus, researchers developed a new integrative clustering method based on a joint latent variable model (iCluster) (*Shen, Olshen & Ladanyi, 2009*) and used it with TCGA data (*Shen et al., 2012*). *iCluster* fits a regularized latent variable model based clustering that generates an integrated cluster assigment based on joint inference across data types. In addition, the implementation in several programming languages is very intuitive. On other hand, an extended version (*iClusterPlus*) was also developed (*Mo et al., 2013*). One of the most important works using this method was (*Curtis et al., 2012*), identifying 12 different breast tumour subtypes.

In addition to beforementioned works, there are a huge examples of iCluster use with TCGA. For instance, in *Xie et al. (2019)* an integrative analysis was carried out with iCluster through RNAseq and proteomics data to analyse the OV subtype. The results showed two clusters with different survival rates; the method identified 18 mRNAs and 38 proteins as distinct molecules among subtypes. Another study proposed a modified iCluster model to discover key processes in the tumour collection through unsupervised integration of multiple types of molecular data and functional annotations (*Bismeijer, Canisius & Wessels, 2018*). Further, *Mo et al. (2017)* described a novel modification (iClusterBayes) capable of jointly modelling omics data of continuous and discrete data types for the identification of tumour subtypes and relevant omics characteristics. In the work of *Kim et al. (2017)*, they modified this procedure to subtype patients using sequential double regularisation. Another pathway-based variant incorporates pathway data to group patients into cancer subtypes (*Mallavarapu et al., 2019*). Additionally, in *Jean-Quartier et al. (2021)* clustered GBM patients into several age subgroups with different age-related biomarkers. Finally, a work developed in *Nguyen et al. (2017)*, named PINS, allows omics data integration and molecular patient stratification automatically.

With the above, the trend in genome research is evident. An increasing number of works are attempting to integrate the greater amount of information provided by the different omic data into their models. Due to the complexity of cancer, stratifying patients according to a single source of information is becoming obsolete. Therefore, it is vitally important to improve models that are capable of multi-omic integration, as is the case with iCluster. Moreover, there is a need of novel approaches to automated medical decision pipelines building on machine learning, information fusion and explainability (*Holzinger et al., 2021*; *Barredo Arrieta et al., 2020*).

### Medical imaging as a data source for ML algorithms

An important event occurred in 2012 during the celebration of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (*Russakovsky et al., 2015*). A deep learning model (specifically, a CNN) halved the second best error rate in the image classification task. The goal of this challenge was the detection of objects and the classification of images using a large-scale database. Furthermore, deep learning algorithms can automatically find the best subset of features that describe the nuances of images. In addition, transfer learning was borne: an attempt to reuse the representation of the learning characteristic of one problem to solve another.

Deep learning techniques are on the rise in cancer research, namely for object detection and image classification. Initiatives such as TCGA offer the possibility of training deep learning models by making a large quantity of biomedical images available for research. Specifically, TCGA provides two types of images: tissue slide and digital imaging and communications in medicine (DICOM) images. DICOM images such as X-rays or computed tomography (CT), are used to extract quantitative characteristics from the images. Algorithms are trained to identify those characteristics. Histopathological images are used for direct image processing.

As discussed in previous sections, the TCGA consortium has used deep learning methods (*Saltz et al., 2018*). Specifically, they used CNN to detect tumour-infiltrating lymphocytes (TILs) based on H&E images in 13 tumour types. They reported a local spatial structure in the TIL patterns and their correlation with overall survival. These data modify densities and spatial structure among tumour types, immune subtypes and molecular tumour subtypes. Spatial infiltration of lymphocytes might reflect particular aberration states of tumour cells.

Based on these findings, several studies have used this and other repositories to create their own models. It is important to distinguish among data types. On the one hand, there are works that have used radiological images for the classification of stages of gliomas (*Park et al., 2019*). In this work, they did not use the radiological images directly; rather, they extracted 250 characteristics from them to train their models, obtaining an area under the receiver operating characteristic curve (AUROC) of 72%. Notably, this model, which was validated with very heterogeneous cohorts such as TCGA, considerably reduced the performance. These results indicate that manual extraction of characteristics does not provide sufficient generalisation.

*Sun et al. (2018b)* utilised contrast-enhanced CT images and RNASeq data to assess CD8 cell infiltration in tumour biopsies. They first extracted features from both types of data to ultimately keep eight features and train an elastic-net regularised regression method. They used this signature to predict the response to anti-programmed cell death protein 1 (PD1) or anti-programmed death-ligand 1 (PDL1) treatments. Magnetic resonance imaging (MRI) was used in to predict the status of MGMT, a promoter of methylation that has been related to better outcomes on GBM patients integrated with expression data (accuracy of 73%; (*Kanas et al., 2017*)).

*Fischer et al. (2018)* reported a new method for histopathological image analysis—sparse coding—using a dictionary optimised for biomedical images. They stated that they generally obtained better performance rates compared to transfer learning. In *Yu et al. (2016)*, they predicted the prognosis of non-small cell lung tumours. Using the CellProfiler software, they extracted 9,879 quantitative characteristics and trained different algorithms, such as SVM or random forest. Finally, with a variant of the SVM algorithm, they achieved an AUROC of 81%. Besides, they developed a low-complexity method for classification and disease grading in histopathological images. This method—discriminative feature-oriented dictionary learning (DFDL)—learns from specific class dictionaries in such a way that under a dispersion restriction, the learned dictionaries allows it to represent a new image in a simplified way. However, it is unable to represent samples from other classes. *Coudray et al. (2018)* used histopathology images of lung cancers to classify squamous cell carcinomas, adenocarcinomas and normal samples with a 97% of AUROC. In the work of *Cheerla & Gevaert (2019)*, they were able to extract information from several datasets and obtain a model capable to predict patient prognosis. *Ertosun & Rubin (2015)* subtyped gliomas with CNN algorithms by using raw images for this task; there was more than 90% accuracy for glioma classification and almost 80% for glioma grade identification. *Rendleman et al. (2019)* used a CNN to evaluate distinct histological tumour growth patterns such as solid, micropapillary, acinar and cribriform (84% accuracy). An important work was developed in *Janowczyk et al. (2019)*. They developed an unsupervised encoder to compress four data modalities, including whole slide images (WSIs), into a single feature vector for each patient. The model was trained with TCGA data and predict single cancer overall survival, achieving a C-index of 0.78 overall.

It is important to highlight the need to pre-process the histopathological images before their analysis. This step is crucial to achieve great performances in the models. The images housed in TCGA are not homogeneous in size, shape and brightness. Therefore, it is necessary to use a pre-processing stage in order to standardise all the images before the analysis. Open source tools as HistoQC (*Janowczyk et al., 2019*) are relevant in the extraction of knowledge and the good use of images in research.

## Biological questions solved by ML algorithms

In addition to all the existing omics data in TCGA, the inclusion of the clinical information from each patient increases the ability to generate analytical models. The dependent variable in supervised learning problems can potentially be any of the 100 clinical outcomes offered by TCGA, depending on the biological response to be answered. For classification problems, researchers have information on the anatomical division of the neoplasm, the clinical stage of the patient, TNM status, MSI status, ethnicity, age and gender, survival and/or relapse of the tumor among others. Thus, we can infer whether we can predict the anatomical division of the tumour or its clinical stage from the methylation marks of the patients (among other possibilities). For regression, we can use the initial age at diagnosis or the prognosis of the patient by means of the Karnofsky Performance Status Scale. Also, independently of clinical data, classificacion and regression models could be created to determinate other omics outcomes. For instance, sobreexpression of driver genes, methylation status or mutation types. In addition, the data storaged in TCGA repository allows any potential researcher to study survival in the cohort: it presents data on life status and the days that have elapsed between events, such as the death of the patient or other events of interest (relapse or disease-free survival).

The quantification of the number of papers published for each cancer subtype is shown in Fig. 3. As shown in this figure, the most used are those with the highest number of samples: BRCA, LUAD and OV. The great number of dimensions and observations, together with the large number of available clinical variables (pathological state, TNM classification status, drug effect, treatment response, etc.) generates an ideal data analysis environment for the use of both supervised and unsupervised ML techniques. For supervised learning problems, contingent on the dependent variable to be predicted, these problems may be regression (patient survival time, expression of a specific gene or individual age) or classification (classification of patients according to some driver gene status, disease or metastasis stages, etc.) problems. In terms of unsupervised learning problems, most work focuses on finding new subtypes of the disease. As for the other tumours, there is a significant decrease in the number of publications, mainly due to the number of samples collected. This fact is due to the intrinsic functioning of the ML algorithms, which, because they work on the basis of examples, are able to generalise more as the number of observations in their training phase increases. We can observe in the Fig. 3 also how there are several works that use different cohorts in the same analysis. After reviewing these papers, two trends have been observed in this type of article. Firstly, there are those that train models to predict cross-sectional and/or basic conditions of tumours. For example, in *Fischer et al. (2018)* they predicts MSI status from

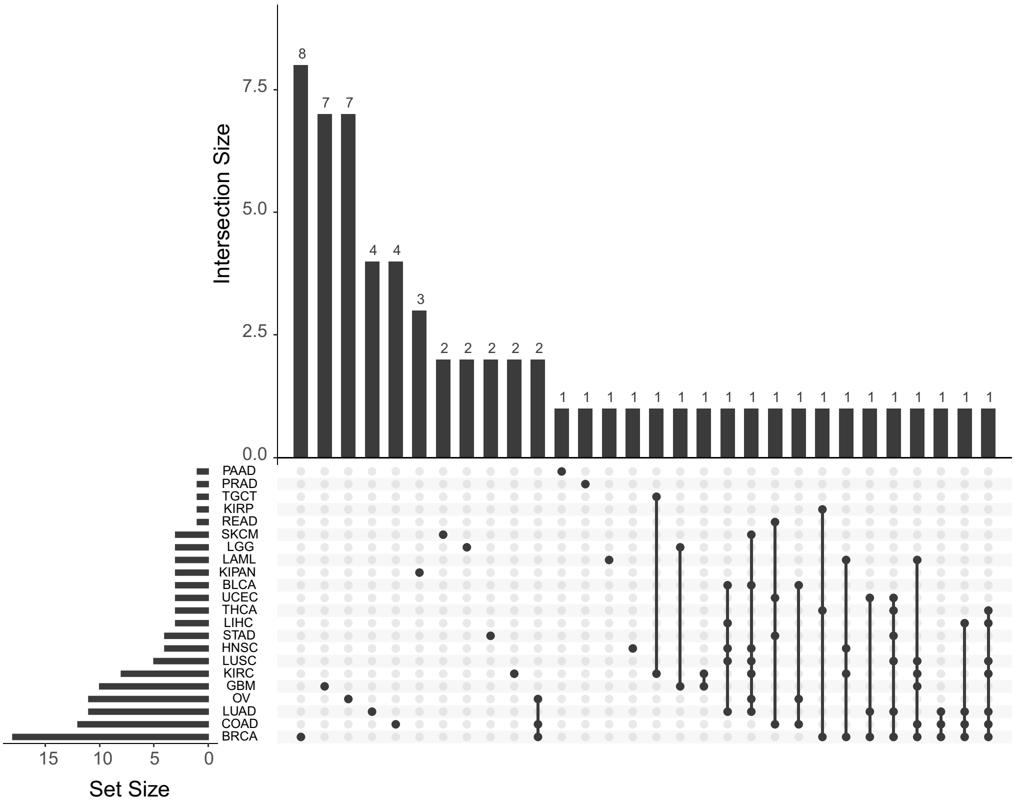
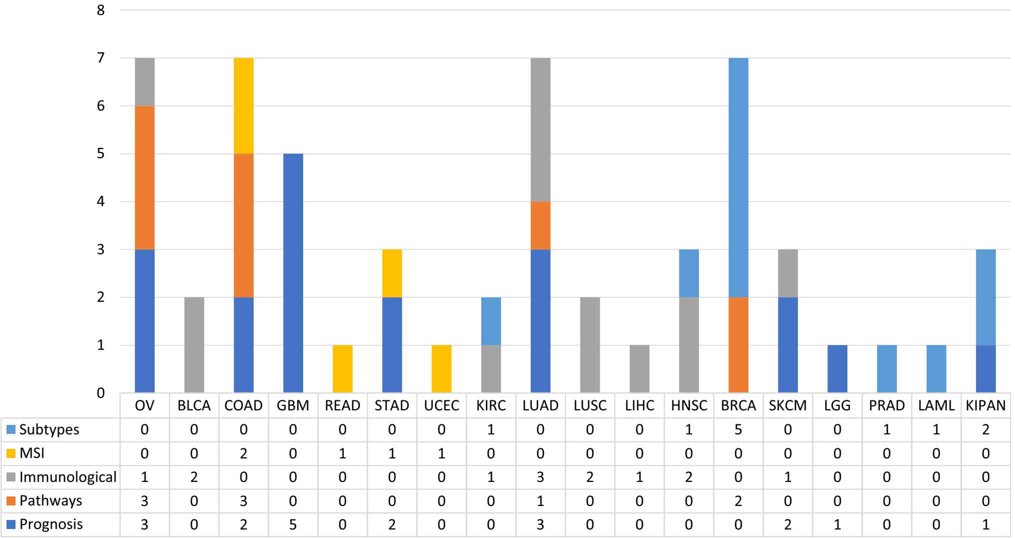


**Figure 3** **Number of papers published with each of the TCGA cohorts.** Upset plot showing the number of works published with each tumor type and their combinations.
Full-size DOI: 10.7717/peerj-cs.584/fig-3

histopathological images. In this case, the different TCGA cohorts are treated together for the training of the models. On the other hand, other works have been identified in which the cohorts are used independently. These works are mainly based on model improvements or development of new technologies that are then tested with each cohort. This is the example of *Chen et al. (2018b)*, where they develop a new model of autoencoders for the search of new genetic signatures. This model is later validated in each of the available TCGA cohorts. Another example is *Cheerla & Gevaert (2017)*, where they obtain a model that recommends the type of treatment from miRNA expression data. This recommendation is validated in the different TCGA cohorts. Therefore, there are many approaches that can be used by researchers to use this type of data.

In this review, we classify the identified works on five major groups according biological problem solved. Although there are more than 100 variables in the TCGA clinical database, there is very little variability observed in the type of analysed problems.

In order to observe the distribution of publications according to this type of classification along the different types of tumours, pay attention to the Fig. 4. Figure 4 shows the distribution of the published papers according to the different types of tumors and the type of biological problem. The different biological problems show a different distribution according to tumor type. It can be seen how prognosis prediction is more



| | OV | BLCA | COAD | GBM | READ | STAD | UCEC | KIRC | LUAD | LUSC | LIHC | HNSC | BRCA | SKCM | LGG | PRAD | LAML | KIPAN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ■ Subtypes | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 5 | 0 | 0 | 1 | 1 | 2 |
| ■ MSI | 0 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ■ Immunological | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 2 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| ■ Pathways | 3 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| ■ Prognosis | 3 | 0 | 2 | 5 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 1 |

**Figure 4 The proportion of published works with ML techniques according to the type of biological problem.** Full-size DOI: 10.7717/peerj-cs.584/fig-4

common in GBM cohorts. In this case GBM is a type of tumour with high mortality rates, so it is a cohort where there are numerous events with which robust ML models can be created. Following GBM, OV and LUAD cohorts were the most used. Furthermore, it is observed how this type of problem is addressed in different cohorts. This is not the case for MSI prediction, as few tumours are defined by MSI status. The most common ones in this case are COAD, READ, STAD and UCEC. Paying attention to the prediction of subtypes, we see that the BRCA cohort is the most used. Regarding the immunological phenotype, the works have used cohorts mainly of solid tumours, which are the ones that present the best response to treatments with immunological therapies. Finally, few tumours have been addressed in the prediction of pathways. The works identified used the OV, COAD, LUAD and BRCA cohorts. The following sections are a review of the works according to the five classes identified.

### Prognosis prediction

The prognosis in the different types of cancer varies greatly due to their heterogeneity, their environment and their unique behaviour in each patient. It is therefore crucial to be able to predict the events that will develop in the patient and have a direct effect on the prognosis of the cancer. These events can be deaths, recurrence and/or relapse events, metastases or the classification of patients into specific stages. Numerous studies have been identified that have addressed this field of study with ML-based analyses.

Within this category, many of the papers identified have aimed to predict events related to patient survival time. Furthermore, it has been observed that expression data are the most used in this type of problem, due to their better performance in predictions, together with methylation data (*Stephen & Lewis, 2013*). In *Wong, Rostomily & Wong (2019)* they use them as input from a deep learning network, while in *Fatai & Gamieldien (2018)* they use the SVM algorithm. In both problems they obtained gene signatures that were highly correlated with the survival events of the patients. Other works have addressed this type of problem by integrating expression data with other data sets. For example, in *Yasser et al. (2018)*, using FS techniques, they obtained subgroups of features from sets of ANC, methylation and expression. In *Zhang et al. (2016)*, they add a layer of complexity, adding to the integration of miRNA data by means of multiple kernel and FS techniques. This technique was also used in *Srivastava et al. (2013)* for the integration of expression and miRNA data. On the other hand, one paper has used only lncRNA data capable of predicting survival events over 19 months (*Cheng, 2018*). Works were also identified that have addressed this problem from histopathological images. In these cases, they extract characteristics from the images in order to train different types of ML algorithms and be able to predict survival times and/or events (*Ing et al., 2017*; *Yu et al., 2016*; *Powell et al., 2017*).

In addition to survival events and times, there are other events that are interesting to predict for clinical practice. In this case, the events of tumour recurrence, which are when the tumour is detected again after a treatment process. Knowing, therefore, the probabilities of a cancer relapse in a given patient is interesting for clinicians. Using the ML approach, this problem has been addressed mainly with transcriptomical expression data. For example, in *Wang et al. (2018)*, from miRNA data, lncRNA and mRNA identified 36 features capable of classifying with 91% accuracy whether a tumour will recur or not. In *Zhou et al. (2018)*, *Sun et al. (2018a)*, *Xu et al. (2017)*, similar performances were obtained with RNASeq data, while in *Wei (2018)* from RNASeq data predict metastasis processes. On the other hand, in *Feng et al. (2018)* they predicted recurrence based on data from the tumour microenvironment.

Usually the different types of tumours are classified in different stages which correlate with different prognoses. Therefore, this has been another problem addressed by the researchers, in which the ML has been able to offer a solution. Again, the RNASeq data were the most used to address this problem. In *Fan et al. (2018)*, they obtained a signature of 12 genes capable of distinguishing patients with lung cancer with different risks, while in *Chen et al. (2017)* they identified pathways of interest capable of classifying the different stages of lung adenocarcinoma. For example, in *Yang, Xu & Zeng (2018)*, they used only the features corresponding to lncRNA, obtaining a signature of six lncRNA capable of classifying patients with melanoma according to their stages.

### Immunological phenotype prediction

Currently, one of the most successful and promising therapies against cancer are drugs that act against immune checkpoint inhibitors (ICI). These drugs block the proteins produced by certain immune cells to prevent immune responses from becoming too strong. The activation of these checkpoints can cause the cells of our immune system not to be able to kill the cancer cells. The treatment of most types of tumours is helped by this type of therapy, although there are some that do not respond in the same way. This is the case of HGSOC tumours. In *Dai et al. (2018)*, they analysed genomic data from HGSOC patients to predict their immune phenotype of the tumour microenvironment. After a comparison with the analysis of other solid tumours, such as BLCA, SKCM, KIRC, LUSC and LUAD, they identified ten dominant factors that determine the immunogenicity of HGSOCs. Using the ML they were able to classify tumours with high and low cytolytic activity, noting also that mutations in BRCA1 may be a good predictive biomarker for guiding ICI therapies of HGSOC patients.

Moreover, they developed and independently validated an eight-feature signature based on CD8 cell radiomic imaging for the response to (PD)-1 and (PD-L1). This imaging predictor provides a promising way to predict the immunologic phenotype of tumors and infer clinical outcomes for cancer patients who had been treated with anti-PD-1 and PD-L1.

### Pathways prediction

Some of the genetic drivers specific to each tumour are well known, as well as certain pathways that influence the process of tumour development. Although the identification of status is a complex issue, it holds a great deal of information in the diagnosis and treatment of patients. This is why researchers have addressed this problem using ML techniques. After the review carried out in this work, works have been detected that were able to model this problem. Most of them are based on RNASeq data, with which they infer the status of different cancer driver pathways (*Rykunov et al., 2016*), damaged pathways (*Klein, Stern & Zhao, 2017*) and level of apoptosis (*Salvucci et al., 2017*). In *Chen et al. (2012)*, RNASeq data and copy number data are used to detect pathways capable of differentiating expression patterns between different phenotypes. In the case of *Ou-Yang et al. (2017)*, they developed a cross-platform method for the identification of new molecular pathways related to tumour types.

### MSI status prediction

Microsatellite instability is the mutation predisposition of certain tumours due to defects in the DNA mismatch repair machinery. It is of great importance to identify MSI status in certain tumours as it is a great predictor and marker for diagnosis and treatment. In this review two papers were identified that have addressed this problem with MSI techniques. The first of these, called *Wang & Liang (2018)*, classified the different MSI subtypes based on mutational annotation data. They used an SVM algorithm and obtained a total accuracy of 0.91 for the COAD, READ, STAD and UCEC cohorts. They used a total of 22 features for the classification, such as the count of SNPs, indels, total mutations, missence mutations or the ratio between mutations and SNPs. On the other hand, in *Chen et al. (2018c)* they made a classification from the expression data. Using ML algorithms and FS techniques they obtained a classifier capable of discerning the different subtypes.

### Subtypes prediction

Finally, another problem that has been addressed by researchers and where ML techniques can contribute significantly is the prediction of the different subtypes of the disease. It is interesting to recognise which are the different omic data sets that hold enough information to build a classification system robust enough to obtain the appropriate yields. As usual, RNASeq was the technique par excellence from which the data were obtained to train the models (*Yang et al., 2014*; *Graudenzi et al., 2017*; *Gao et al., 2017*). In addition, the expression data were combined with other sets such as miRNA (*Wilop et al., 2016*; *Nair et al., 2015*), methylation (*List et al., 2014*) or miRNA and methylation (*Nguyen et al., 2017*).

In addition to expression data, two papers have used exclusively image data to classify subtypes of the disease. Firstly from MRI images (*Sutton et al., 2017*) and with qCT-TA data (*Kocak et al., 2018*). Other work, for example, used mutation data (*Vural, Wang & Guda, 2016*) and miRNA data (*Muhamed Ali et al., 2018*).

It is logical to think that the ML algorithms now attempt to analyse the most studied problems to determine whether they can reach the same conclusions as conventional statistical approximations. In general, ML approximations analyse the importance of each of the variables in the dataset without making any a priori assumptions, so the generalisation of the model does not have to be based on inherent biological knowledge of the data. Although there are ML approximations that base the selection of genes from each data platform to certain pathways of interest (*Seoane et al., 2013*), this field is still open field for new approximations.

One study observed that the ML algorithms reached similar conclusions and also provided a certain degree of diversity in the results (*Liñares Blanco et al., 2019*). This outcome aids the examination of new omics variables that might be of interest to study the development of cancer. Cancer is a multifactorial and complex disease, so it makes sense that the analysis should consider the differences that characterise the patients as a whole and not individually.

### A deeper examination of the BRCA cohort

The TCGA consortium jointly analysed genomic DNA copy number arrays, DNA methylation, exome sequencing, mRNA arrays, miRNA sequencing and reverse-phase protein arrays (*The Cancer Genome Atlas Network, 2012b*). In this study, they demonstrated the existence of four main classes of breast cancer by combining data from five platforms; there was great heterogeneity. Mutations in only three genes (TP53, PIK3CA and GATA3) occurred in more than 10% of all the samples. In addition, they identified two new subgroups defined by protein expression—produced primarily by the tumour microenvironment. Besides, the comparison of basal-type breast tumours with high-grade serous ovarian tumours showed a myriad of molecular similarities, a finding that indicates a related aetiology and similar therapeutic opportunities.

In one study (*Ciriello et al., 2015*), the authors discovered that invasive lobular carcinoma (ILC) is a clinically and molecularly distinct disease. In this case, patients with ILC show CDH1 and PTEN loss, AKT activation and mutations in TBX3 and FOXA1. The proliferation and expression of genes related to the immune system defined three ILC subtypes.

The findings made by TCGA are leading the way in the search for new treatment and diagnostic opportunities for patients, in this case, with breast cancer. Although the work of the TCGA has been exhaustive, the possibilities offered by giving free access to its data are enormous. For this reason, many researchers have taken these data as a reference and have reported results of great interest to the community.

We identified several publications that utilised ML to analyse TCGA BRCA data. There are published works using miRNA data (*Sherafatian, 2018*), methylation data (*Hao et al., 2017*), expression data (*Wen, Li & Chang, 2018*), integrative analysis of expression and methylation data (*Cappelli, Felici & Weitschek, 2018*) and even expression data from isomiRs (*Liao et al., 2018*). These works achieved prominent outcomes, notably the ability to infer that the problems of classification for diagnosis (healthy or disease patients) are problems that the ML algorithms solve quite easily, even with different types of data.

Several papers have been published to address this patient stratification. For example, to classify the subtypes of PR, ER and HER2 with miRNA data (*Sherafatian, 2018*; *Liao et al., 2018*), the status of the basal subtype through the analysis of images with deep learning algorithms (*Chidester, Do & Ma, 2018*) and the different subtypes of BRCA by the expression of molecular pathways (*Graudenzi et al., 2017*), mutation data (*Vural, Wang & Guda, 2016*) or even the integration of expression and methylation data (*List et al., 2014*). Cancer subtypes can be studied by unsupervised learning techniques and the integration of different data (expression, methylation, miRNA and CNV) (*Nguyen et al., 2017*).

Finally, other works have studied the interaction between miRNA and mRNA (*Koo, Zhang & Chaterji, 2018*; *Ghoshal et al., 2018*), the identification of altered pathways by mRNA expression data (*Klein, Stern & Zhao, 2017*) or by integrating expression and mutation data (*Rykunov et al., 2016*), the response to drugs in different cell lines (*Daemen et al., 2013*; *Geeleher et al., 2017*) and the identification of variants by means of genomic data (*Dong et al., 2016*) and by means of images with artificial vision techniques (*Sutton et al., 2017*).

## CONCLUSIONS

Many studies on cancer have been performed in recent years with ML that uses molecular data. These data have mainly included diagnostic studies, prognosis or patient stratification. More recently, there have been promising results in response to drugs or genetic interactions. In this review, we investigated and identified those relevant works that have used TCGA data through algorithms or pipelines of analysis based on ML.

ML techniques can extract the underlying knowledge from a set of data, so it is relevant to understand the appropriateness of the data. In other words, these techniques must be used with certain precautions. Indeed, researchers should be aware that the conclusions they obtain may be biased due to poor data selection or analytical methodology. Among the different learning techniques, supervised learning has analysed the most problems using TCGA data. This endeavour has emphasised the use of genetic expression data through different variants of the SVM algorithm. There are still infinite opportunities and possibilities for the exploitation of TCGA data with ML. ML techniques can reach conclusions that are similar to conventional approaches and also to obtain a degree of variability that is extremely useful when searching for novel predictors.

It is clear that we are still at an early stage in the analysis of this pathology and it is necessary to develop and use more complex algorithms. For example, the use of kernel-based models can integrate different datasets in the same process. The integration of data in the analysis of complex and multifactorial diseases continues to be a challenge for which it is necessary to invest even more time and money in finding better algorithms. As discussed above, the quantity of existing data will not stop growing and all derive from the same biological sample. Thus, it is expected that the connection between omics platforms can improve the performance of the models. It is still necessary to take a step forward in the development of multidimensional ML models for cancer research.

Complex problems, such as the prediction of different cell statuses (methylation, apoptosis or mutation), are already being tackled with promising results. We and others hope that the links between biological information extracted from the same patient will be further explored in order to elucidate the origin of the disease by ML techniques. Currently, the focus is on certain types and subtypes of cancer (e.g., BRCA, LUAD or OV), usually due to the number of people afflicted with it and the importance attributed by society. It is also necessary to increase investment in the generation of data that is related to relatively minor or especially aggressive cancer types in order to provide the algorithms with sufficient information in their learning phase and to avoid biases in their learning.

In this work, we exhaustively reviewed studies that have used ML techniques for the analysis of different types of cancer using TCGA data. In our opinion, the era of individual analysis has passed and we are entering the era of data integration studies—at the clinical-genomic level as well as medical imaging or evolution analysis by means of time series. We are working on the development of complex data integration algorithms in different fields, one of which is artificial intelligence. There are currently ML models that are demonstrating great effectiveness and are gaining followers. These methods include the aforementioned deep learning techniques, but research is required to render the results understandable and explain why a certain prediction is made, especially from a clinical point of view. The great challenge of the integration techniques is the incessant increase in the number of dimensions and the heterogeneity of the data sets generated from the same patient/biological process (*Kristensen et al., 2014*).

Finally, we hope that this review will serve as a starting point for researchers in bioinformatics and computer science who are interested in studying cancer, as well as those researchers who are more focused on the use of ML techniques to know the potential of their algorithms with TCGA data. More research and the development of new algorithms are required to overcome the disease.

## ABBREVIATIONS

| | |
|---|---|
| **ML** | Machine Learning |
| **TCGA** | The Cancer Genome Atlas |
| **MSI** | Microsatellite Instability |
| **BRCA** | Breast Cancer Adenocarcinoma |
| **GBM** | Glioblastoma Multiforme |
| **SNP** | single nucleotide polymorphisms |
| **LUSC** | Lung squamous cell carcinoma |
| **OV** | Ovarian Cancer |

| | |
|---|---|
| **SVM** | Support Vector Machines |
| **RF** | Random Forest |
| **WGS** | Whole Genome Sequencing |
| **CNN** | Convolutional Neural Networks |
| **TIL** | Tumour-infiltrating lymphocytes |
| **MRI** | Magnetic resonance imaging |

## ACKNOWLEDGEMENTS

We thank Dr. Jose A. Seoane for comments during the preparation of this review.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

This work was supported by 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.” and the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431D 2017/16), the “Galician Network for Colorectal Cancer Research” (Ref. ED431D 2017/23) and Competitive Reference Groups (Ref. ED431C 2018/49). CITIC, as Research Center accredited by Galician University System, is funded by “Consellería de Cultura, Educación e Universidades from Xunta de Galicia”, supported in an 80% through ERDF Funds, ERDF Operational Programme Galicia 2014–2020, and the remaining 20% by “Secretaría Xeral de Universidades” (Grant ED431G 2019/01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:
Carlos III Health Institute: PI17/01826.
European Regional Development Funds (FEDER).
General Directorate of Culture, Education and University Management of Xunta de Galicia: ED431D 2017/16.
Galician Network for Colorectal Cancer Research: ED431D 2017/23.
Competitive Reference Groups: ED431C 2018/49.
Consellería de Cultura, Educación e Universidades from Xunta de Galicia.
Secretaría Xeral de Universidades: ED431G 2019/01.

### Competing Interests

The authors declare that they have no competing interests.

## Author Contributions

- Jose Liñares-Blanco conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Alejandro Pazos conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Carlos Fernandez-Lozano conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

## Data Availability

The following information was supplied regarding data availability:

This research is a Literature Review; there are no raw data or code files.

## REFERENCES

**Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, Annala M, Aprikian A, Armenia J, Arora A, Auman JT, Balasundaram M, Balu S, Barbieri CE, Bauer T, Benz CC, Bergeron A, Beroukhim R, Berrios M, Bivol A, Bodenheimer T, Boice L, Bootwalla MS, Borges dos Reis R, Boutros PC, Bowen J, Bowlby R, Boyd J, Bradley RK, Breggia A, Brimo F, Bristow CA, Brooks D, Broom BM, Bryce AH, Bubley G, Burks E, Butterfield YSN, Button M, Canes D, Carlotti CG, Carlsen R, Carmel M, Carroll PR, Carter SL, Cartun R, Carver BS, Chan JM, Chang MT, Chen Y, Cherniack AD, Chevalier S, Chin L, Cho J, Chu A, Chuah E, Chudamani S, Cibulskis K, Ciriello G, Clarke A, Cooperberg MR, Corcoran NM, Costello A J, Cowan J, Crain D, Curley E, David K, Demchok J A, Demichelis F, Dhalla N, Dhir R, Doueik A, Drake B, Dvinge H, Dyakova N, Felau I, Ferguson M L, Frazer S, Freedland S, Fu Y, Gabriel S B, Gao J, Gardner J, Gastier-Foster JM, Gehlenborg N, Gerken M, Gerstein MB, Getz G, Godwin AK, Gopalan A, Graefen M, Graim K, Gribbin T, Guin R, Gupta M, Hadjipanayis A, Haider S, Hamel L, Hayes DN, Heiman DI, Hess J, Hoadley KA, Holbrook AH, Holt RA, Holway A, Hovens CM, Hoyle AP, Huang M, Hutter CM, Ittmann M, Iype L, Jefferys SR, Jones CD, Jones SJM, Juhl H, Kahles A, Kane CJ, Kasaian K, Kerger M, Khurana E, Kim J, Klein RJ, Kucherlapati R, Lacombe L, Ladanyi M, Lai PH, Laird PW, Lander ES, Latour M, Lawrence MS, Lau K, LeBien T, Lee D, Lee S, Lehmann K-V, Leraas KM, Leshchiner I, Leung R, Libertino JA, Lichtenberg TM, Lin P, Linehan WM, Ling S, Lippman S M, Liu J, Liu W, Lochovsky L, Loda M, Logothetis C, Lolla L, Longacre T, Lu Y, Luo J, Ma Y, Mahadeshwar H S, Mallery D, Mariamidze A, Marra MA, Mayo M, McCall S, McKercher G, Meng S, Mes-Masson A-M, Merino MJ, Meyerson M, Mieczkowski PA, Mills GB, Shaw KRM, Minner S, Moinzadeh A, Moore R A, Morris S, Morrison C, Mose LE, Mungall AJ, Murray BA, Myers J B, Naresh R, Nelson J, Nelson M A, Nelson P S, Newton Y, Noble M S, Noushmehr H, Nykter M, Pantazi A, Parfenov M, Park PJ, Parker J S, Paulauskis J, Penny R, Perou C M, Piché A, Pihl T, Pinto P A, Prandi D, Protopopov A, Ramirez NC, Rao A, Rathmell WK, et al. 2015.** The molecular taxonomy of primary prostate cancer. *Cell* **163(4)**:1011–1025 DOI 10.1016/j.cell.2015.10.025.

**Agrawal N, Akbani R, Aksoy BA, Ally A, Arachchi H, Asa SL, Auman JT, Balasundaram M, Balu S, Baylin SB, Behera M, Bernard B, Beroukhim R, Bishop JA, Black AD, Bodenheimer T, Boice L, Bootwalla MS, Bowen J, Bowlby R, Bristow CA, Brookens R, Brooks D, Bryant R, Buda E, Butterfield YSN, Carling T, Carlsen R, Carter SL, Carty SE, Chan TA, Chen AY, Cherniack AD, Cheung D, Chin L, Cho J, Chu A, Chuah E, Cibulskis K, Ciriello G, Clarke A, Clayman GL, Cope L, Copland J A, Covington K, Danilova L, Davidsen T, Demchok JA,**

**DiCara D, Dhalla N, Dhir R, Dookran SS, Dresdner G, Eldridge J, Eley G, El-Naggar AK, Eng S, Fagin JA, Fennell T, Ferris RL, Fisher S, Frazer S, Frick J, Gabriel SB, Ganly I, Gao J, Garraway LA, Gastier-Foster JM, Getz G, Gehlenborg N, Ghossein R, Gibbs RA, Giordano TJ, Gomez-Hernandez K, Grimsby J, Gross B, Guin R, Hadjipanayis A, Harper HA, Hayes DN, Heiman DI, Herman JG, Hoadley KA, Hofree M, Holt RA, Hoyle AP, Huang FW, Huang M, Hutter CM, Ideker T, Iype L, Jacobsen A, Jefferys SR, Jones CD, Jones SJM, Kasaian K, Kebebew E, Khuri FR, Kim J, Kramer R, Kreisberg R, Kucherlapati R, Kwiatkowski DJ, Ladanyi M, Lai PH, Laird PW, Lander E, Lawrence MS, Lee D, Lee E, Lee S, Lee W, Leraas KM, Lichtenberg TM, Lichtenstein L, Lin P, Ling S, Liu J, Liu W, Liu Y, LiVolsi VA, Lu Y, Ma Y, Mahadeshwar HS, Marra MA, Mayo M, McFadden DG, Meng S, Meyerson M, Mieczkowski PA, Miller M, Mills G, Moore RA, Mose LE, Mungall AJ, Murray BA, Nikiforov YE, Noble MS, Ojesina AI, Owonikoko TK, Ozenberger BA, Pantazi A, Parfenov M, Park PJ, Parker JS, Paull EO, Pedamallu CS, Perou CM, Prins JF, Protopopov A, Ramalingam SS, Ramirez NC, Ramirez R, Raphael B J, Rathmell WK, Ren X, Reynolds SM, Rheinbay E, Ringel MD, Rivera M, Roach J, Robertson AG, Rosenberg MW, Rosenthal M, Sadeghi S, Saksena G, Sander C, Santoso N, Schein JE, Schultz N, Schumacher SE, Seethala RR, Seidman J, Senbabaoglu Y, Seth S, Sharpe S, Shaw KRM, Shen JP, Shen R, Sherman S, Sheth M, Shi Y, Shmulevich I, Sica GL, Simons JV, Sinha R, Sipahimalani P, Smallridge RC, Sofia HJ, Soloway MG, Song X, Sougnez C, Stewart C, Stojanov P, Stuart JM, Sumer SO, Sun Y, Tabak B, Tam A, Tan D, et al. 2014.** Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159(3)**:676–690 DOI 10.1016/j.cell.2014.09.050.

**Akbani R, Akdemir K C, Aksoy BA, Albert M, Ally A, Amin S B, Arachchi H, Arora A, Auman JT, Ayala B, Baboud J, Balasundaram M, Balu S, Barnabas N, Bartlett J, Bartlett P, Bastian BC, Baylin SB, Behera M, Belyaev D, Benz C, Bernard B, Beroukhim R, Bir N, Black AD, Bodenheimer T, Boice L, Boland GM, Bono R, Bootwalla MS, Bosenberg M, Bowen J, Bowlby R, Bristow CA, Brockway-Lunardi L, Brooks D, Brzezinski J, Bshara W, Buda E, Burns WR, Butterfield YSN, Button M, Calderone T, Cappellini GA, Carter C, Carter SL, Cherney L, Cherniack AD, Chevalier A, Chin L, Cho J, Cho RJ, Choi Y-L, Chu A, Chudamani S, Cibulskis K, Ciriello G, Clarke A, Coons S, Cope L, Crain D, Curley E, Danilova L, D'Atri S, Davidsen T, Davies MA, Delman KA, Demchok JA, Deng QA, Deribe YL, Dhalla N, Dhir R, DiCara D, Dinikin M, Dubina M, Ebrom JS, Egea S, Eley G, Engel J, Eschbacher JM, Fedosenko KV, Felau I, Fennell T, Ferguson ML, Fisher S, Flaherty KT, Frazer S, Frick J, Fulidou V, Gabriel SB, Gao J, Gardner J, Garraway LA, Gastier-Foster JM, Gaudioso C, Gehlenborg N, Genovese G, Gerken M, Gershenwald JE, Getz G, Gomez-Fernandez C, Gribbin T, Grimsby J, Gross B, Guin R, Gutschner T, Hadjipanayis A, Halaban R, Hanf B, Haussler D, Haydu LE, Hayes DN, Hayward NK, Heiman DI, Herbert L, Herman JG, Hersey P, Hoadley KA, Hodis E, Holt RA, Hoon DSB, Hoppough S, Hoyle AP, Huang FW, Huang M, Huang S, Hutter CM, Ibbs M, Iype L, Jacobsen A, Jakrot V, Janning A, Jeck WR, Jefferys SR, Jensen MA, Jones CD, Jones SJM, Ju Z, Kakavand H, Kang H, Kefford RF, Khuri FR, Kim J, Kirkwood J M, Klode J, Korkut A, Korski K, Krauthammer M, Kucherlapati R, Kwong L N, Kycler W, Ladanyi M, Lai PH, Laird PW, Lander E, Lawrence MS, Lazar AJ, Łaźniak R, Lee D, Lee J E, Lee J, Lee K, Lee S, Lee W, Leporowska E, Leraas K M, Li HI, Lichtenberg T M, Lichtenstein L, Lin P, Ling S, Liu J, Liu O, Liu W, Long G V, Lu Y, Ma S, Ma Y, Mackiewicz A, Mahadeshwar H S, Malke J, Mallery D, Manikhas GM, Mann GJ, Marra M A, Matejka B, Mayo M, Mehrabi S, Meng S, Meyerson M, Mieczkowski PA, Miller JP, Miller ML, Mills G B, Moiseenko F, Moore RA, Morris S, Morrison C, Morton D, Moschos S, et al. 2015.** Genomic classification of cutaneous melanoma. *Cell* **161(7)**:1681–1696 DOI 10.1016/j.cell.2015.05.044.

**Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, Holt RA, Jones SJM, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Schein JE, Sipahimalani P, Tam A, Thiessen N, Cheung D, Wong T, Brooks D, Robertson AG, Bowlby R, Mungall K, Sadeghi S, Xi L, Covington K, Shinbrot E, Wheeler DA, Gibbs RA, Donehower LA, Wang L, Bowen J, Gastier-Foster JM, Gerken M, Helsel C, Leraas KM, Lichtenberg TM, Ramirez NC, Wise L, Zmuda E, Gabriel SB, Meyerson M, Cibulskis C, Murray BA, Shih J, Beroukhim R, Cherniack AD, Schumacher SE, Saksena G, Pedamallu CS, Chin L, Getz G, Noble M, Zhang H, Heiman D, Cho J, Gehlenborg N, Saksena G, Voet D, Lin P, Frazer S, Defreitas T, Meier S, Lawrence M, Kim J, Creighton CJ, Muzny D, Doddapaneni HV, Hu J, Wang M, Morton D, Korchina V, Han Y, Dinh H, Lewis L, Bellair M, Liu X, Santibanez J, Glenn R, Lee S, Hale W, Parker JS, Wilkerson MD, Hayes DN, Reynolds SM, Shmulevich I, Zhang W, Liu Y, Iype L, Makhlouf H, Torbenson MS, Kakar S, Yeh MM, Jain D, Kleiner DE, Jain D, Dhanasekaran R, El-Serag HB, Yim SY, Weinstein JN, Mishra L, Zhang J, Akbani R, Ling S, Ju Z, Su X, Hegde AM, Mills GB, Lu Y, Chen J, Lee J-S, Sohn BH, Shim JJ, Tong P, Aburatani H, Yamamoto S, Tatsuno K, Li W, Xia Z, Stransky N, Seiser E, Innocenti F, Gao J, Kundra R, Zhang H, Heins Z, Ochoa A, Sander C, Ladanyi M, Shen R, Arora A, Sanchez-Vega F, Schultz N, Kasaian K, Radenbaugh A, Bissig K-D, Moore DD, Totoki Y, Nakamura H, Shibata T, Yau C, Graim K, Stuart J, Haussler D, Slagle BL, Ojesina AI, Katsonis P, Koire A, Lichtarge O, Hsu T-K, Ferguson ML, Demchok JA, Felau I, Sheth M, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Hutter CM, Sofia HJ, Verhaak RGW, Zheng S, Lang F, Chudamani S, Liu J, Lolla L, Wu Y, Naresh R, Pihl T, Sun C, Wan Y, Benz C, Perou AH, Thorne LB, Boice L, Huang M, Rathmell WK, Noushmehr H, Saggioro FP, Tirapelli DPDC, Junior CGC, Mente ED, Silva ODC Jr, Trevisan FA, Kang KJ, Ahn KS, Giama NH, Moser CD, Giordano TJ, Vinco M, Welling TH, Crain D, Curley E, Gardner J, Mallery D, Morris S, Paulauskis J, Penny R, et al. 2017.** Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169(7)**:1327–1341 DOI 10.1016/j.cell.2017.05.046.

**Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK-S, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang W-W, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L,**

**Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle KP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, et al. 2018.** Comprehensive characterization of cancer driver genes and mutations. *Cell* **173(2)**:371–385.

**Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. 2020.** Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* **58**:82–115.

**Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, Rao A, Schultz A, Li X, Sumazin P, Williams C, Mestdagh P, Gunaratne PH, Yau C, Bowlby R, Robertson AG, Tiezzi DG, Wang C, Cherniack AD, Godwin AK, Kuderer NM, Rader JS, Zuna RE, Sood AK, Lazar AJ, Ojesina AI, Adebamowo C, Adebamowo SN, Baggerly KA, Chen T-W, Chiu H-S, Lefever S, Liu L, MacKenzie K, Orsulic S, Roszik J, Shelley CS, Song Q, Vellano CP, Wentzensen N, Weinstein JN, Mills GB, Levine DA, Akbani R, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, et al. 2018.** A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33(4)**:690–705.

**Bismeijer T, Canisius S, Wessels LF. 2018.** Molecular characterization of breast and lung tumors by integration of multiple data types with functional sparse-factor analysis. *PLOS Computational Biology* **14(10)**:e1006520 DOI 10.1371/journal.pcbi.1006520.

**Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu C-J, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla S A, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L, Benz C, Barnholtz-Sloan J, Barrett W, Ostrom Q, Wolinsky Y, Black KL, Bose B, Boulos PT, Boulos M, Brown J, Czerinski C, Eppley M, Iacocca M, Kempista T, Kitko T, Koyfman Y, Rabeno B, Rastogi P, Sugarman M, Swanson P, Yalamanchii K, Otey IP, Liu YS, Xiao Y, Auman JT, Chen P-C,

**Hadjipanayis A, Lee E, Lee S, Park PJ, Seidman J, Yang L, Kucherlapati R, Kalkanis S, Mikkelsen T, Poisson LM, Raghunathan A, Scarpace L, Bernard B, Bressler R, Eakin A, Iype L, Kreisberg RB, Leinonen K, Reynolds S, Rovira H, Thorsson V, Shmulevich I, Annala MJ, Penny R, Paulauskis J, Curley E, Hatfield M, Mallery D, Morris S, Shelton T, Shelton C, Sherman M, Yena P, Cuppini L, DiMeco F, Eoli M, Finocchiaro G, Maderna E, Pollo B, Saini M, Balu S, Hoadley KA, Li L, Miller CR, Shi Y, Topal MD, Wu J, Dunn G, Giannini C, O'Neill BP, Aksoy BA, Antipin Y, Borsu L, Berman SH, Brennan CW, Cerami E, Chakravarty D, Ciriello G, Gao J, Gross B, Jacobsen A, Ladanyi M, Lash A, Liang Y, Reva B, Sander C, Schultz N, Shen R, Socci ND, Viale A, Ferguson ML, Chen Q-R, Demchok JA, Dillon LAL, Shaw KRM, Sheth M, Tarnuzzer R, Wang Z, Yang L, Davidsen T, Guyer MS, Ozenberger BA, Sofia HJ, Bergsten J, Eckman J, Harr J, Myers J, Smith C, Tucker K, Winemiller C, Zach LA, Ljubimova JY, Eley G, Ayala B, Jensen MA, Kahn A, Pihl TD, Pot DA, Wan Y, Eschbacher J, Foltz G, Hansen N, Hothi P, Lin B, Shah N, Yoon J-G, Lau C, Berens M, Ardlie K, Beroukhim R, Carter SL, Cherniack AD, Noble M, Cho J, Cibulskis K, DiCara D, et al. 2013.** The somatic genomic landscape of glioblastoma. *Cell* **155(2)**:462–477 DOI 10.1016/j.cell.2013.09.034.

**Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, Shukla SA, Guo G, Brooks AN, Murray BA, Imielinski M, Hu X, Ling S, Akbani R, Rosenberg M, Cibulskis C, Ramachandran A, Collisson EA, Kwiatkowski DJ, Lawrence MS, Weinstein JN, Verhaak RGW, Wu CJ, Hammerman PS, Cherniack AD, Getz G, Artyomov MN, Schreiber R, Govindan R, Meyerson M, Cancer Genome Atlas Research Network. 2016.** Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics* **48(6)**:607–616 DOI 10.1038/ng.3564.

**Campbell JD, Yau C, Bowlby R, Liu Y, Brennan K, Fan H, Taylor AM, Wang C, Walter V, Akbani R, Byers LA, Creighton CJ, Coarfa C, Shih J, Cherniack AD, Gevaert O, Prunello M, Shen H, Anur P, Chen J, Cheng H, Hayes DN, Bullman S, Pedamallu CS, Ojesina AI, Sadeghi S, Mungall KL, Robertson AG, Benz C, Schultz A, Kanchi RS, Gay CM, Hegde A, Diao L, Wang J, Ma W, Sumazin P, Chiu H-S, Chen T-W, Gunaratne P, Donehower L, Rader JS, Zuna R, Al-Ahmadie H, Lazar AJ, Flores ER, Tsai KY, Zhou JH, Rustgi AK, Drill E, Shen R, Wong CK, Stuart JM, Laird PW, Hoadley KA, Weinstein JN, Peto M, Pickering CR, Chen Z, Van Waes C, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, et al. 2018.** Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell Reports* **23(1)**:194–212.

**Cappelli E, Felici G, Weitschek E. 2018.** Combining dna methylation and rna sequencing data of cancer for supervised knowledge extraction. *BioData Mining* **11(1)**:22 DOI 10.1186/s13040-018-0184-6.

**Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, Anjum S, Wang J, Manyam G, Zoppoli P, Ling S, Rao AA, Grifford M, Cherniack AD, Zhang H, Poisson L, Carlotti CG, Tirapelli DPDC, Rao A, Mikkelsen T, Lau CC, Yung WKA, Rabadan R, Huse J, Brat DJ, Lehman NL, Barnholtz-Sloan JS, Zheng S, Hess K, Rao G, Meyerson M, Beroukhim R, Cooper L, Akbani R, Wrensch M, Haussler D, Aldape KD, Laird PW, Gutmann DH, Noushmehr H, Iavarone A, Verhaak RGW, Anjum S, Arachchi H, Auman JT, Balasundaram M, Balu S, Barnett G, Baylin S, Bell S, Benz C, Bir N, Black KL, Bodenheimer T, Boice L, Bootwalla MS, Bowen J, Bristow CA, Butterfield YSN, Chen Q-R, Chin L, Cho J, Chuah E, Chudamani S, Coetzee SG, Cohen ML, Colman H, Couce M, D'Angelo F, Davidsen T, Davis A, Demchok JA, Devine K, Ding L, Duell R, Elder JB, Eschbacher JM, Fehrenbach A, Ferguson M, Frazer S, Fuller G, Fulop J, Gabriel SB, Garofano L, Gastier-Foster JM, Gehlenborg N, Gerken M, Getz G, Giannini C, Gibson WJ, Hadjipanayis A, Hayes DN, Heiman DI, Hermes B, Hilty J, Hoadley KA, Hoyle AP, Huang M, Jefferys SR, Jones CD, Jones SJM, Ju Z, Kastl A, Kendler A, Kim J, Kucherlapati R, Lai PH, Lawrence MS, Lee S, Leraas KM, Lichtenberg TM, Lin P, Liu Y, Liu J, Ljubimova JY, Lu Y, Ma Y, Maglinte DT, Mahadeshwar HS, Marra M A, McGraw M, McPherson C, Meng S, Mieczkowski PA, Miller CR, Mills GB, Moore RA, Mose LE, Mungall AJ, Naresh R, Naska T, Neder L, Noble MS, Noss A, O'Neill BP, Ostrom QT, Palmer C, Pantazi A, Parfenov M, Park PJ, Parker JS, Perou CM, Pierson CR, Pihl T, Protopopov A, Radenbaugh A, Ramirez NC, Rathmell WK, Ren X, Roach J, Robertson AG, Saksena G, Schein JE, Schumacher SE, Seidman J, Senecal K, Seth S, Shen H, Shi Y, Shih J, Shimmel K, Sicotte H, Sifri S, Silva T, Simons JV, Singh R, Skelly T, Sloan AE, Sofia HJ, Soloway MG, Song X, Sougnez C, Souza C, Staugaitis SM, Sun H, Sun C, Tan D, Tang J, Tang Y, Thorne L, Trevisan FA, Triche T, Van Den Berg DJ, Veluvolu U, Voet D, Wan Y, Wang Z, Warnick R, Weinstein JN, Weisenberger DJ, Wilkerson MD, Williams F, Wise L, Wolinsky Y, Wu J, Xu AW, et al. 2016.** Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164(3)**:550–563 DOI 10.1016/j.cell.2015.12.028.

**Cheerla A, Gevaert O. 2019.** Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35(14)**:i446–i454 DOI 10.1093/bioinformatics/btz342.

**Cheerla N, Gevaert O. 2017.** Microrna based pan-cancer diagnosis and treatment recommendation. *BMC Bioinformatics* **18(1)**:32 DOI 10.1186/s12859-016-1421-y.

**Chen H, Li C, Peng X, Zhou Z, Weinstein JN, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J,**

**Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, Liu X, Morgan M, Morton D, Fulton LA, Fulton RS, Kandoth C, Mardis ER, McLellan MD, Miller CA, et al. 2018a.** A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell* **173(2)**:386–399.

**Chen H-IH, Chiu Y-C, Zhang T, Zhang S, Huang Y, Chen Y. 2018b.** Gsae: an autoencoder with embedded gene-set nodes for genomics functional characterization. *BMC Systems Biology* **12(8)**:142.

**Chen L, Pan X, Hu X, Zhang Y-H, Wang SP, Huang T, Cai Y-D. 2018c.** G ene expression differences among different msi statuses in colorectal cancer. *International Journal of Cancer* **143(7)**:1731–1740.

**Chen L, Xuan J, Gu J, Wang Y, Zhang Z, Wang TL, Shih IM. 2012.** Integrative network analysis to identify aberrant pathway networks in ovarian cancer. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 17th Pacific Symposium on Biocomputing, PSB 2012*. 31–42.

**Chen X, Duan Q, Xuan Y, Sun Y, Wu R. 2017.** Possible pathways used to predict different stages of lung adenocarcinoma. *Medicine* **96(17)**.

**Cheng P. 2018.** A prognostic 3-long noncoding rna signature for patients with gastric cancer. *Journal of Cellular Biochemistry* **119(11)**:9261–9269 DOI 10.1002/jcb.27195.

**Cherniack AD, Shen H, Walter V, Stewart C, Murray BA, Bowlby R, Hu X, Ling S, Soslow RA, Broaddus RR, Zuna RE, Robertson G, Laird PW, Kucherlapati R, Mills GB, Weinstein JN, Zhang J, Akbani R, Levine DA, Akbani R, Ally A, Auman JT, Balasundaram M, Balu S, Baylin SB, Beroukhim R, Bodenheimer T, Bogomolniy F, Boice L, Bootwalla MS, Bowen J, Bowlby R, Broaddus R, Brooks D, Carlsen R, Cherniack AD, Cho J, Chuah E, Chudamani S, Cibulskis K, Cline M, Dao F, David M, Demchok JA, Dhalla N, Dowdy S, Felau I, Ferguson ML, Frazer S, Frick J, Gabriel S, Gastier-Foster JM, Gehlenborg N, Gerken M, Getz G, Gupta M, Haussler D, Hayes DN, Heiman DI, Hess J, Hoadley KA, Hoffmann R, Holt RA, Hoyle AP, Hu X, Huang M, Hutter CM, Jefferys SR, Jones SJM, Jones CD, Kanchi RS, Kandoth C, Kasaian K, Kerr S, Kim J, Lai PH, Laird PW, Lander E, Lawrence MS, Lee D, Leraas KM, Leshchiner I, Levine DA, Lichtenberg TM, Lin P, Ling S, Liu J, Liu W, Liu Y, Lolla L, Lu Y, Ma Y, Maglinte DT, Marra MA, Mayo M, Meng S, Meyerson M, Mieczkowski PA, Mills GB, Moore RA, Mose LE, Mungall AJ, Mungall K, Murray BA, Naresh R, Noble MS, Olvera N, Parker JS, Perou CM, Perou AH, Pihl T, Radenbaugh AJ, Ramirez NC, Rathmell WK, Roach J, Robertson AG, Sadeghi S, Saksena G, Salvesen HB, Schein JE, Schumacher SE, Shen H, Sheth M, Shi Y, Shih J, Simons JV, Sipahimalani P, Skelly T, Sofia HJ, Soloway MG, Soslow RA, Sougnez C, Stewart C, Sun C, Tam A, Tan D, Tarnuzzer R, Thiessen N, Thorne LB, Tse K, Tseng J, Van Den Berg DJ, Veluvolu U, Verhaak RGW, Voet D, von Bismarck A, Walter V, Wan Y, Wang Z, Wang C, Weinstein JN, Weisenberger DJ, Wilkerson MD, Winterhoff B,**

**Wise L, Wong T, Wu Y, Yang L, Zenklusen JC, Zhang J, Zhang H, Zhang W, Zhu J-C, Zmuda E, Zuna RE, et al. 2017.** Integrated molecular characterization of uterine carcinosarcoma. *Cancer Cell* **31(3)**:411–423 DOI 10.1016/j.ccell.2017.02.010.

**Chidester B, Do MN, Ma J. 2018.** Discriminative bag-of-cells for imaging-genomics. In: *PSB*. World Scientific, 319–330.

**Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, Bowlby R, Shen H, Hayat S, Fieldhouse R, Lester SC, Tse GMK, Factor RE, Collins LC, Allison KH, Chen Y-Y, Jensen K, Johnson NB, Oesterreich S, Mills GB, Cherniack AD, Robertson G, Benz C, Sander C, Laird PW, Hoadley KA, King TA, Perou CM, Akbani R, Auman JT, Balasundaram M, Balu S, Barr T, Beck A, Benz C, Benz S, Berrios M, Beroukhim R, Bodenheimer T, Boice L, Bootwalla MS, Bowen J, Bowlby R, Brooks D, Cherniack AD, Chin L, Cho J, Chudamani S, Ciriello G, Davidsen T, Demchok JA, Dennison JB, Ding L, Felau I, Ferguson ML, Frazer S, Gabriel S B, Gao JJ, Gastier-Foster JM, Gatza ML, Gehlenborg N, Gerken M, Getz G, Gibson WJ, Hayes DN, Heiman DI, Hoadley KA, Holbrook A, Holt RA, Hoyle AP, Hu H, Huang M, Hutter CM, Hwang ES, Jefferys SR, Jones SJM, Ju Z, Kim J, Lai PH, Laird PW, Lawrence MS, Leraas KM, Lichtenberg TM, Lin P, Ling S, Liu J, Liu W, Lolla L, Lu Y, Ma Y, Maglinte DT, Mardis E, Marks J, Marra MA, McAllister C, McLellan M, Meng S, Meyerson M, Mills GB, Moore RA, Mose LE, Mungall AJ, Murray BA, Naresh R, Noble MS, Oesterreich S, Olopade O, Parker JS, Perou CM, Pihl T, Saksena G, Schumacher SE, Shaw KRM, Ramirez NC, Rathmell WK, Rhie SK, Roach J, Robertson AG, Saksena G, Sander C, Schein JE, Schultz N, Shen H, Sheth M, Shi Y, Shih J, Shelley CS, Shriver C, Simons J V, Sofia HJ, Soloway MG, Sougnez C, Sun C, Tarnuzzer R, Tiezzi DG, Van Den Berg DJ, Voet D, Wan Y, Wang Z, Weinstein JN, Weisenberger DJ, Wilkerson MD, Wilson R, Wise L, Wiznerowicz M, Wu J, Wu Y, Yang L, Yau C, Zack TI, Zenklusen JC, Zhang H, Zhang J, Zmuda E, et al. 2015.** Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163(2)**:506–519 DOI 10.1016/j.cell.2015.09.033.

**Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A. 2018.** Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine* **24(10)**:1559–1567 DOI 10.1038/s41591-018-0177-5.

**Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale A-L, Brenton JD, Tavaré S, Caldas C, Aparicio S. 2012.** The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486(7403)**:346.

**Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, Pepin F, Durinck S, Korkola JE, Griffith M, Hur JS, Huh N, Chung J, Cope L, Fackler M, Umbricht C, Sukumar S, Seth P, Sukhatme VP, Jakkula LR, Lu Y, Mills GB, Cho RJ, Collisson EA, van't Veer LJ, Spellman PT, Gray JW. 2013.** Modeling precision treatment of breast cancer. *Genome Biology* **14(10)**:R110 DOI 10.1186/gb-2013-14-10-r110.

**Dai Y, Sun C, Feng Y, Jia Q, Zhu B. 2018.** Potent immunogenicity in brca 1-mutated patients with high-grade serous ovarian carcinoma. *Journal of Cellular and Molecular Medicine* **22(8)**:3979–3986.

**Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, Buhay C, Kang H, Kim SC, Fahey CC, Hacker KE, Bhanot G, Gordenin DA, Chu A, Gunaratne PH, Biehl M, Seth S, Kaipparettu BA, Bristow CA, Donehower LA, Wallen EM, Smith AB, Tickoo SK, Tamboli P, Reuter V, Schmidt LS, Hsieh JJ, Choueiri TK, Hakimi AA, Chin L, Meyerson M,**

**Kucherlapati R, Park W-Y, Robertson AG, Laird PW, Henske EP, Kwiatkowski DJ, Park PJ, Morgan M, Shuch B, Muzny D, Wheeler DA, Linehan WM, Gibbs RA, Rathmell WK, Creighton CJ, Creighton CJ, Davis CF, Morgan M, Gunaratne PH, Donehower LA, Kaipparettu BA, Wheeler DA, Gibbs RA, Signoretti S, Cherniack AD, Robertson AG, Chu A, Choueiri TK, Henske EP, Kwiatkowski DJ, Reuter V, Hsieh JJ, Hakimi AA, Tickoo SK, Ricketts C, Linehan WM, Schmidt LS, Gordenin DA, Bhanot G, Seiler M, Tamboli P, Rathmell WK, Fahey CC, Hacker KE, Smith AB, Wallen EM, Shen H, Laird PW, Shuch B, Muzny D, Buhay C, Wang M, Chao H, Dahdouli M, Xi L, Kakkar N, Reid JG, Downs B, Drummond J, Morton D, Doddapaneni H, Lewis L, English A, Meng Q, Kovar C, Wang Q, Hale W, Hawes A, Kalra D, Walker K, Murray BA, Sougnez C, Saksena G, Carter SL, Schumacher SE, Tabak B, Zack TI, Getz G, Beroukhim R, Gabriel SB, Meyerson M, Ally A, Balasundaram M, Birol I, Brooks D, Butterfield YSN, Chuah E, Clarke A, Dhalla N, Guin R, Holt RA, Kasaian K, Lee D, Li HI, Lim E, Ma Y, Mayo M, Moore RA, Mungall AJ, Schein JE, Sipahimalani P, Tam A, Thiessen N, Wong T, Jones SJM, Marra M A, Auman JT, Tan D, Meng S, Jones CD, Hoadley KA, Mieczkowski PA, Mose LE, Jefferys SR, Roach J, Veluvolu U, Wilkerson MD, Waring S, Buda E, Wu J, Bodenheimer T, Hoyle AP, Simons JV, Soloway MG, Balu S, Parker JS, Hayes DN, Perou CM, Weisenberger DJ, Bootwalla MS, Triche T Jr, Lai PH, Van Den Berg DJ, Baylin SB, Chen F, Coarfa C, Noble MS, DiCara D, Zhang H, Cho J, Heiman DI, Gehlenborg N, Voet D, Lin P, Frazer S, Stojanov P, Liu Y, Zou L, Kim J, Lawrence MS, Chin L, Yang L, Seth S, Bristow CA, Protopopov A, Song X, Zhang J, Pantazi A, Hadjipanayis A, Lee E, Luquette LJ, Lee S, Parfenov M, Santoso N, Seidman J, Xu AW, Kucherlapati R, Park PJ, Kang H, et al. 2014.** The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26(3)**:319–330 DOI 10.1016/j.ccr.2014.07.014.

**Dong C, Guo Y, Yang H, He Z, Liu X, Wang K. 2016.** icages: integrated cancer genome score for comprehensively prioritizing driver genes in personal cancer genomes. *Genome Medicine* **8(1)**:135.

**Ellrott K, Bailey MH, Saksena G, Covingt, andoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, Sofia HJ, Hutter C, Getz G, Wheeler D, Ding L, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker WMD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA,**

**Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, et al. 2018.** Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Systems* **6(3)**:271–281.

**Ertosun MG, Rubin DL. 2015.** Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. In: *AMIA Annual Symposium Proceedings*. **2015**:American Medical Informatics Association, 1899.

**Fan Z, Xue W, Li L, Zhang C, Lu J, Zhai Y, Suo Z, Zhao J. 2018.** Identification of an early diagnostic biomarker of lung adenocarcinoma based on co-expression similarity and construction of a diagnostic model. *Journal of Translational Medicine* **16(1)**:205.

**Farshidfar F, Zheng S, Gingras M-C, Newton Y, Shih J, Robertson AG, Hinoue T, Hoadley KA, Gibb EA, Roszik J, Covington KR, Wu C-C, Shinbrot E, Stransky N, Hegde A, Yang JD, Reznik E, Sadeghi S, Pedamallu CS, Ojesina AI, Hess JM, Auman JT, Rhie SK, Bowlby R, Borad MJ, Zhu AX, Stuart JM, Sander C, Akbani R, Cherniack AD, Deshpande V, Mounajjed T, Foo WC, Torbenson MS, Kleiner DE, Laird PW, Wheeler DA, McRee AJ, Bathe OF, Andersen JB, Bardeesy N, Roberts LR, Kwong LN, Akbani R, Allotey LK, Ally A, Alvaro D, Andersen JB, Appelbaum EL, Arora A, Auman JT, Balasundaram M, Balu S, Bardeesy N, Bathe OF, Baylin SB, Beroukhim R, Berrios M, Bodenheimer T, Boice L, Bootwalla MS, Borad MJ, Bowen J, Bowlby R, Bragazzi MC, Brooks D, Cardinale V, Carlsen R, Carpino G, Carvalho AL, Chaiteerakij R, Chandan VC, Cherniack AD, Chin L, Cho J, Choe G, Chuah E, Chudamani S, Cibulskis C, Cordes MG, Covington KR, Crain D, Curley E, De Rose AM, Defreitas T, Demchok JA, Deshpande V, Dhalla N, Ding L, Evason K, Farshidfar F, Felau I, Ferguson ML, Foo WC, Franchitto A, Frazer S, Fronick CC, Fulton LA, Fulton RS, Gabriel SB, Gardner J, Gastier-Foster JM, Gaudio E, Gehlenborg N, Genovese G, Gerken M, Getz G, Giama NH, Gibbs RA, Gingras M-C, Giuliante F, Grazi GL, Hayes DN, Hegde AM, Heiman DI, Hess JM, Hinoue T, Hoadley KA, Holbrook A, Holt RA, Hoyle AP, Huang M, Hutter CM, Jefferys SR, Jones SJM, Jones CD, Kasaian K, Kelley RK, Kim J, Kleiner DE, Kocher J-PA, Kwong LN, Lai PH, Laird PW, Lawrence MS, Leraas KM, Lichtenberg TM, Lin P, Liu W, Liu J, Lolla L, Lu Y, Ma Y, Mallery D, Mardis ER, Marra MA, Matsushita MM, Mayo M, McLellan MD, McRee AJ, Meier S, Meng S, Meyerson M, Mieczkowski PA, Miller CA, Mills GB, Moore RA, Morris S, Mose LE, Moser CD, Mounajjed T, Mungall AJ, Mungall K, Murray BA, Naresh R, Newton Y, Noble MS, O'Brien DR, Ojesina AI, Parker JS, Patel TC, Paulauskis J, Pedamallu CS, Penny R, Perou CM, Perou AH, Pihl T, Radenbaugh AJ, Ramirez NC, Rathmell WK, Reznik E, Rhie SK, Roach J, Roberts LR, Robertson AG, Sadeghi S, Saksena G, Sander C, Schein JE, Schmidt HK, Schumacher SE, Shelton C, Shelton T, Shen R, Sheth M, Shi Y, Shih J, Shinbrot E, Shroff R, Simons JV, et al. 2017.** Integrative genomic analysis of cholangiocarcinoma identifies distinct idh-mutant molecular profiles. *Cell Reports* **18(11)**:2780–2794 DOI 10.1016/j.celrep.2017.02.033.

**Fatai AA, Gamieldien J. 2018.** A 35-gene signature discriminates between rapidly-and slowly-progressing glioblastoma multiforme and predicts survival in known subtypes of the cancer. *BMC Cancer* **18(1)**:377 DOI 10.1186/s12885-018-4103-5.

**Feng Y, Dai Y, Gong Z, Cheng J-N, Zhang L, Sun C, Zeng X, Jia Q, Zhu B. 2018.** Association between angiogenesis and cytotoxic signatures in the tumor microenvironment of gastric cancer. *OncoTargets and Therapy* **11**:2725 DOI 10.2147/OTT.

**Fernandez-Lozano C, Gestal M, Munteanu CR, Dorado J, Pazos A. 2016.** A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ* **4**:e2721.

**Fischer W, Moudgalya SS, Cohn JD, Nguyen NTT, Kenyon GT. 2018.** Sparse coding of pathology slides compared to transfer learning with deep neural networks. *BMC Bioinformatics* **19(18)**:489.

**Fishbein L, Leshchiner I, Walter V, Danilova L, Robertson AG, Johnson AR, Lichtenberg TM, Murray BA, Ghayee HK, Else T, Ling S, Jefferys SR, de Cubas AA, Wenz B, Korpershoek E, Amelio AL, Makowski L, Rathmell WK, Gimenez-Roqueplo A-P, Giordano TJ, Asa SL, Tischler AS, Pacak K, Nathanson KL, Wilkerson MD, Akbani R, Ally A, Amar L, Amelio AL, Arachchi H, Asa SL, Auchus RJ, Auman JT, Baertsch R, Balasundaram M, Balu S, Bartsch DK, Baudin E, Bauer T, Beaver A, Benz C, Beroukhim R, Beuschlein F, Bodenheimer T, Boice L, Bowen J, Bowlby R, Brooks D, Carlsen R, Carter S, Cassol CA, Cherniack AD, Chin L, Cho J, Chuah E, Chudamani S, Cope L, Crain D, Curley E, Danilova L, de Cubas AA, de Krijger RR, Demchok JA, Deutschbein T, Dhalla N, Dimmock D, Dinjens WNM, Else T, Eng C, Eschbacher J, Fassnacht M, Felau I, Feldman M, Ferguson ML, Fiddes I, Fishbein L, Frazer S, Gabriel SB, Gardner J, Gastier-Foster JM, Gehlenborg N, Gerken M, Getz G, Geurts J, Ghayee HK, Gimenez-Roqueplo A-P, Giordano TJ, Goldman M, Graim K, Gupta M, Haan D, Hahner S, Hantel C, Haussler D, Hayes DN, Heiman DI, Hoadley KA, Holt RA, Hoyle AP, Huang M, Hunt B, Hutter CM, Jefferys SR, Johnson AR, Jones SJM, Jones CD, Kasaian K, Kebebew E, Kim J, Kimes P, Knijnenburg T, Korpershoek E, Lander E, Lawrence MS, Lechan R, Lee D, Leraas KM, Lerario A, Leshchiner I, Lichtenberg TM, Lin P, Ling S, Liu J, LiVolsi VA, Lolla L, Lotan Y, Lu Y, Ma Y, Maison N, Makowski L, Mallery D, Mannelli M, Marquard J, Marra MA, Matthew T, Mayo M, Méatchi T, Meng S, Merino MJ, Mete O, Meyerson M, Mieczkowski PA, Mills GB, Moore RA, Morozova O, Morris S, Mose LE, Mungall AJ, Murray BA, Naresh R, Nathanson KL, Newton Y, Ng S, Ni Y, Noble MS, Nwariaku F, Pacak K, Parker JS, Paul E, Penny R, Perou CM, Perou AH, Pihl T, Powers J, Rabaglia J, Radenbaugh A, Ramirez NC, Rao A, Rathmell WK, Riester A, Roach J, Robertson AG, Sadeghi S, Saksena G, Salama S, Saller C, Sandusky G, Sbiera S, Schein JE, Schumacher SE, Shelton C, Shelton T, Sheth M, Shi Y, Shih J, Shmulevich I, Simons JV, Sipahimalani P, Skelly T, Sofia HJ, Sokolov A, Soloway MG, Sougnez C, Stuart J, Sun C, Swatloski T, Tam A, Tan D, Tarnuzzer R, Tarvin K, et al. 2017.** Comprehensive molecular characterization of pheochromocytoma and paraganglioma. *Cancer Cell* **31(2)**:181–193 DOI 10.1016/j.ccell.2017.01.001.

**Gao S, Qiu Z, Song Y, Mo C, Tan W, Chen Q, Liu D, Chen M, Zhou H. 2017.** Unsupervised clustering reveals new prostate cancer subtypes. *Translational Cancer Research* **6(3)**:561–572.

**Ge Z, Leighton JS, Wang Y, Peng X, Chen Z, Chen H, Sun Y, Yao F, Li J, Zhang H, Liu J, Shriver CD, Hu H, Piwnica-Worms H, Ma L, Liang H, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ,**

**Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, et al. 2018.** Integrated genomic analysis of the ubiquitin pathway across cancer types. *Cell Reports* **23(1)**:213–226.

**Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, Bhutra S, Grossman RL, Huang RS. 2017.** Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Research* **27(10)**:1743–1751.

**Ghoshal A, Zhang J, Roth MA, Xia KM, Grama AY, Chaterji S. 2018.** A distributed classifier for microrna target prediction with validation through tcga expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **15(4)**:1037–1051.

**Graudenzi A, Cava C, Bertoli G, Fromm B, Flatmark K, Mauri G, Castiglioni I. 2017.** Pathway-based classification of breast cancer subtypes. *Frontiers in Bioscience* **22**:1697–1712.

**Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, Flagg K, Hou J, Zhang H, Yi S, Jafari M, Lin D, Chung C, Caughey BA, Li G, Dhar D, Shi W, Zheng L, Hou R, Zhu J, Zhao L, Fu X, Zhang E, Zhang C, Zhu J-K, Karin M, Xu R-H, Zhang K. 2017.** Dna methylation markers for diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences of the United States of America* **114(28)**:7414–7419 DOI 10.1073/pnas.1703577114.

**Hmeljak J, Sanchez-Vega F, Hoadley KA, Shih J, Stewart C, Heiman D, Tarpey P, Danilova L, Drill E, Gibb EA, Bowlby R, Kanchi R, Osmanbeyoglu HU, Sekido Y, Takeshita J, Newton Y, Graim K, Gupta M, Gay CM, Diao L, Gibbs DL, Thorsson V, Iype L, Kantheti H, Severson DT, Ravegnini G, Desmeules P, Jungbluth AA, Travis WD, Dacic S, Chirieac LR, Fçoise G-Sallé, Fujimoto J, Husain AN, Silveira HC, Rusch VW, Rintoul RC, Pass H, Kindler H, Zauderer MG, Kwiatkowski DJ, Bueno R, Tsao AS, Creaney J, Lichtenberg T, Leraas K, Bowen J, Zenklusen JC, Akbani R, Cherniack AD, Byers LA, Noble MS, Fletcher JA, Robertson AG, Shen R, Aburatani H, Robinson BW, Campbell P, Ladanyi M. 2018.** Integrative molecular characterization of malignant pleural mesothelioma. *Cancer Discovery* **8(12)**:1548–1565.

**Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, Akbani R, Bowlby R, Wong CK, Wiznerowicz M, Sanchez-Vega F, Robertson AG, Schneider BG, Lawrence MS, Noushmehr H, Malta TM, Stuart JM, Benz CC, Laird PW, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DAkbani R, Bowlby R, Wong CK, Wiznerowicz M, Sanchez-Vega F, Robertson AG, Schneider BG, Lawrence MS, Noushmehr H, Malta TM, Stuart JM, Benz CC, Laird PW, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y,**

**Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, et al. 2018.** Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173(2)**:291–304.

**Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, Zhang J, Kandoth C, Akbani R, Shen H, Omberg L, Chu A, Margolin AA, van't Veer LJ, Lopez-Bigas N, Laird PW, Raphael BJ, Ding L, Robertson AG, Byers LA, Mills GB, Weinstein JN, Van Waes C, Chen Z, Collisson EA, Benz CC, Perou CM, Stuart JM. 2014.** Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158(4)**:929–944.

**Holzinger A, Malle B, Saranti A, Pfeifer B. 2021.** Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. *Information Fusion* **71(7639)**:28–37 DOI 10.1016/j.inffus.2021.01.008.

**Huang K-l, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, Scott AD, Krassowski M, Cherniack AD, Houlahan KE, Jayasinghe R, Wang L-B, Zhou DC, Liu D, Cao S, Kim YW, Koire A, McMichael JF, Hucthagowder V, Kim T-B, Hahn A, Wang C, McLellan MD, Al-Mulla F, Johnson KJ, Lichtarge O, Boutros PC, Raphael B, Lazar AJ, Zhang W, Wendl MC, Govindan R, Jain S, Wheeler D, Kulkarni S, Dipersio JF, Jüri R, Meric-Bernstam F, Chen K, Shmulevich I, Plon SE, Chen F, Ding L, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, et al. 2018.** Pathogenic germline variants in 10,389 adult cancers. *Cell* **173(2)**:355–370.

**Ing N, Huang F, Conley A, You S, Ma Z, Klimov S, Ohe C, Yuan X, Amin MB, Figlin R, Gertych A, Knudsen BS. 2017.** A novel machine learning approach reveals latent vascular phenotypes predictive of renal cancer outcome. *Scientific Reports* **7(1)**:13190 DOI 10.1038/s41598-017-13196-4.

**Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. 2019.** Histoqc: an open-source quality control tool for digital pathology slides. *JCO Clinical Cancer Informatics* **3(3)**:1–7 DOI 10.1200/CCI.18.00157.

**Jean-Quartier C, Jeanquartier F, Ridvan A, Kargl M, Mirza T, Stangl T, Markaĉ R, Jurada M, Holzinger A. 2021.** Mutation-based clustering and classification analysis reveals distinctive age groups and age-related biomarkers for glioma. *BMC Medical Informatics and Decision Making* **21(1)**:1–14.

**Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O, Sander C, The Cancer Genome Atlas Research Network. 2018.** Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* **34(2)**:211–224.

**Kanas VG, Zacharaki EI, Thomas GA, Zinn PO, Megalooikonomou V, Colen RR. 2017.** Learning mri-based classification models for mgmt methylation status prediction in glioblastoma. *Computer Methods and Programs in Biomedicine* **140**:249–257.

**Karczewski KJ, Snyder MP. 2018.** Integrative omics for health and disease. *Nature Reviews Genetics* **19(5)**:299.

**Kim S, Oesterreich S, Kim S, Park Y, Tseng GC. 2017.** Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics* **18(1)**:165–179 DOI 10.1093/biostatistics/kxw039.

**Klein MI, Stern DF, Zhao H. 2017.** Grape: a pathway template method to characterize tissue-specific functionality from gene expression profiles. *BMC Bioinformatics* **18(1)**:317 DOI 10.1186/s12859-017-1711-z.

**Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, Greene CS, Liu Y, Akbani R, Feng B, Donehower LA, Miller C, Shen Y, Karimi M, Chen H, Kim P, Jia P, Shinbrot E, Zhang S, Liu J, Hu H, Bailey MH, Yau C, Wolf D, Zhao Z, Weinstein JN, Li L, Ding L, Mills GB, Laird PW, Wheeler DA, Shmulevich I, Monnat RJ, Xiao Y, Wang C, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, et al. 2018.** Genomic and molecular landscape of dna damage repair deficiency across The Cancer Genome Atlas. *Cell Reports* **23(1)**:239–254.

**Kocak B, Yardimci AH, Bektas CT, Turkcanoglu MH, Erdim C, Yucetas U, Koca SB, Kilickesmez O. 2018.** Textural differences between renal cell carcinoma subtypes: machine learning-based quantitative computed tomography texture analysis with independent external validation. *European Journal of Radiology* **107**:149–157.

**Koo J, Zhang J, Chaterji S. 2018.** Tiresias: context-sensitive approach to decipher the presence and strength of microrna regulatory interactions. *Theranostics* **8(1)**:277–291 DOI 10.7150/thno.22065.

**Kristensen V, Lingjærde O, Russnes H, Vollan H, Frigessi A, Børresen-Dale A. 2014.** Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer* **14**:299–313.

**Leung MK, Delong A, Alipanahi B, Frey BJ. 2015.** Machine learning in genomic medicine: a review of computational problems and data sets. *Proceedings of the IEEE* **104(1)**:176–197 DOI 10.1109/JPROC.2015.2494198.

**Levine DA, The Cancer Genome Atlas Research Network. 2013.** Integrated genomic characterization of endometrial carcinoma. *Nature* **497(7447)**:67–73 DOI 10.1038/nature12113.

**Liao Z, Li D, Wang X, Li L, Zou Q. 2018.** Cancer diagnosis through isomir expression with machine learning method. *Current Bioinformatics* **13(1)**:57–63.

**Liñares Blanco J, Gestal M, Dorado J, Fernandez-Lozano C. 2019.** *Differential gene expression analysis of RNA-seq data using machine learning for cancer research.* Cham: Springer International Publishing, 27–65.

**List M, Hauschild A-C, Tan Q, Kruse TA, Baumbach J, Batra R. 2014.** Classification of breast cancer subtypes by combining gene expression and dna methylation data. *Journal of Integrative Bioinformatics* **11(2)**:1–14.

**Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, Omberg L, Wolf DM, Shriver CD, Thorsson V, Hu H, Cancer Genome Atlas Research Network. 2018a.** An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173(2)**:400–416.

**Liu Y, Sethi NS, Hinoue T, Schneider BG, Cherniack AD, Sanchez-Vega F, Seoane JA, Farshidfar F, Bowlby R, Islam M, Kim J, Chatila W, Akbani R, Kanchi RS, Rabkin CS, Willis JE, Wang KK, McCall SJ, Mishra L, Ojesina AI, Bullman S, Pedamallu CA, Lazar AJ, Sakai R, Thorsson V, Bass AJ, Laird RW, Cancer Genome Atlas Research Network. 2018b.** Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell* **33(4)**:721–735.

**Mallavarapu T, Hao J, Kim Y, Oh JH, Kanga M. 2019.** Pathway-based deep clustering for molecular subtyping of cancer. *Methods* **173**:24–31 DOI 10.1016/j.ymeth.2019.06.017.

**Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O, Colaprico A, Czerwińska P, Mazurek S, Mishra L, Heyn H, Krasnitz A, Godwin AK, Lazar AJ, The Cancer Genome Atlas Research Network. 2018.** Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* **173(2)**:338–354.

**Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. 2017.** A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19(1)**:71–86.

**Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. 2013.** Pattern discovery and cancer gene identification in integrated cancer genomic data.

*Proceedings of the National Academy of Sciences of the United States of America* **110(11)**:4245–4250 DOI 10.1073/pnas.1208949110.

**Muhamed Ali A, Zhuang H, Ibrahim A, Rehman O, Huang M, Wu A. 2018.** A machine learning approach for the classification of kidney cancer subtypes using mirna genome data. *Applied Sciences* **8(12)**:2422.

**Nair J, Jain P, Chandola U, Palve V, Vardhan NRH, Reddy RB, Kekatpure VD, Suresh A, Kuriakose MA, Panda B. 2015.** Gene and mirna expression changes in squamous cell carcinoma of larynx and hypopharynx. *Genes & Cancer* **6(7–8)**:328–340 DOI 10.18632/genesandcancer.69.

**Nguyen T, Tagett R, Diaz D, Draghici S. 2017.** A novel approach for data integration and disease subtyping. *Genome Research* **27(12)**:2025–2039 DOI 10.1101/gr.215129.116.

**Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, Verhaak RGW, Hoadley KA, Hayes DN, Perou CM, Schmidt HK, Ding L, Wilson RK, Van Den Berg D, Shen H, Bengtsson H, Neuvial P, Cope LM, Buckley J, Herman JG, Baylin SB, Laird PW, Aldape K. 2010.** Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17(5)**:510–522 DOI 10.1016/j.ccr.2010.03.017.

**Ou-Yang L, Zhang X-F, Wu M, Li X-L. 2017.** Node-based learning of differential networks from multi-platform gene expression data. *Methods* **129(1)**:41–49 DOI 10.1016/j.ymeth.2017.05.014.

**Park YW, Choi YS, Ahn SS, Chang JH, Kim SH, Lee S-K. 2019.** Radiomics mri phenotyping with machine learning to predict the grade of lower-grade gliomas: A study focused on nonenhancing tumors. *Korean Journal of Radiology* **20(9)**:1381–1389.

**Peng X, Chen Z, Farshidfar F, Xu X, Lorenzi PL, Wang Y, Cheng F, Tan L, Mojumdar K, Du D, Ge Z, Li J, Thomas GV, Birsoy K, Liu L, Zhang H, Zhao Z, Marchand C, Weinstein JN, Bathe OF, Liang H, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, et al. 2018.** Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. *Cell Reports* **23(1)**:255–269.

**Powell RT, Olar A, Narang S, Rao G, Sulman E, Fuller GN, Rao A. 2017.** Identification of histological correlates of overall survival in lower grade gliomas using a bag-of-words paradigm: A preliminary analysis based on hematoxylin & eosin stained slides from the lower grade glioma cohort of The Cancer Genome Atlas. *Journal of Pathology Informatics* **8**:9.

**Radovich M, Pickering CR, Felau I, Ha G, Zhang H, Jo H, Hoadley KA, Anur P, Zhang J, McLellan M, Bowlby R, Matthew T, Danilova L, Hegde AM, Kim J, Leiserson MDM, Sethi G, Lu C, Ryan M, Su X, Cherniack AD, Robertson G, Akbani R, Spellman P, Weinstein JN, Hayes DN, Raphael B, Lichtenberg T, Leraas K, Zenklusen JC, Fujimoto J, Scapulatempo-Neto C, Moreira AL, Hwang D, Huang J, Marino M, Korst R, Giaccone G, Gokmen-Polar Y, Badve S, Rajan A, Ströbel P, Girard N, Tsao MS, Marx A, Tsao AS, Loehrer PJ, Ally A, Appelbaum EL, Auman JT, Balasundaram M, Balu S, Behera M, Beroukhim R, Berrios M, Blandino G, Bodenheimer T, Bootwalla MS, Bowen J, Brooks D, Carcano FM, Carlsen R, Carvalho AL, Castro P, Chalabreysse L, Chin L, Cho J, Choe G, Chuah E, Chudamani S, Cibulskis C, Cope L, Cordes MG, Crain D, Curley E, Defreitas T, Demchok JA, Detterbeck F, Dhalla N, Dienemann H, Edenfield WJ, Facciolo F, Ferguson ML, Frazer S, Fronick CC, Fulton LA, Fulton RS, Gabriel SB, Gardner J, Gastier-Foster JM, Gehlenborg N, Gerken M, Getz G, Heiman DI, Hobensack S, Holbrook A, Holt RA, Hoyle AP, Hutter CM, Ittmann M, Jefferys SR, Jones CD, Jones SJM, Kasaian K, Kim J, Kimes PK, Lai PH, Laird PW, Lawrence MS, Lin P, Liu J, Lolla L, Lu Y, Ma Y, Maglinte DT, Mallery D, Mardis ER, Marra MA, Martin J, Mayo M, Meier S, Meister M, Meng S, Meyerson M, Mieczkowski PA, Miller CA, Mills GB, Moore RA, Morris S, Mose LE, Muley T, Mungall AJ, Mungall K, Naresh R, Newton Y, Noble MS, Owonikoko T, Parker JS, Paulaskis J, Penny R, Perou CM, Perrin C, Pihl T, Radenbaugh A, Ramalingam S, Ramirez N, Rieker R, Roach J, Sadeghi S, Saksena G, Schein JE, Schmidt HK, Schumacher SE, Shelton C, Shelton T, Shi Y, Shih J, Sica G, Silveira HCS, Simons JV, Sipahimalani P, Skelly T, Sofia HJ, Soloway MG, Stuart J, Sun Q, Tam A, Tan D, Tarnuzzer R, Thiessen N, Van Den Berg DJ, Vasef MA, Veluvolu U, Voet D, Walter V, Wan Y, Wang Z, Warth A, Weis C-A, Weisenberger DJ, Wilkerson MD, Wise L, Wong T, Wu H-T, Wu Y, Yang L, Zhang J, Zmuda E, et al. 2018.** The integrated genomic landscape of thymic epithelial tumors. *Cancer Cell* **33(2)**:244–258 DOI 10.1016/j.ccell.2018.01.003.

**Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, Robertson AG, Cherniack AD, Gupta M, Getz G, Gabriel SB, Meyerson M, Cibulskis C, Fei SS, Hinoue T, Shen H, Laird PW, Ling S, Lu Y, Mills GB, Akbani R, Loher P, Londin ER, Rigoutsos I, Telonis AG, Gibb EA, Goldenberg A, Mezlini AM, Hoadley KA, Collisson E, Lander E, Murray BA, Hess J, Rosenberg M, Bergelson L, Zhang H, Cho J, Tiao G, Kim J, Livitz D, Leshchiner I, Reardon B, Van Allen E, Kamburov A, Beroukhim R, Saksena G, Schumacher SE, Noble MS, Heiman DI, Gehlenborg N, Kim J, Lawrence MS, Adsay V, Petersen G, Klimstra D, Bardeesy N, Leiserson MDM, Bowlby R, Kasaian K, Birol I, Mungall KL, Sadeghi S, Weinstein JN, Spellman PT, Liu Y, Amundadottir LT, Tepper J, Singhi AD, Dhir R, Paul D, Smyrk T, Zhang L, Kim P, Bowen J, Frick J, Gastier-Foster JM, Gerken M, Lau K, Leraas KM, Lichtenberg TM, Ramirez NC, Renkel J, Sherman M, Wise L, Yena P, Zmuda E, Shih J, Ally A, Balasundaram M, Carlsen R, Chu A, Chuah E, Clarke A, Dhalla N, Holt RA, Jones SJM, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Brooks D, Auman JT, Balu S, Bodenheimer T, Hayes DN, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou CM, Perou AH, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Parker JS, Wilkerson MD, Korkut A, Senbabaoglu Y, Burch P, McWilliams R, Chaffee K, Oberg A, Zhang W, Gingras M-C, Wheeler DA, Xi L, Albert M, Bartlett J, Sekhon H, Stephen Y, Howard Z, Judy M, Breggia A, Shroff RT, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Jennifer S, Roggin K, Becker K-F, Behera M, Bennett J, Boice L, Burks E, Carlotti Junior

**CG, Chabot J, Pretti da Cunha Tirapelli D, Sebastião dos Santos J, Dubina M, Eschbacher J, Huang M, Huelsenbeck-Dill L, Jenkins R, Karpov A, Kemp R, Lyadov V, Maithel S, Manikhas G, Montgomery E, Noushmehr H, Osunkoya A, Owonikoko T, Paklina O, Potapova O, Ramalingam S, Rathmell WK, Rieger-Christ K, Saller C, Setdikova G, Shabunin A, Sica G, Su T, Sullivan T, Swanson P, Tarvin K, Tavobilov M, Thorne LB, Urbanski S, Voronina O, Wang T, Crain D, et al. 2017.** Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32(2)**:185–203 DOI 10.1016/j.ccell.2017.07.007.

**Rendleman MC, Buatti JM, Braun TA, Smith BJ, Nwakama C, Beichel RR, Brown B, Casavant TL. 2019.** Machine learning with the tcga-hnsc dataset: improving usability by addressing inconsistency, sparsity, and high-dimensionality. *BMC Bioinformatics* **20(1)**:339 DOI 10.1186/s12859-019-2929-8.

**Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, Hinoue T, Laird PW, Hoadley KA, Akbani R, Castro MAA, Gibb EA, Kanchi RS, Gordenin DA, Shukla SA, Sanchez-Vega F, Hansel DE, Czerniak BA, Reuter VE, Su X, de Sa Carvalho B, Chagas VS, Mungall KL, Sadeghi S, Pedamallu CS, Lu Y, Klimczak LJ, Zhang J, Choo C, Ojesina AI, Bullman S, Leraas KM, Lichtenberg TM, Wu CJ, Schultz N, Getz G, Meyerson M, Mills GB, McConkey DJ, Weinstein JN, Kwiatkowski DJ, Lerner SP, Akbani R, Al-Ahmadie H, Albert M, Alexopoulou I, Ally A, Antic T, Aron M, Balasundaram M, Bartlett J, Baylin SB, Beaver A, Bellmunt J, Birol I, Boice L, Bootwalla MS, Bowen J, Bowlby R, Brooks D, Broom BM, Bshara W, Bullman S, Burks E, Cárcano FM, Carlsen R, Carvalho BS, Carvalho AL, Castle EP, Castro MAA, Castro P, Catto JW, Chagas VS, Cherniack AD, Chesla DW, Choo C, Chuah E, Chudamani S, Cortessis VK, Cottingham SL, Crain D, Curley E, Czerniak BA, Daneshmand S, Demchok JA, Dhalla N, Djaladat H, Eckman J, Egea SC, Engel J, Felau I, Ferguson ML, Gardner J, Gastier-Foster JM, Gerken M, Getz G, Gibb EA, Gomez-Fernandez CR, Gordenin DA, Guo G, Hansel DE, Harr J, Hartmann A, Herbert LM, Hinoue T, Ho TH, Hoadley KA, Holt RA, Hutter CM, Jones SJM, Jorda M, Kahnoski RJ, Kanchi RS, Kasaian K, Kim J, Klimczak LJ, Kwiatkowski DJ, Lai PH, Laird PW, Lane BR, Leraas KM, Lerner SP, Lichtenberg TM, Liu J, Lolla L, Lotan Y, Lu Y, Lucchesi FR, Ma Y, Machado RD, Maglinte DT, Mallery D, Marra MA, Martin SE, Mayo M, McConkey DJ, Meraney A, Meyerson M, Mills GB, Moinzadeh A, Moore RA, Mora Pinero EM, Morris S, Morrison C, Mungall KL, Mungall AJ, Myers JB, Naresh R, O'Donnell PH, Ojesina AI, Parekh DJ, Parfitt J, Paulauskis JD, Sekhar Pedamallu C, Penny RJ, Pihl T, Porten S, Quintero-Aguilo ME, Ramirez NC, Rathmell WK, Reuter VE, Rieger-Christ K, Robertson AG, Sadeghi S, Saller C, Salner A, Sanchez-Vega F, Sandusky G, Scapulatempo-Neto C, Schein JE, Schuckman AK, Schultz N, Shelton C, Shelton T, Shukla SA, Simko J, Singh P, Sipahimalani P, Smith ND, Sofia HJ, Sorcini A, Stanton ML, Steinberg GD, Stoehr R, Su X, Sullivan T, Sun Q, Tam A, Tarnuzzer R, Tarvin K, Taubert H, Thiessen N, Thorne L, Tse K, Tucker K, Van Den Berg DJ, van Kessel KE, Wach S, Wan Y, Wang Z, et al. 2017a.** Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* **171(3)**:540–556 DOI 10.1016/j.cell.2017.09.007.

**Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, Hess JM, Uzunangelov V, Walter V, Danilova L, Lichtenberg TM, Kucherlapati M, Kimes PK, Tang M, Penson A, Babur O, Akbani R, Bristow CA, Hoadley KA, Iype L, Chang MT, Cherniack AD, Benz C, Mills GB, Verhaak RGW, Griewank KG, Felau I, Zenklusen JC, Gershenwald JE, Schoenfield L, Lazar AJ, Abdel-Rahman MH, Roman-Roman S, Stern M-H, Cebulla CM, Williams MD, Jager MJ, Coupland SE, Esmaeli B, Kandoth C, Woodman SE, Abdel-Rahman MH, Akbani R, Ally A, Auman JT, Babur O, Balasundaram M, Balu S, Benz C, Beroukhim R, Birol I, Bodenheimer T, Bowen J, Bowlby R, Bristow CA, Brooks D, Carlsen R, Cebulla CM, Chang MT, Cherniack AD, Chin L, Cho J, Chuah E, Chudamani S, Cibulskis C, Cibulskis K, Cope L, Coupland SE, Danilova L, Defreitas T, Demchok JA, Desjardins L, Dhalla N, Esmaeli B,**

**Felau I, Ferguson ML, Frazer S, Gabriel SB, Gastier-Foster JM, Gehlenborg N, Gerken M, Gershenwald JE, Getz G, Gibb EA, Griewank KG, Grimm EA, Hayes DN, Hegde AM, Heiman DI, Helsel C, Hess JM, Hoadley KA, Hobensack S, Holt RA, Hoyle AP, Hu X, Hutter CM, Jager MJ, Jefferys SR, Jones CD, Jones SJM, Kandoth C, Kasaian K, Kim J, Kimes PK, Kucherlapati M, Kucherlapati R, Lander E, Lawrence MS, Lazar AJ, Lee S, Leraas KM, Lichtenberg TM, Lin P, Liu J, Liu W, Lolla L, Lu Y, Iype L, Ma Y, Mahadeshwar HS, Mariani O, Marra MA, Mayo M, Meier S, Meng S, Meyerson M, Mieczkowski PA, Mills GB, Moore RA, Mose LE, Mungall AJ, Mungall KL, Murray BA, Naresh R, Noble MS, Oba J, Pantazi A, Parfenov M, Park PJ, Parker JS, Penson A, Perou CM, Pihl T, Pilarski R, Protopopov A, Radenbaugh A, Rai K, Ramirez NC, Ren X, Reynolds SM, Roach J, Robertson AG, Roman-Roman S, Roszik J, Sadeghi S, Saksena G, Sastre X, Schadendorf D, Schein JE, Schoenfield L, Schumacher SE, Seidman J, Seth S, Sethi G, Sheth M, Shi Y, Shields C, Shih J, Shmulevich I, Simons JV, Singh AD, Sipahimalani P, Skelly T, Sofia H, Soloway MG, Song X, Stern M-H, Stuart J, Sun Q, Sun H, Tam A, Tan D, Tang M, Tang J, Tarnuzzer R, Taylor BS, Thiessen N, Thorsson V, Tse K, Uzunangelov V, Veluvolu U, Verhaak RGW, Voet D, Walter V, Wan Y, Wang Z, Weinstein JN, Wilkerson MD, Williams MD, et al. 2017b.** Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer cell* **32(2)**:204–220 DOI 10.1016/j.ccell.2017.07.003.

**Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. 2015.** ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)* **115(3)**:211–252 DOI 10.1007/s11263-015-0816-y.

**Rykunov D, Beckmann ND, Li H, Uzilov A, Schadt EE, Reva B. 2016.** A new molecular signature method for prediction of driver cancer pathways from transcriptional data. *Nucleic Acids Research* **44(11)**:e110.

**Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, Van Arnam J, Shmulevich I, Rao AUK, Lazar AJ, Sharma A, Vésteinn T, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J,**

**Korchina V, Lee S, et al. 2018.** Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Reports* **23(1)**:181–193.

**Salvucci M, Würstle ML, Morgan C, Curry S, Cremona M, Lindner AU, Bacon O, Resler AJ, Murphy ÁC, O'Byrne R, Flanagan L, Dasgupta S, Rice N, Pilati C, Zink E, Schöller LM, Toomey S, Lawler M, Johnston PG, Wilson R, Camilleri-Broët S, Salto-Tellez M, McNamara DA, Kay EW, Laurent-Puig P, Van Schaeybroeck S, Hennessy BT, Longley DB, Rehm M, Prehn JHM. 2017.** A stepwise integrated approach to personalized risk predictions in stage iii colorectal cancer. *Clinical Cancer Research* **23(5)**:1200–1212 DOI 10.1158/1078-0432.CCR-16-1084.

**Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, Chakravarty D, Daian F, Gao Q, Bailey MH, W-W Liang W-W, Foltz SM, Shmulevich I, Ding L, Heins Z, Ochoa A, Gross B, Gao J, Zhang H, Kundra R, Kandoth C, Bahceci I, Dervishi L, Dogrusoz U, Zhou W, Shen H, Laird P, Way GP, Greene CS, Liang H, Xiao Y, Wang C, Iavarone A, Berger AH, Bivona TG, Lazar AJ, Hammer GD, Giordano T, Kwong LN, McArthur G, Huang C, Tward AD, Frederick MJ, McCormick F, Meyerson M, Van Allen EM, Cherniack AD, Ciriello G, Sander C, Schultz N, Cancer Genome Atlas Research Network. 2018.** Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173(2)**:321–337.

**Schaub FX, Dhankani V, Berger AC, Trivedi M, Richardson AB, Shaw R, Zhao W, Zhang X, Ventura A, Liu Y, Ayer DE, Hurlin PJ, Cherniack AD, Eisenman RN, Bernard B, Grandori C, Network Cancer Genome Atlas. 2018.** Pan-cancer alterations of the myc oncogene and its proximal network across The Cancer Genome Atlas. *Cell Systems* **6(3)**:282–300.

**Seoane JA, Day INM, Gaunt TR, Campbell C. 2013.** A pathway-based data integration framework for prediction of disease progression. *Bioinformatics* **30(6)**:838–845.

**Shen H, Shih J, Hollern DP, Wang L, Bowlby R, Tickoo SK, Thorsson V, Mungall AJ, Newton Y, Hegde AM, Armenia J, Sánchez-Vega F, Pluta J, Pyle LC, Mehra R, Reuter VE, Godoy G, Jones J, Shelley CS, Feldman DR, Vidal DO, Lessel D, Kulis T, Cárcano FM, Leraas KM, Lichtenberg TM, Brooks D, Cherniack AD, Cho J, Heiman DI, Kasaian K, Liu M, Noble MS, Xi L, Zhang H, Zhou W, ZenKlusen JC, Hutter CM, Felau I, Zhang J, Schultz N, Getz G, Meyerson M, Stuart JM, Akbani R, Wheeler DA, Laird PW, Nathanson KL, Cortessis VK, Hoadley KA, Cancer Genome Atlas Research Network. 2018.** Integrated molecular characterization of testicular germ cell tumors. *Cell Reports* **23(11)**:3392–3406.

**Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, Ladanyi M, Sander C. 2012.** Integrative subtype discovery in glioblastoma using icluster. *PLOS ONE* **7(4)**:e35236.

**Shen R, Olshen AB, Ladanyi M. 2009.** Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25(22)**:2906–2912 DOI 10.1093/bioinformatics/btp543.

**Sherafatian M. 2018.** Tree-based machine learning algorithms identified minimal set of mirna biomarkers for breast cancer diagnosis and molecular subtyping. *Gene* **677(2)**:111–118 DOI 10.1016/j.gene.2018.07.057.

**Srivastava S, Wang W, Manyam G. 2013.** et al.Integrating multi-platform genomic data using hierarchical bayesian relevance vector machines. *EURASIP Journal on Bioinformatics and Systems Biology* **2013(1)**:9 DOI 10.1186/1687-4153-2013-9.

**Stephen RP, Lewis JF. 2013.** Clinical and molecular models of glioblastoma multiforme survival. *International Journal of Data Mining and Bioinformatics* **7(3)**:245–265 DOI 10.1504/IJDMB.2013.053310.

**Sun D, Chen J, Liu L, Zhao G, Dong P, Wu B, Wang J, Dong L. 2018a.** Establishment of a 12-gene expression signature to predict colon cancer prognosis. *PeerJ* **6**:e4942.

**Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, Verlingue Lïc, Brandao D, Lancia A, Ammari S, Hollebecque A, Scoazec J-Y, Marabelle A, Massard C, Soria J-C, Robert C, Paragios N, Deutsch E, Ferté C. 2018b.** A radiomics approach to assess tumour-infiltrating cd8 cells and response to anti-pd-1 or anti-pd-l1 immunotherapy: an imaging biomarker, retrospective multicohort study. *The Lancet Oncology* **19(9)**:1180–1191 DOI 10.1016/S1470-2045(18)30413-3.

**Sutton EJ, Huang EP, Drukker K, Burnside ES, Li H, Net JM, Rao A, Whitman GJ, Zuley M, Ganott M, Bonaccio E, Giger ML, Morris EA, On behalf of the TCGA group. 2017.** Breast mri radiomics: comparison of computer-and human-extracted imaging phenotypes. *European Radiology Experimental* **1(1)**:22 DOI 10.1186/s41747-017-0025-2.

**Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, Lazar AJ, Cherniack AD, Beroukhim R, Meyerson M, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, et al. 2018.** Genomic and functional approaches to understanding cancer aneuploidy. *Cancer cell* **33(4)**:676–689.

**The Cancer Genome Atlas Network. 2012a.** Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487(7407)**:330–337 DOI 10.1038/nature11252.

**The Cancer Genome Atlas Network. 2012b.** Comprehensive molecular portraits of human breast tumours. *Nature* **490(7418)**:61–70 DOI 10.1038/nature11412.

**The Cancer Genome Atlas Network. 2015.** Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517(7536)**:576–582 DOI 10.1038/nature14129.

**The Cancer Genome Atlas Research Network. 2013.** Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine* **368(22)**:2059–2074 DOI 10.1056/NEJMoa1301689.

**The Cancer Genome Atlas Research Network. 2015.** Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine* **372(26)**:2481–2498 DOI 10.1056/NEJMoa1402121.

**The Cancer Genome Atlas Research Network. 2016.** Comprehensive molecular characterization of papillary renal-cell carcinoma. *New England Journal of Medicine* **374(2)**:135–145 DOI 10.1056/NEJMoa1505917.

**The Cancer Genome Atlas Research Network. 2008.** Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455(7216)**:1061–1068 DOI 10.1038/nature07385.

**The Cancer Genome Atlas Research Network. 2011.** Integrated genomic analyses of ovarian carcinoma. *Nature* **474(7353)**:609–615 DOI 10.1038/nature10166.

**The Cancer Genome Atlas Research Network. 2012c.** Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489(7417)**:519–525 DOI 10.1038/nature11404.

**The Cancer Genome Atlas Research Network. 2013.** Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499(7456)**:43–49 DOI 10.1038/nature12222.

**The Cancer Genome Atlas Research Network. 2014a.** Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513(7517)**:202–209 DOI 10.1038/nature13480.

**The Cancer Genome Atlas Research Network. 2014b.** Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507(7492)**:315–322 DOI 10.1038/nature12965.

**The Cancer Genome Atlas Research Network. 2014c.** Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511(7511)**:543–550 DOI 10.1038/nature13385.

**The Cancer Genome Atlas Research Network. 2017a.** Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* **171(4)**:950–965.e28 DOI 10.1016/j.cell.2017.10.014.

**The Cancer Genome Atlas Research Network. 2017b.** Integrated genomic and molecular characterization of cervical cancer. *Nature* **543(7645)**:378–384 DOI 10.1038/nature21386.

**The Cancer Genome Atlas Research Network. 2017c.** Integrated genomic characterization of oesophageal carcinoma. *Nature* **541(7636)**:169–175 DOI 10.1038/nature20805.

**Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Yang T-HO, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Mokrab Y, Newman AM, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedamallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CS, Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich I, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M,**

**Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Laird PW, Shen H, et al. 2018.** The immune landscape of cancer. *Immunity* **48(4)**:812–830.

**Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN. 2010.** Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell* **17(1)**:98–110 DOI 10.1016/j.ccr.2009.12.020.

**Vural S, Wang X, Guda C. 2016.** Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Systems Biology* **10(3)**:62 DOI 10.1186/s12918-016-0306-z.

**Wang C, Liang C. 2018.** Msipred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine. *Scientific Reports* **8(1)**:17546 DOI 10.1038/s41598-018-35682-z.

**Wang X, Han L, Zhou L, Wang L, Zhang L-M. 2018.** Prediction of candidate rna signatures for recurrent ovarian cancer prognosis by the construction of an integrated competing endogenous rna network. *Oncology Reports* **40(5)**:2659–2673 DOI 10.3892/or.2018.6707.

**Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, Sander C, Cherniack AD, Mina M, Ciriello G, Schultz N, Sanchez Y, Greene CS, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V,**

**Lee S, Lewis L, Li W, Liu X, et al. 2018.** Machine learning detects pan-cancer ras pathway activation in The Cancer Genome Atlas. *Cell Reports* **23(1)**:172–180.

**Wei D. 2018.** A multigene support vector machine predictor for metastasis of cutaneous melanoma. *Molecular Medicine Reports* **17(2)**:2907–2914.

**Wen J-X, Li X-Q, Chang Y. 2018.** Signature gene identification of cancer occurrence and pattern recognition. *Journal of Computational Biology* **25(8)**:907–916 DOI 10.1089/cmb.2017.0261.

**Wilop S, Chou W-C, Jost E, Crysandt M, Panse J, Chuang M-K, Brümmendorf TH, Wagner W, Tien H-F, Kharabi Masouleh B. 2016.** A three-gene expression-based risk score can refine the european leukemianet aml classification. *Journal of Hematology & Oncology* **9(1)**:78 DOI 10.1186/s13045-016-0308-8.

**Wong KK, Rostomily R, Wong ST. 2019.** Prognostic gene discovery in glioblastoma patients using deep learning. *Cancers* **11(1)**:53 DOI 10.3390/cancers11010053.

**Xie H, Xu H, Hou Y, Cai Y, Rong Z, Song W, Wang W, Li K. 2019.** Integrative prognostic subtype discovery in high-grade serous ovarian cancer. *Journal of Cellular Biochemistry* **120(11)**:18659–18666 DOI 10.1002/jcb.29049.

**Xu G, Zhang M, Zhu H, Xu J. 2017.** A 15-gene signature for prediction of colon cancer recurrence and prognosis based on svm. *Gene* **604**:33–40 DOI 10.1016/j.gene.2016.12.016.

**Yang S, Xu J, Zeng X. 2018.** A six-long non-coding rna signature predicts prognosis in melanoma patients. *International Journal of Oncology* **52(4)**:1178–1188 DOI 10.3892/ijo.2018.4268.

**Yang W, Yoshigoe K, Qin X, Liu JS, Yang JY, Niemierko A, Deng Y, Liu Y, Dunker AK, Chen Z, Wang L, Xu D, Arabnia HR, Tong W, Yang MQ. 2014.** Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinformatics* **15(17)**:S2 DOI 10.1186/1471-2105-15-S17-S2.

**Yasser E-M, Hsieh T-Y, Shivakumar M, Kim D, Honavar V. 2018.** Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Medical Genomics* **11(3)**:71.

**Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M. 2016.** Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications* **7**:12474.

**Zhang Y, Li A, Peng C, Wang M. 2016.** Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13(5)**:825–835 DOI 10.1109/TCBB.2016.2551745.

**Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, Lerario AM, Else T, Knijnenburg TA, Ciriello G, Kim S, Assie G, Morozova O, Akbani R, Shih J, Hoadley KA, Choueiri TK, Waldmann J, Mete O, Robertson AG, Wu H-T, Raphael B J, Shao L, Meyerson M, Demeure MJ, Beuschlein F, Gill AJ, Sidhu S B, Almeida M Q, Fragoso MCBV, Cope LM, Kebebew E, Habra MA, Whitsett TG, Bussey KJ, Rainey WE, Asa SL, Bertherat J, Fassnacht M, Wheeler DA, Hammer GD, Giordano TJ, Verhaak RGW, Zheng S, Verhaak RGW, Giordano TJ, Hammer GD, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, Lerario AM, Else T, Knijnenburg TA, Ciriello G, Kim S, Assié G, Morozova O, Akbani R, Shih J, Hoadley KA, Choueiri TK, Waldmann J, Mete O, Robertson AG, Wu H-T, Raphael BJ, Meyerson M, Demeure MJ, Beuschlein F, Gill AJ, Sidhu SB, Almeida M, Barisson Fragoso MC, Cope LM, Kebebew E, Habra MA, Whitsett TG, Bussey KJ, Rainey WE, Asa SL, Bertherat J, Fassnacht M, Wheeler DA, Benz C, Ally A, Balasundaram M, Bowlby R, Brooks D, Butterfield YSN, Carlsen R, Dhalla N, Guin R, Holt RA, Jones SJM, Kasaian K, Lee D, Li HI, Lim L, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Park PJ, Kroiss M, Gao J, Sander C, Schultz N, Jones**

**CD, Kucherlapati R, Mieczkowski PA, Parker JS, Perou CM, Tan D, Veluvolu U, Wilkerson MD, Hayes DN, Ladanyi M, Quinkler M, Auman JT, Latronico AC, Mendonca BB, Sibony M, Sanborn Z, Bellair M, Buhay C, Covington K, Dahdouli M, Dinh H, Doddapaneni H, Downs B, Drummond J, Gibbs R, Hale W, Han Y, Hawes A, Hu J, Kakkar N, Kalra D, Khan Z, Kovar C, Lee S, Lewis L, Morgan M, Morton D, Muzny D, Santibanez J, Xi L, Dousset B, Groussin L, Libé R, Chin L, Reynolds S, Shmulevich I, Chudamani S, Liu J, Lolla L, Wu Y, Yeh JJ, Balu S, Bodenheimer T, Hoyle AP, Jefferys SR, Meng S, Mose LE, Shi Y, Simons JV, Soloway MG, Wu J, Zhang W, Mills Shaw KR, Demchok JA, Felau I, Sheth M, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Davidsen T, Crawford C, Hutter CM, Sofia HJ, Roach J, Bshara W, Gaudioso C, Morrison C, Soon P, Alonso S, Baboud J, Pihl T, et al. 2016.** Comprehensive pan-genomic characterization of adrenocortical carcinoma. *Cancer Cell* **29(5)**:723–736 DOI 10.1016/j.ccell.2016.04.002.

**Zhou J, Li L, Wang L, Li X, Xing H, Cheng L. 2018.** Establishment of a svm classifier to predict recurrence of ovarian cancer. *Molecular Medicine Reports* **18(4)**:3589–3598.

## 3.2 Búsqueda de nuevas dianas terapéuticas para el tratamiento del cáncer de colon

Este artículo se desarrolla con el fin de identificar nuevos biomarcadores y/o *pathways* alterados que ayuden a la estratificación de pacientes con el fin de buscar tratamientos específicos. A partir de la clasificación de artículos presentados en la sección 3.1, se realizó una búsqueda de firmas de expresión genética en cáncer de colon previamente publicadas. Se identificaron tres firmas que cumplían con los criterios de inclusión: generadas a partir de la cohorte del TCGA, mediante técnicas de selección de características y con valor prognóstico en cáncer de colon. Las tres firmas fueron combinadas en una única, denominada *meta-signature*. Siguiendo una metodología robusta de ML, establecida tras el estudio en [1], se validó la *meta-signature* en diferentes problemas de clasificación: predicción del *N stage*, del *tumor stage* y de la presencia de tumor. Se utilizan los datos de la cohorte TCGA-COAD para el desarrollo experimental.

La *meta-signature* presenta rendimientos muy altos para diferenciar tejido tumoral de tejido adyacente normal. Con el fin de identificar nuevos biomarcadores y dianas terapéuticas, se realiza un reposicionamiento de fármacos contra los productos de los genes de la *meta-signature*. El reposicionamiento de fármacos se realiza mediante técnicas de *Docking Molecular*. Gracias a esta técnica, se estudia la interacción *in silico* de 81 fármacos anti-tumorales aprobados por la FDA con los productos de los genes dentro de la *meta-signature*.

Se identificaron cuatro interacciones de interés: GLTP - Nilotinib, PTPRN - Venetoclax, VEGFA - Venetoclax y FABP6 - Abemaciclib. La ausencia de expresión en tejido sano adyacente, la alta interacción con la droga y su rol metabólico hacen que FABP6 sea un potencial candidato a ser considerado diana terapéutica.

El reposicionamiento de fármacos, de esta manera, es capaz de abaratar enormemente los costes en la fase de desarrollo experimental. También cabe resaltar que el reposicionamiento de fármacos anti-tumorales presentan ciertas ventajas. A diferencia del reposicionamiento de otras sustancias como pueden ser las estatinas, los bloqueadores del receptor de la angiotensina ó la aspirina, los fármacos que han sido aprobados ya para el tratamiento de otros tumores funcionan como terapias no combinadas. Por lo tanto, el estudio de la efectividad y desarrollo será menos tedioso en estos casos.

Fui el autor principal del artículo. Realicé el diseño experimental, ejecuté la metodología de ML, hice la selección de fármacos y la búsqueda de las firmas genéticas. También fui el que redactó el artículo y creó las tablas y las figuras. La información de participación de los otros autores está disponible en el artículo.

Liñares-Blanco *et al. BMC Molecular and Cell Biology* (2020) 21:52
https://doi.org/10.1186/s12860-020-00295-w

BMC Molecular and Cell Biology

RESEARCH ARTICLE Open Access

# Molecular docking and machine learning analysis of Abemaciclib in colon cancer

Jose Liñares-Blanco[1], Cristian R. Munteanu[1,2], Alejandro Pazos[1,2] and Carlos Fernandez-Lozano[1,2*]

## Abstract

**Background:** The main challenge  in cancer research is the identification of different omic variables that present a prognostic value and personalised diagnosis for each tumour. The fact that the diagnosis is personalised opens the doors to the design and discovery of new specific treatments for each patient. In this context, this work offers new ways to reuse existing databases and work to create added value in research. Three published signatures with significante prognostic value in Colon Adenocarcinoma (COAD) were indentified. These signatures were combined in a new meta-signature and validated with main Machine Learning (ML) and conventional statistical techniques. In addition, a drug repurposing experiment was carried out through Molecular Docking (MD) methodology in order to identify new potential treatments in COAD.

**Results:** The prognostic potential of the signature was validated by means of ML algorithms and differential gene expression analysis. The results obtained supported the possibility that this meta-signature could harbor genes of interest for the prognosis and treatment of COAD. We studied drug repurposing following a molecular docking (MD) analysis, where the different protein data bank (PDB) structures of the genes of the meta-signature (in total 155) were confronted with 81 anti-cancer drugs approved by the FDA. We observed four interactions of interest: GLTP - Nilotinib, PTPRN - Venetoclax, VEGFA - Venetoclax and FABP6 - Abemaciclib. The FABP6 gene and its role within different metabolic pathways were studied in tumour and normal tissue and we observed the capability of the FABP6 gene to be a therapeutic target. Our in silico results showed a significant specificity of the union of the protein products of the FABP6 gene as well as the known action of Abemaciclib as an inhibitor of the CDK4/6 protein and therefore, of the cell cycle.

**Conclusions:** The results of our ML and differential expression experiments have first shown the FABP6 gene as a possible  new cancer biomarker due to its specificity in colonic tumour tissue and no expression in healthy adjacent tissue. Next, the MD analysis showed that the drug Abemaciclib characteristic affinity for the different protein

(Continued on next page)

*Correspondence: carlos.fernandez@udc.es
[1] Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, CITIC, Campus Elviña s/n, 15071 A Coruña, Spain
[2] Grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos. Imagen Médica y Diagnóstico Radiológico (RNASA-IMEDIR). Instituto de Investigación Biomédica de A Coruña (INIBIC). Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. Universidade da Coruña (UDC), Xubias de arriba, 84, 15006 A Coruña, Spain

(Continued from previous page)
structures of the FABP6 gene. Therefore, in silico experiments have shown a new opportunity that should be validated experimentally, thus helping to reduce the cost and speed of drug screening. For these reasons, we propose the validation of the drug Abemaciclib for the treatment of colon cancer.

**Keywords:** Machine learning, Molecular docking, Colon cancer, Prognosis, Drug repurposing, FABP6, Abemaciclib, TCGA

## Background

Colon adenocarcinomas (COAD) significantly contributes to mortality and morbidity [1] of cancer in the world population. In 2018, of the approximately 18 million new cases, about 10% were colorectal cancer (1.8 million cases), according to data from the World Cancer Research Fund. This type of cancer gains significance when we focus on data within Spain, as it is the primary cause of hospital stay in the country. It is estimated that by 2019 there will be around 44,000 new cases of COAD. Moreover, this problem is also alarming in Galicia, which has the fifth highest number of COAD cases, with 2,500 new cases each year according to data from the Spanish Cancer Association [2]. All studies indicate that early detection and targeted treatment are the best weapons to reduce these devastating statistics.

To achieve this goal, the scientific mass is, on the one hand, generating different types of omics data to define diseases molecularly and, on the other hand, designing different data analysis models to extract valuable information from these data.

In such context, extensive scientific contributions have based their research on data reported by international initiatives such as The Cancer Genome Atlas (TCGA) [3]. The TCGA was born with the objective of obtaining a multidimensional genomic map of the main genomic changes in a wide variety of tumours. Analysis of the data hosted by the TCGA offers scientists new opportunities to obtain highly reproducible results that can be extrapolated to most of the world's populations. With a sample size of over 11,000 patients categorised into 33 different tumour types, this repository offers the possibility of creating models sufficiently robust for the extraction of statistically reliable results and conclusions.

With access to such an amount of data, an ideal environment is created for the use of new computational methods capable of extracting information from the data and simulating complex biological processes. Computational methods such as machine learning (ML) and molecular docking (MD) are examples that can provide new and different visions in the fight against complex diseases, such as COAD. Both techniques have already been used extensively in recent years to bring about new results and conclusions [4–9].

As far as therapeutic targets are concerned, an immense investment is being made in terms of time, personnel, resources and money, to experimentally validate new biological targets and new drugs that act efficiently on them. Once the drug to be validated has been identified, experiments are carried out to test and validate whether the drug has the expected effect on cellular and animal models. Subsequently, a clinical trial is necessary, where a significant number of patients must be recruited, all adverse aspects must be analysed and all quality controls must be passed. It is here, in the clinical trials phase, that budgets skyrocket. Therefore, it is necessary to pre-screen interactions, in a in silico way, to obtain potential candidates that can be tested later in the laboratory. This prior in silico step greatly helps to reduce experimental costs. It should be noted that one out of every 5,000 drugs goes to the clinic [10], which implies a disproportionate investment on the part of the pharmaceutical companies. Therefore, finding pathways and shortcuts from basic research to clinical research through translational research would offer a significant advantage to this field of research.

With this in mind, the present work has used public data from TCGA and has employed the latest ML and MD techniques to predict new biomarkers in COAD and simulate the effect of drugs already approved for repurposing to this type of tumour.

For some years now, promising results have been reported on the effect of several drugs on diseases for which they were not designed. For example, the drug Zidovudine, which was originally designed for the treatment of cancer, has subsequently been used for HIV/AIDS. Another example is Rituximab, which had originally been indicated for various types of cancers and has subsequently been approved for rheumatoid arthritis, or Raloxifene, which went from being used for osteoporosis to being used for breast cancer [11]. Experimentally testing all the drugs used today in all diseases is unfeasible. Fortunately, with the increase in computational capacity, techniques of drug repurposing can give a realistic approximation of what might occur in nature.

Both the results obtained by previous researchers reported in the scientific literature and the results of the in silico analysis of the present work seem to coincide that the FABP6 gene presents all the necessary characteristics to be proposed as a potential candidate for an early diagnostic marker in COAD patients. Moreover, the results of MD indicate a strong interaction between the drug Abemacicib and the different protein conformations of the FABP6 gene, leading to a possible inhibition of the protein activity.

It is known that FABP6 belongs to a group of low-molecular-weight proteins related to the transport of long-chain bioactive fatty acids in cells. In humans, there are nine different subgroups (FABP1-9). This group of proteins play a role in the development of different types of cancer cells [12–16], and have also been proposed as diagnostic markers and therapeutic targets [17–19]. Specifically, FABP6 is highly expressed in the ileum and is an intracellular transporter of bile acids in ileal epithelial cells, contributing to the catalysis and metabolism of cholesterol. In relation to COAD cells, previous works have observed that there are high concentrations of faecal bile acids, in particular, secondary bile acids [20–22]. Furthermore, the involvement of FABP6 in the development of colon cancer has been addressed in previous publications [23, 24].

This work presents two distinct phases. Firstly, a search has been carried out for previously published gene signatures obtained from TCGA data using ML algorithms. The use of these algorithms for this type of problem offers the possibility of finding patterns and identifying important variables that have not been identified with the classical techniques. Thus, after a thorough review of the different papers published under these requirements, three gene signatures with prognostic value have been identified for colon cancer [25–27]. Secondly, once our meta-signature was created, the objective was to search for and identify those genes that could behave as therapeutic targets in colon cancer and to carry out an in-depth study for their validation and contribution to new treatment approaches, in a in silico way. To this end, a repurposing of anti-cancer drugs, already approved by the FDA for use in different types of tumour s, has been carried out. The results of this work show a strong interaction, in in silico experiments, between the PDB structures of the FABP6 gene and the drug *Abemaciclib*. An in-depth study of this interaction, which is detailed in this work, offers hopeful results on a possible new treatment against colon cancer, which must be validated experimentally.

## Results

### New gene signature for COAD prognosis prediction

The first objective of this work was to search for previously published gene signatures that could predict the prognosis of COAD. In order to do this, we opted to search for these signatures based on works using Machine Learning techniques. We hypothesised that Machine Learning techniques, together with different techniques for selecting characteristics, could find variables (in this case genes) of interest that have not previously been identified by conventional techniques such as mutation analysis or differential expression analysis.

In order to avoid biases between different cohorts, such as data normalisation, as the data have been generated on different platforms, the search for the papers focused on those that used data from the TCGA repository, mainly due to its heterogeneity and its large number of samples and results published using Machine Learning techniques.

Three papers were identified [25–27] that satisfied all the requirements: they used colon cancer data from the TCGA repository, applied some type of dimensionality reduction techniques linked to Machine Learning techniques, and reported a gene signature that was capable of predicting disease prognosis with great precision.

The three signatures published by each of the selected papers are shown in Table 1.

The three signatures were combined for later experiments, generating a meta-signature of 34 genes. None of these genes, after verification with the repository Intogen [28], was previously catalogued as a genetic driver for any type of cancer. Therefore, we considered that this set of genes could harbor some previously unidentified biomarker or therapeutic target that could predict the appearance of COAD, how the disease will develop, or inhibit its growth.

Therefore, the hypothesis of the present work is that the identification of different gene signatures, reported by different research groups, working under the same data and the same problem, may harbour genes of interest that can be used as new therapeutic targets. We will validate the signatures in different biological problems using ML techniques and conventional statistical techniques, and we will use MD techniques to identify if there is an approved drug that strongly interacts with any of the protein products of the identified genes, and coul be used for the treatment of COAD. Therefore, an experiment of repurposing drugs against these targets would be carried out to identify new therapeutic targets and their respective treatments in colon cancer.

**Table 1** Gene signatures obtained from previous works

| Work | Gene signature |
|---|---|
| Sun et al. 2018 | TREML2, PADI4, NCKIPSD, PTPRN, PGLYRP1, C5orf53,<br>TREML3, NOG, VIP, RIMKLB, NKAIN4, FAM171B |
| Xu et al. 2017 | HES5, ZNF417, GLRA2, OR8D2, HOXA7, FABP6, MUSK,<br>HTR6, GRIP2, KLRK1, VEGFA, AKAP12, RHEB, NCRNA00152, PMEPA1 |
| Wen et al. 2018 | GLTP, METTL7A, PPAP2A, CITED2, SCARA5, CDH3,<br>IL6R, PKIB, GLP2R, LINC00974, EPB41L3, NR3C2 |

Below are the results obtained after testing the hypothesis put forward in this section. Two experiments were carried out. Firstly, Machine Learning experiments for different classifications in order to clarify and validate the real predictive value of the signature obtained. Secondly, Molecular Docking experiments in order to search for candidates that could be possible therapeutic targets and the corresponding drugs that interact with them.

### Machine learning and statistical analysis

To validate whether the meta-signature obtained is capable of predicting different clinical outcomes, three types of Machine Learning experiments were designed: 1) classification of different stages of cancer with expression data; 2) classification of the metastatic stage in lymph nodes with expression data; 3) classification expression data of tumour and healthy adjacent samples.

These three experiments were mainly designed to observe the predictive potential of the meta-signature obtained. For the first two experiments, the obtained signature was compared with others to determine if there was sufficient information to predict advanced carcinogenic aspects. As for the third experiment, the aim was to observe whether the signature was capable of differentiating, in a significant way, between tumour and adjacent normal tissue, thus being able to identify specific omic variables in tumour tissue, and generate the possibility of finding new biomarkers or therapeutic targets specific to the tumour. The three experiments were compared with two algorithms and three different signatures. The Random Forest and Glmnet algorithms were trained for three different expression data subsets of the TCGA COAD cohort: 1) the 34 genes of the meta-signature obtained from the aforementioned studies, which will be the object of study in this work; 2) a random signature of 34 genes and 3) a signature that houses the genetic drivers for colon cancer, obtained from the Intogen repository.

#### *Classification of different stages of cancer*

Available data from the COAD cohort of the TCGA were downloaded. As mentioned, three different datasets were generated for each signature to study. As a dependent variable, patients were classified according to their stages. Patients were grouped into two classes (stage I-II; stage III-IV) representing the good and bad prognosis of the patients, respectively.

Figure 1 shows the results achieved by each data-algorithm binomial. The worst results were obtained with algorithms trained with random signature, as expected. As for the other two signatures, those obtained a slightly higher yield, around 2.5 points more than the best performance of the random signature. Due to the small difference in the three signatures, it can be deduced that this is an extremely complicated problem and that both the meta-signature and the genetic drivers downloaded from the Intogen repository do not present significant information about the problem to be solved. To confirm this assumption, Friedman's statistical test was performed to see if there was any significant difference between the models. A $p$-value=0.2208 indicate that no model existed that was significantly better than the others.

#### *Classification of patients according to their metastatic stage in lymph nodes*

The next problem addressed was the prediction of the metastatic stage of patients. In the same way as the previous one, three different datasets were created with the same signatures of the previous problem. In this case, the patients were classified into two groups according to their metastatic stage (N0 and N1-3). This problem was established to obtain information about the very early metastasis development. To date, there is still great uncertainty about the omic variables involved in the process of metastasis, so it is also considered a very complex problem to solve.

In this case, Fig. 2 shows that the signature containing the genetic drivers is superior to the other two signatures. In this graph, we can deduce that our meta-signature does not contain any type of useful information for discerning different metastatic stages, since it has yields even lower than the random signature in the training with certain algorithms. After performing Friedman's statistical test, a $p$-value=0.27 was obtained, indicating that no model was significantly better than the others.

#### *Classification of tumour and adjacent healthy tissue*

In the same way as in the previous experiments, three datasets were obtained corresponding to the three signatures used. In this case, the samples were classified between tumour samples and healthy adjacent tissue samples.

Unlike the previous experiments, we observed in Fig. 3a a greater general precision in this problem. The dataset formed by the genetic drivers presented an almost perfect performance in both algorithms. As for the random signature, its performance dropped considerably. It was also unstable and irregular between the algorithms used and presented a randomness of the results. On the other hand, the meta-signature studied in this work presented a perfect prediction of the samples, surpassing in performance to the signature presented by the genetic drivers. Again, Friedman's statistical test was performed to observe if there were significant differences between the models. The $p$-value of the test was significant, with a value of $2.2e^{-16}$. Because the test was significant, a multiple comparison test was performed to see which models had a significant difference. The PostHoc Friedman Conover Test variant was used. Assuming a significance level of

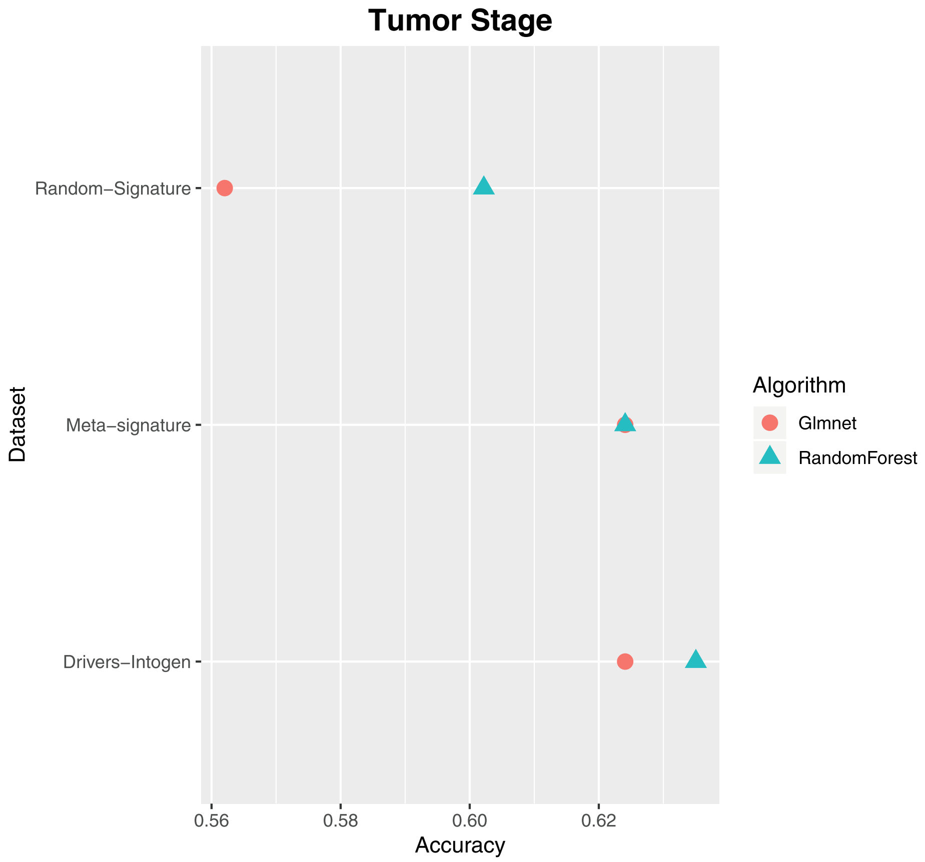


**Fig. 1** Classification according to the stage of the patients. A comparative experiment was carried out with different datasets and different algorithms for the classification of patients according to their stage. The classification consisted of a binary classification, grouping the patients in two classes (stage I-II and stage III-IV)

0.01, it was determined that the datasets conformed by the meta-signature and the genetic drivers were significantly better than the dataset formed by the random signature. As for the comparison between the first two, there were no significant differences in performance.

We can infer, therefore, that the meta-signature obtained is useful when differentiating patients in a very early stage of the tumour. It is interesting to know at this point of the analysis which of the genes presented a greater weight within the model and a greater discriminatory capacity in the classification between healthy and tumour tissue.

For a further study of the models that were trained with the meta-signature, the importance of the variables in the Random Forest and Glmnet models were extracted. The importance of the variables (standardized to have a mean of zero and a standard deviation of one) within the Glmnet and Random Forest models is shown in Fig. 3c. In addition, Table 2 shows, in descending order, the top 15 most important genes obtained in both algorithms. A differential analysis of gene expression using the package edgeR [29] was also performed on the datasets that presented the variables of the meta-signature. The results obtained in this analysis were compared with those obtained in the ML models (see Fig. 3c and Table 2). In addition, in Fig. 3b, a graph of differential expression obtained by means of the classical approximation is observed. The figure shows how this approximation detected the FABP6 and CDH3 genes as the most significant according to the log fold change. This conventional approach models the data under a negative binomial distribution, calculates the overdispersion coefficient, and performs the exact Fischer Test to obtain the most significant variables. Figure 3d shows a Venn diagram with the coincidences of the three approximations, and it can be seen that the three approximations reach very similar conclusions.

The results obtained in the analysis of the importance of the variables indicate that the two algorithms and the classical approximation reached almost the same conclusions, and gave importance to the same variables (genes), although there is a small degree of variability. Specifically, among the top 15 variables (genes)

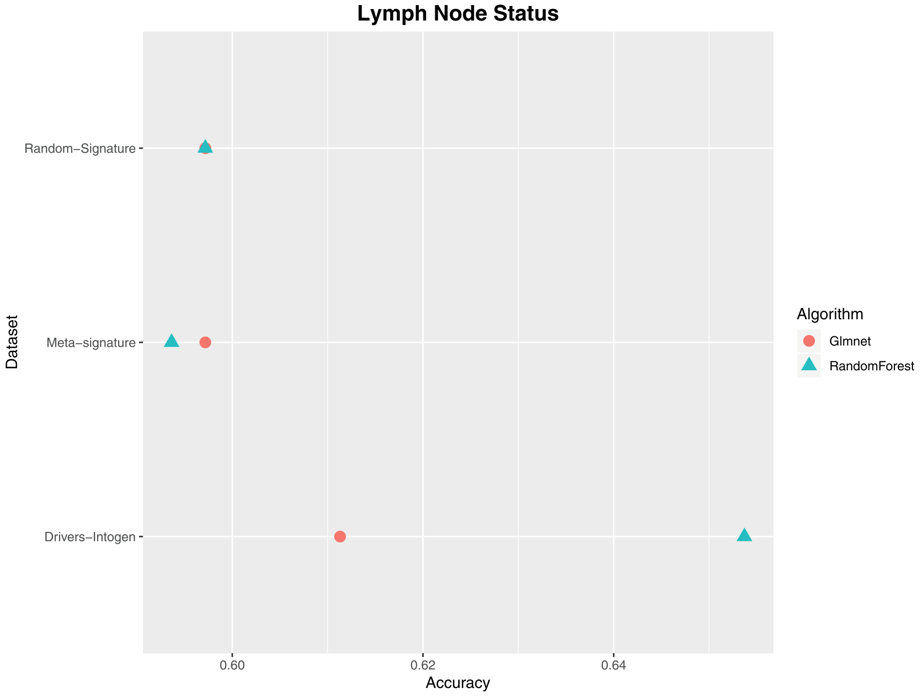


**Fig. 2** Classification by metastatic stage of patients. A comparative experiment was carried out with different datasets and different algorithms for the classification of patients according to their metastatic stage of lymphatic node. The classification consisted of a binary classification, grouping patients into two classes (stage n0 and stage 1-3)

of the three approximations, there was coincidence in 80% of them. All of them agree that the genes CDH3, MUSK, SCARA5, NR3C2, GLP2R, EPB41L3, PKIB, IL6R, METTL7A, VEGFA, FABP6 and VIP have great importance when differentiating between healthy samples and tumour samples.

By defining this meta-signature as a predictor in the diagnosis of the disease, and after the results obtained in the different approximations, it is theoretically possible that there exists in this meta-signature, a gene that may have an important role in the development of the tumour and may be a new target for future treatments. For this reason, a molecular affinity study was carried out between the different protein products of these genes and anti-cancer drugs that have been previously approved. In this way, a more in-depth study can be carried out on the results obtained and new specific therapeutic targets for colon cancer can be proposed.

### Molecular docking - drug repurposing

At this point in the work, we found a meta-signature of genes that was able to classify with great precision, healthy and colon cancer tissues. Another important aspect was the importance of the variables in the different models. The three approaches (Random Forest, Glmnet and edgeR) showed great coincidence in the most significant variables, indicating that these genes could have an important role in the disease. In this context, we consider it necessary to conduct an experiment in silico to observe the interactions between the protein products of these genes and various anti-cancer drugs previously approved by the FDA.

The 34 genes of the signature studied were converted into their different 3D structures, annotated in the PDB repository [30] (only those structures with a validated 3D annotation were chosen). Of these 34 genes, only 16 were 3D annotated: PADI4, VIP, GRIP2, NCKIPSD, PGLYRP1, FABP6, CDH3, VEGFA, NOG, EPB41L3, IL6R, CITED2, NR3C2, RHEB, PTPRN and GLTP. These genes represent 60% of those indicated in Table 2. These 16 genes resulted in a total of 155 PDB structures. Figure 4 shows a diagram representing the number of PDB structures obtained for each gene.

A list of anti-cancer drugs that had been approved by the FDA was selected. In the Supplementary information file S1, a complete list of the name of the 81 drugs corresponding to the compounds downloaded from the ChEMBL repository, which have been used for the Molecular Docking experiment, are shown. All of them are marketed for use in different cancer treatments.

After the execution of the Molecular Docking experiment of confronting the 155 PDB structures with the 81

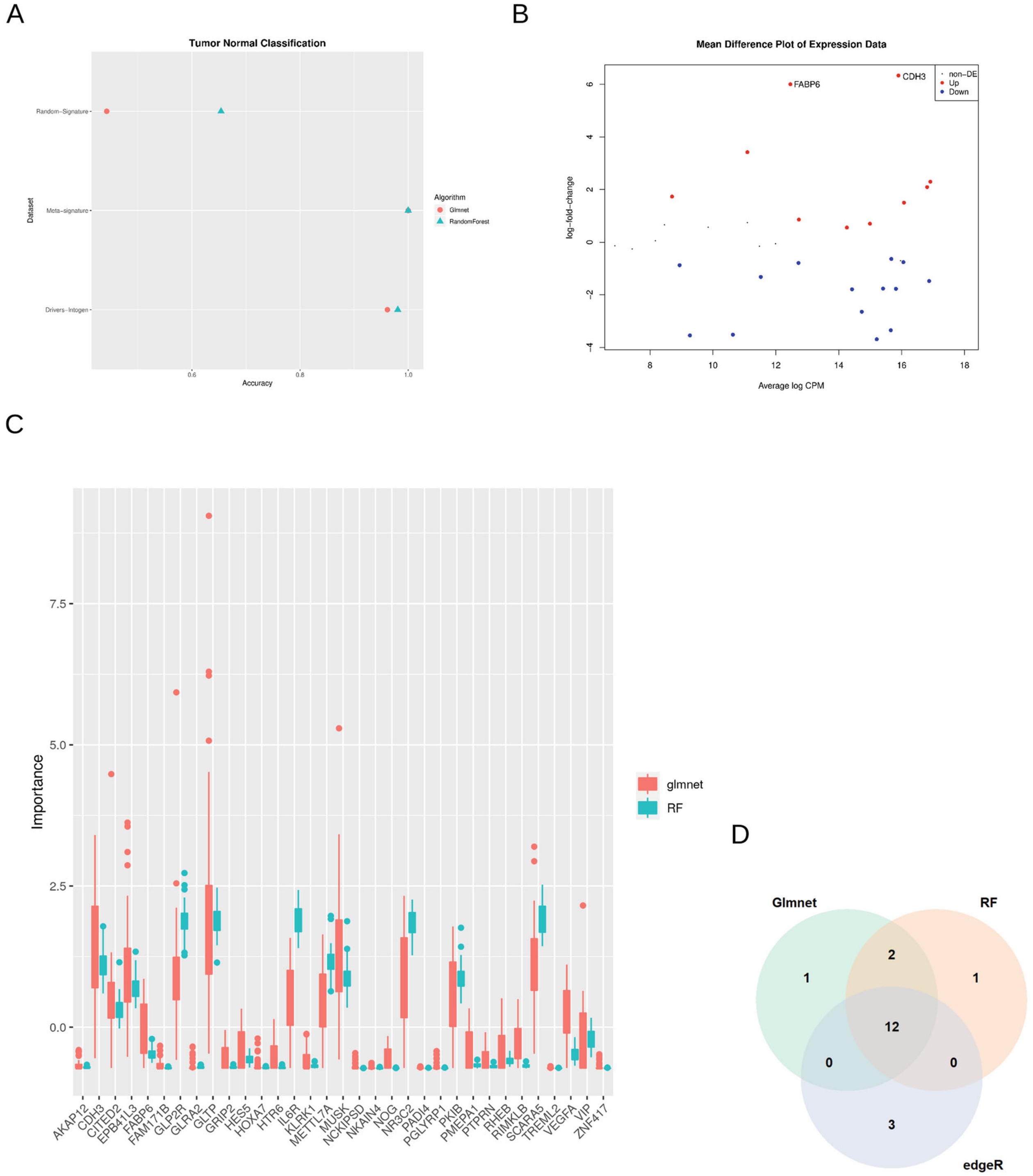


**Fig. 3** Results of analyisis prediction from tumor and helath tissues. **a**) A comparative ML task was carried out with three different signatures (Random signature, Meta signature and Drivers Intogen) to predict between tumor and helth tissues. TCGA expression values of these three signatures were the input in training phase for two ML algorithms (Random Forest and glmnet). The accuracy of the models for each signature is shown. **b**) Mean difference plot after differencial gene expresión is shown. Up and Down expression genes are highlighted in red and blue respectively. FABP6 and CDH3 were the genes with major gene expression differences. **c**) Comparative variable importance for metasignature in Random Forest and glment algorithms. Values were scaled for comparative analysis. **d**) Pie chart with intersections of same genes obtained by two ML approaches and differential gene expression. The three approaches obtained very similar conclusions

**Table 2** Top 15 Variable Importance obtained through Glmnet and Random Forest algorithm. In addition, we have compared these results with a classical analysis aproach for differential expression analysis with edgeR package

| Glmnet | Random Forest | edgeR |
|---|---|---|
| GLTP | GLP2R | CDH3 |
| CDH3 | GLTP | GLP2R |
| MUSK | IL6R | VEGFA |
| SCARA5 | SCARA5 | MUSK |
| NR3C2 | NR3C2 | PKIB |
| GLP2R | CDH3 | SCARA5 |
| EPB41L3 | METTL7A | PMEPA1 |
| PKIB | MUSK | FABP6 |
| IL6R | PKIB | RHEB |
| METTL7A | EPB41L3 | IL6R |
| CITED2 | CITED2 | NR3C2 |
| VEGFA | VIP | VIP |
| FABP6 | FABP6 | EPB41L3 |
| VIP | VEGFA | GRIP2 |
| RIMKLB | HES5 | METTL7A |

drugs, the results were obtained for each PDB structure-drug binomial, indicating the value of the interaction in $\frac{kcal}{mol}$. As a result of the study, the 50 strongest interactions (see Supplementary information file S2) were evaluated and only 4 different genes were identified among them, shown in Table 3. For this experiment, a significant interaction was considered for values that were lower than $-7\frac{kcal}{mol}$.

It is important to point out that among the 50 strongest interactions (see Supplementary information file S2), 92% involved some structure of the GLTP gene. In position 40 and 44 of the ranking, we found PDB structures of the PTPRN gene, in position 48, the PDB structure of the VEGFA gene and in position 49, the PDB structure of the FABP6 gene. Although there is a predominance of PDB structures of the GLTP gene, there is little difference in the force of interaction, showing a decrease of only $-1.6\frac{kcal}{mol}$ between the strongest interaction (3SOI-Nilotinib) and the one in position 49 (2MM3-Abemaciclib).

A review study was made for each of these four genes to see which might be possible therapeutic targets.

### Study of each of the genes

A comparative study of the four genes was carried out to analyse if any of them could behave as a possible therapeutic target. In Fig. 5 the expression between tumour tissue

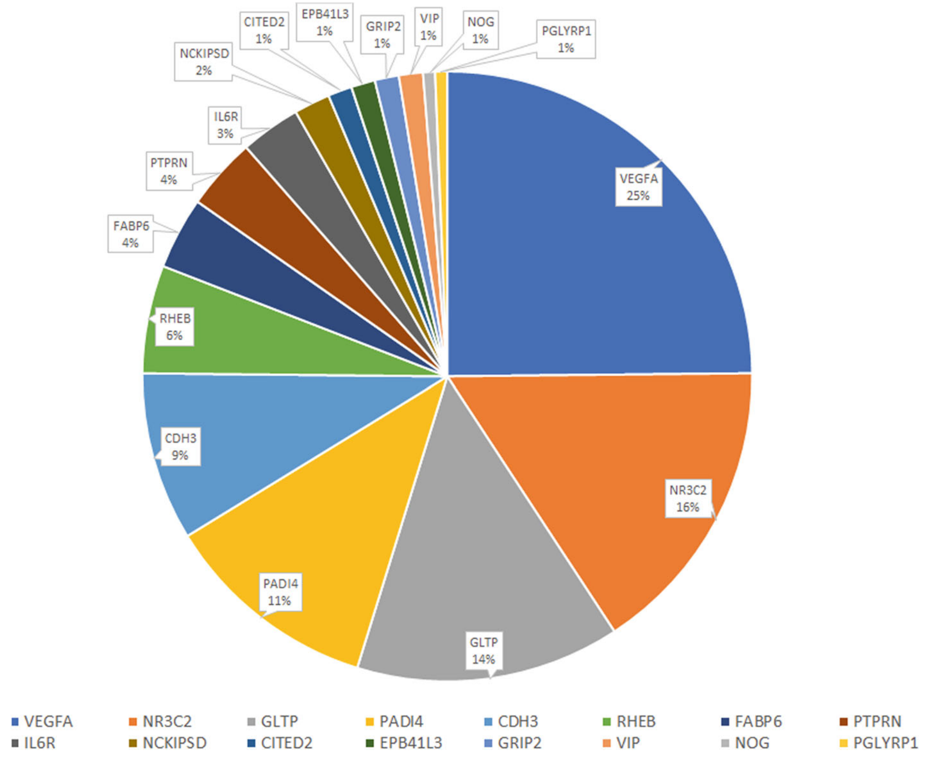


**Fig. 4** Percentage of 3D PDB structures for each gene obtained

**Table 3** Top interactions of the 4 genes that have appeared among the 50 best interactions

| Gene | Protein | Drug | AE (kcal/mol) |
|---|---|---|---|
| GLTP | 3S0I | Nilotinib | -13.7 |
| PTPRN | 3NP5 | Venetoclax | -12.3 |
| VEGFA | 4GLS | Venetoclax | -12.2 |
| FABP6 | 2MM3 | Abemaciclib | -12.1 |

and adjacent healthy tissue of each of the four genes from the COAD cohort of the TCGA is seen.

As can be seen, the genes GLTP and PTPRN present an underexpression in tumour tissue, so attacking it through inhibitor drugs will not produce a positive consequence when slowing tumour development. On the other hand, VEGFA and FABP6 genes are overexpressed in tumour tissue, which makes them possible candidates for inhibitory therapy. This is an important step because in addition to observing whether the gene is over- or underexpressed in tumour tissue, it is crucial to know what its status is in normal tissue. As shown in Fig. 5, VEGFA has significant expression in normal tissue. Whereas the FABP6 gene showed no expression in normal tissue, which is beneficial if our objective was to propose it as a possible biomarker and therapeutic target. Therefore, the biological function of this gene has been deepened.

Docking studies show that the drugs *venetoclax* and *abemaciclib* (previously known as LY2835219) have a significant interaction with the VEGFA and FABP6 genes, respectively. As for the drug *venetoclax*, itwas approved in 2016 as therapy for patients with Chronic Lymphocytic Leukemia (CLL). The mechanism of action of this drug focuses on inhibition of the apoptosis regulator Bcl-2, which is a 'single protein' [31]. Moreover, textitabemaciclib was approved in 2017 for breast cancer patients. Like the previous drug, this is an inhibitor against cyclin-dependent kinase 6, which is also a 'single protein' [32].

In this way we ruled out genes underexpressed in tumour due to the type of drugs we tested. As for the VEGFA and FABP6 genes, the first of them (Vascular Endothelial Growth Factor A - VEGFA) is a specific growth factor for vascular endothelial cells, capable of inducing angiogenesis in vivo [33]. This gene is the central axis in tumour angiogenesis, and there are already different experimental therapies tested against this gene [34, 35]. In addition, different studies are working to predict the different peptides, in silico form, that act against this target [36–39].

The FABP6 gene produces Ileal lipid-binding protein (ILBP) which is a member of a family of fatty acid binding proteins, retinoic acids, and intracellular bile acids [40]. In relation to cancer, the FABP family has been reported to play a role in the development and pathogenesis of cancer [41], and as a possible therapeutic target in clear renal cell carcinoma [42]. Specifically, the FABP6 gene has been suggested as a potential drug discovery target [24, 43], although to date no therapy directed against this gene and/or protein product has been approved.

Our findings are in line with the conclusions shown in the work of Ohmachi et al. [23] published in 2006 by a high impact journal such as Clinical Cancer Research. In this work, they observed that the expression of FABP6 was higher in primary colorectal cancers and adenomas than in normal epithelium, thus suggesting that FABP6 plays an important role in early carcinogenesis. The results of our research are linked to this conclusion, firstly by observing how our signature, in which FABP6 was present, was able to predict more accurately, even more than previously identified genetic drivers, between healthy and tumoural tissue. In addition, analysis of the importance of variables in ML models and differential expression analysis showed that FABP6 was at the top of both lists (see 3 b and c).

Omachi et al. [23] also focused their research on the FABP6 gene because of the large difference in gene expression between healthy tissue and tumor tissue. In addition, the results of [23] were based on a Chinese population, while ours are from the USA. It can be inferred that the function of this gene could be cross-sectional in different world populations. These differences are explained by the high concentration of secondary bile acids present in patients with colonic adenoma.

Reinforcing our hypothesis that FABP6 may be an interesting biomarker for colon cancer, in the same work Ohmachi et al.[23] found that tumours expressing higher levels of FABP6 were smaller, supporting that theory that FABP6 could be a biomarker for the early stage of carcinogenesis.

In addition to being an early stage marker in COAD, we believe that FABP6 may also behave as a therapeutic target because: 1) it is known that each of the nine types of FABP proteins shows tissue specificity, with FABP6 being the ileum, thus generating specificity in future treatment; and 2) the expression of FABP6 in tumour tissue is due to an increase in secondary bile acids, and it is known that the action of these bile acids triggers cellular apoptosis [44]. Therefore, avoiding the metabolisation of these acids would cause apoptosis in cells of the cancerous tissue, abruptly stopping their growth.

Therefore, the data we found in the literature led us to design a drug repurposing experiment to find an already approved drug that could specifically bind to this possible therapeutic target.

### Deepening in Abemaciclib and FABP6 interaction

The Molecular Docking results presented in this work show a significant specificity of all protein PDB structures of the FABP6 gene with the drug *Abemaciclib*. Our

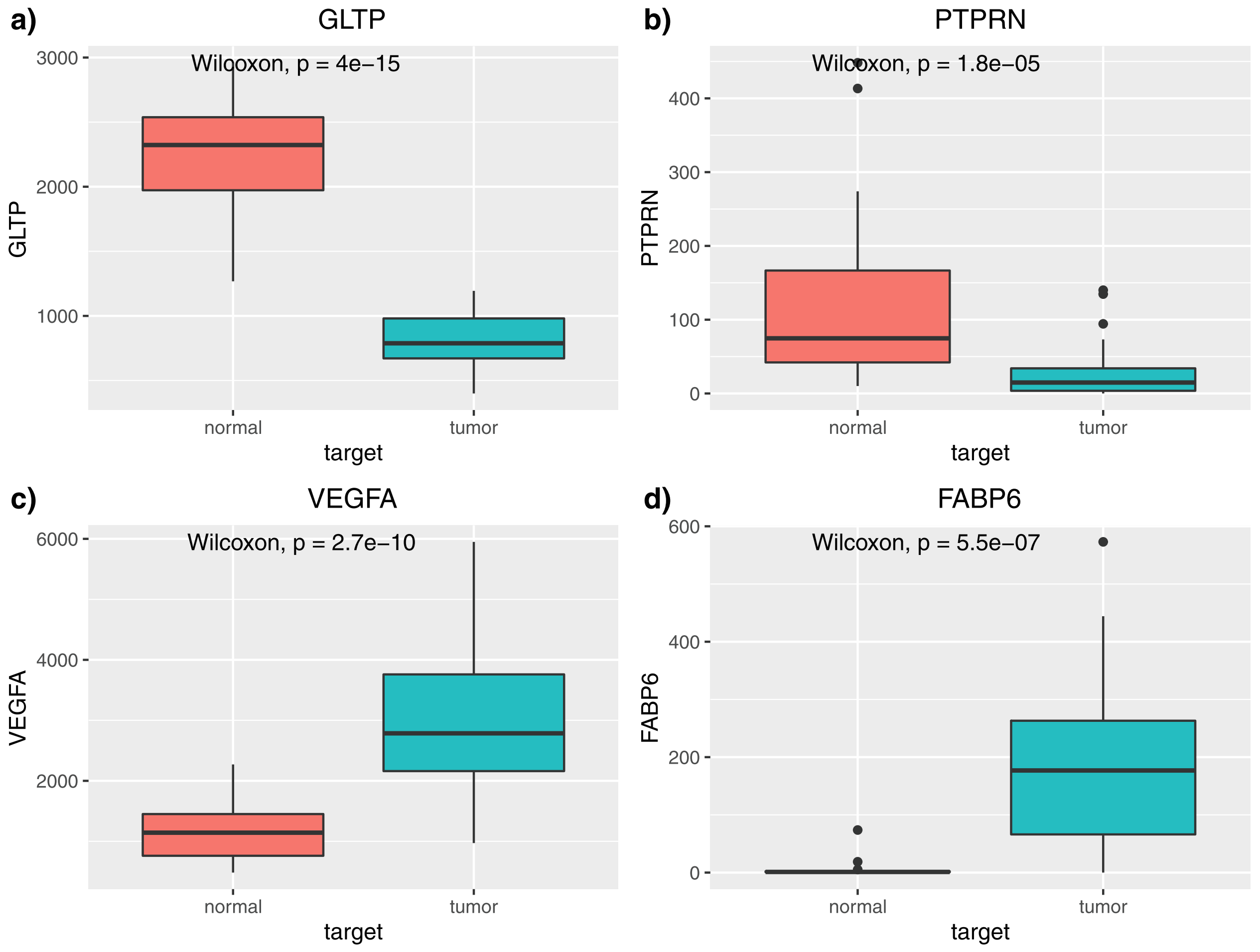


**Fig. 5** Box diagram of the expression of the four genes between healthy and diseased tissues of the COAD cohort of the TCGA

experiment took into account a total of six PDB structures (2MM3, 1O1U, 1O1V, 5L8I, 5L8N, 5L8O). In Table 4, the interaction force of the *Abemaciclib* with all the PDB structures annotated in the FABP6 gene is shown. As shown in the Table 4, all interaction forces have a value of less than -7 $kcal/mol$, and all are considered significant.

In order to understand the docking details, a new open science Web tool was introduced as COAD-DRD: Colon Adenocarcinoma Drug Repurposing with Docking, available at https://muntisa.github.io/COAD-DRD/. There are six different sections: Abemaciclib-FABP6 - dedicated to the interactions of Abemaciclib with the six PDB structures, Selected - with the most interesting interactions between drugs and genes in COAD, Top50 - with statistical plots about the docking signature of the best 50 drugs in COAD, Full by Genes - with statistical box plot for the interactions of each gene with all the drugs (without any cutoff for the AE), Full by Drugs - with graphics that show the drug signature on all the genes in COAD, and Full DB - a pivot table with graphics that give the possibility to represent all docking results by any criteria.

**Table 4** Interaction force of Abemaciclib with all PDB structures of the FABP6 gene

| Gene | PDB structure | Drug | AE (kcal/mol) |
|---|---|---|---|
| FABP6 | 2MM3 | Abemaciclib | -12.1 |
| FABP6 | 1O1U | Abemaciclib | -8.0 |
| FABP6 | 1O1V | Abemaciclib | -10 |
| FABP6 | 5L8I | Abemaciclib | -9.0 |
| FABP6 | 5L8N | Abemaciclib | -9.5 |
| FABP6 | 5L8O | Abemaciclib | -10 |

COAD-DRD sections provide interactive graphics, interactive 3D structures for the complexes that provide direct visualisation with the binding poses, interactive tables with specific datasets for each section (with filtering, searching), interactive pivot tables with a high degree of flexibility to visualize the entire dataset for this study, and direct visualisation of important docking information (docking outputs, search box configuration, pdbqt files, pictures of interactions, contact atoms, or hydrogen bonds, etc.). The source code and all the other files, including the script to generate the dynamic elements, are available as an open GitHub repository at https://github.com/muntisa/muntisa.github.io/tree/master/COAD-DRD.

Based on our findings, the FABP6 gene and, specifically, its protein products, are proposed as therapeutic targets for the development of colon cancer. In addition, owing to the drug repurposing experiment, we present the drug *abemaciclib* as a possible drug that may interact specifically against the protein products of this gene.

Blind molecular docking means that the search of the best Abemaciclib interaction uses the entire surface of the FABPs without defining an active site of the natural ligands (lipids). This could generate docking results where Abemaciclib could interact out of the active site without implications in the FABP activity. Therefore, Fig. 6 presents the three FABP structures with the natural ligands and the best interaction of Abemaciclib with the FABPs in order to check the location of these interactions. FABPs are represented using ribbons (white), the natural ligand using lines (violet) and Abemaciclib using sticks and balls (blue-green).

Figure 6 presents FABP with PDB ID 1O1V: human ileal lipid-binding protein (ILBP) in complex with cholyltaurine (ligand). This protein has a single ligand active site defined by 10 amino acids: TYR14, MET18, ILE23, VAL27, TRP49, TYR53, ASN61, MET74, LEU90, ARG121. 1O1V amino acid – Abemaciclib atom interactions are SER54:HG, MET59:CE, ILE23:CD1, TYR53:CE2, VAL27:CG1, ILE23:CG2, LYS77:CD, GLN51:HE21, VAL27:CG2, MET74:CG, SER24:H,

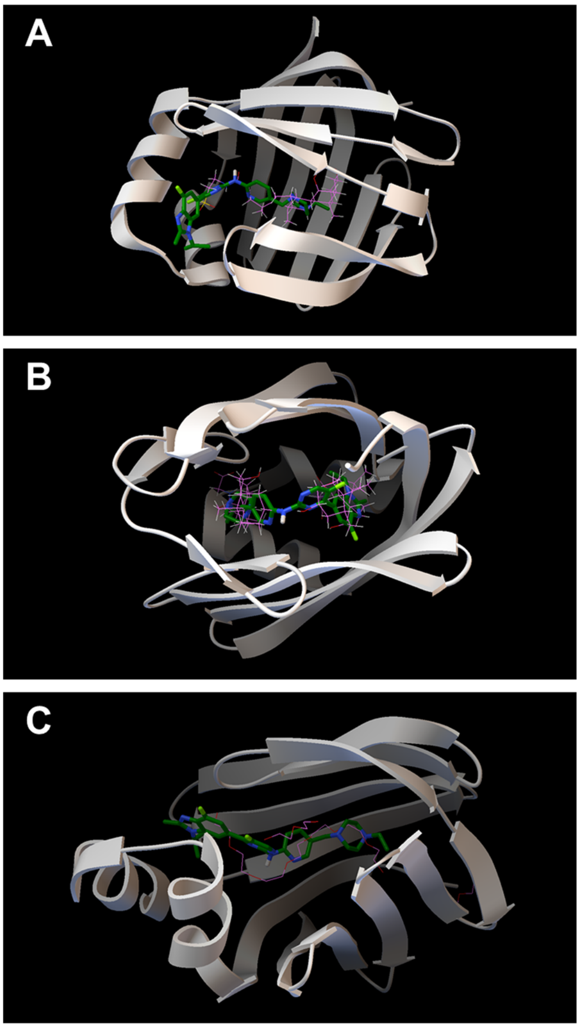


**Fig. 6** Three FABP structures (white ribbon) with the natural ligands (violet lines) and Abemaciclib (blue-green sticks and balls): 1O1V **a**, 2MM3 **b**, and 5L8N **c**

TYR53:CD2, TYR14:OH, VAL27:CB. Thus, Abemaciclib interacts with TYR14, ILE23, SER24, VAL27, GLN51, TYR53, SER54, MET59, MET74, LYS77. From these amino acids, five of them are defining the active site of 1O1V: TYR14, ILE23, VAL27, TYR53, and MET74. In addition, the visualization of Fig. 6a demonstrated that Abemaciclib occupy the FABP active site, with consequences in the lipid transport activity.

Figure 6b presents FABP with PDB ID 2MM3: human ileal bile acid-binding protein with glycocholate and glycochenodeoxycholate (two ligands). This protein has two active sites for the two ligands. AC1 site for glycocholate ligand (CHO202) is defined by 22 amino acids: PHE2, PHE6, MET8, MET18, ALA31, ILE36, THR38, VAL40, PHE47, TRP49, GLN51, MET74, LEU90, SER101, GLU102, LEU108, VAL109, GLU110, TYR119, ARG121, and SER123. AC2 site for glycochenodeoxycholate ligand (GCH201) is defined by 17 amino acids: LEU21, ILE23, TRP49, ASN61, PHE63, GLN72, THR73, MET74, GLY75, LYS77, PHE79, VAL83, LEU90, VAL92, TYR97, and GLN99. 2MM3 amino acid – Abemaciclib atom interactions are PHE47:CZ, GLN99:HE21, MET8:CB, PHE47:CE1, GLN99:NE2, PHE2:CB, PHE47:CE2, THR38:CB, GLN51:HE21, PHE2:CD2, PHE79:CZ, LEU90:CB, GLN51:NE2, GLN99:CD, TYR97:CE2, ARG121:CD, VAL109:N, SER101:CB, TRP49:CZ3, VAL109:C, MET8:CG, SER101:HG, THR38:CG2, PHE2:C. Thus, Abemaciclib interacts with PHE2, MET8, THR38, PHE47, TRP49, GLN51, PHE79, LEU90, TYR97, GLN99, SER101, VAL109, and ARG121. All these amino acids are defining both active sites for both natural ligands in 2MM3: PHE2, MET8, THR38, PHE47, GLN51, SER101, VAL109, ARG121 from AC1 active site and TRP49, PHE79, LEU90, TYR97, GLN99 from AC2 active site. Figure 6b shows that Abemaciclib occupy both FABP active site in the same time. This interaction should disturb the both lipid transport activity.

Figure 6c presents FABP with PDB ID 5L8N: human FABP6 protein with fragment 1 + 3,6,9,12,15,18-hexaoxaicosane-1,20-diol (P33); 5,6-dimethyl-1 H-benzimidazol-2-amine; di(hydroxyethyl)ether (PEG). This protein has several active sites and only the one for P33 and PEG will be compared with the Abemaciclib interaction preference: AC5 site is defined by 11 amino acids - PHE18, TRP50, ILE72, THR74, GLY76, LEU91, TYR98, GLN100, THR101, SER102, ARG122. 5L8N amino acid – Abemaciclib atom interactions are ALA32:CB, PHE64:CZ, MET19:CE, LEU91:CB, VAL28:CG1, PHE64:CE2, GLN100:NE2, TRP50:CE2, TRP50:CG, GLY76:CA, ILE72:CD1, TRP50:CD2, and MET75:CB. Thus, Abemaciclib interacts with MET19, VAL28, ALA32, TRP50, PHE64, ILE72, MET75, GLY76, LEU91, and GLN100. Five of these amino acids are defining AC5 active sites in 5L8N: TRP50, ILE72, GLY76, LEU91, and GLN100. Fig. 6c shows that Abemaciclib occupy the FABP active site where normally interacts both ligands: P33 and PEG. This interaction should modify the ability of FABP to transport lipids.

In conclusion, Abemaciclib prefers interactions inside the active site of FABPs using more nonpolar/aliphatic/hydrophobic amino acids (GLY, ALA, VAL, LEU, ILE, MET, TRP, PHE) than hydrophilic uncharged amino acids (SER, THR, TYR, GLN) or charged/basic amino acids (LYS, ARG). This is explained by the FABP preference for aliphatic interactions to link natural lipids for transport.

The results will then be discussed, focusing mainly on the interaction of FABP6 and *abemaciclib* protein products. Proposing in this way, a new potential candidate to be validated experimentally.

## Discussion

The results obtained in this work report the FABP6 gene as a possible therapeutic target and the drug *abemaciclib* as a possible drug directed towards the gene products of this gene.

As mentioned above, this gene functions as a receptor for lipids and bile acids. It is curious to observe how its genetic expression is almost totally specific to the small intestine - terminal ileum, with average values around 500 RPKM. In other tissues, as can be seen in the data from the Genome Browser [45], its expression is practically null. Therefore, the data shown in Fig. 5 indicate an abnormal function of that gene in carcinogenic tissue.

It is interesting to note the expression behaviour of this gene in different tumour. With the available TCGA data, the expression of the FABP6 gene in healthy and tumour tissue has been compared (see Fig. 7).

The results show a significant difference in certain types of tumour. Firstly, we observed how the expression is practically null in all the healthy tissues of each one of the different patients, being indifferent to the type of cancer. However, the PCPG cohort shows some contradictory results to what was previously proposed, which could harbour new functions and roles of the FABP6 gene. On the other hand, although there is a significant difference in certain tumours (as may be the case of the breast adenocarcinoma (BRCA), stomach adenocarcinoma (STAD) or cholangiocarcinoma (CHOL) cohort, for example), this difference is mainly due to the outliers, as shown in the different box diagrams. This does not occur in the READ cohort (rectum adenocarcinoma), which presents high levels of FABP6 in tumour tissues throughout the sample. This fact coincides with the results shown in this work, supporting the idea of specificity of FABP6 expression in colorectal tissue.

In this context, we can conclude that FABP6 is a specific biomarker for COAD and READ, so the action of

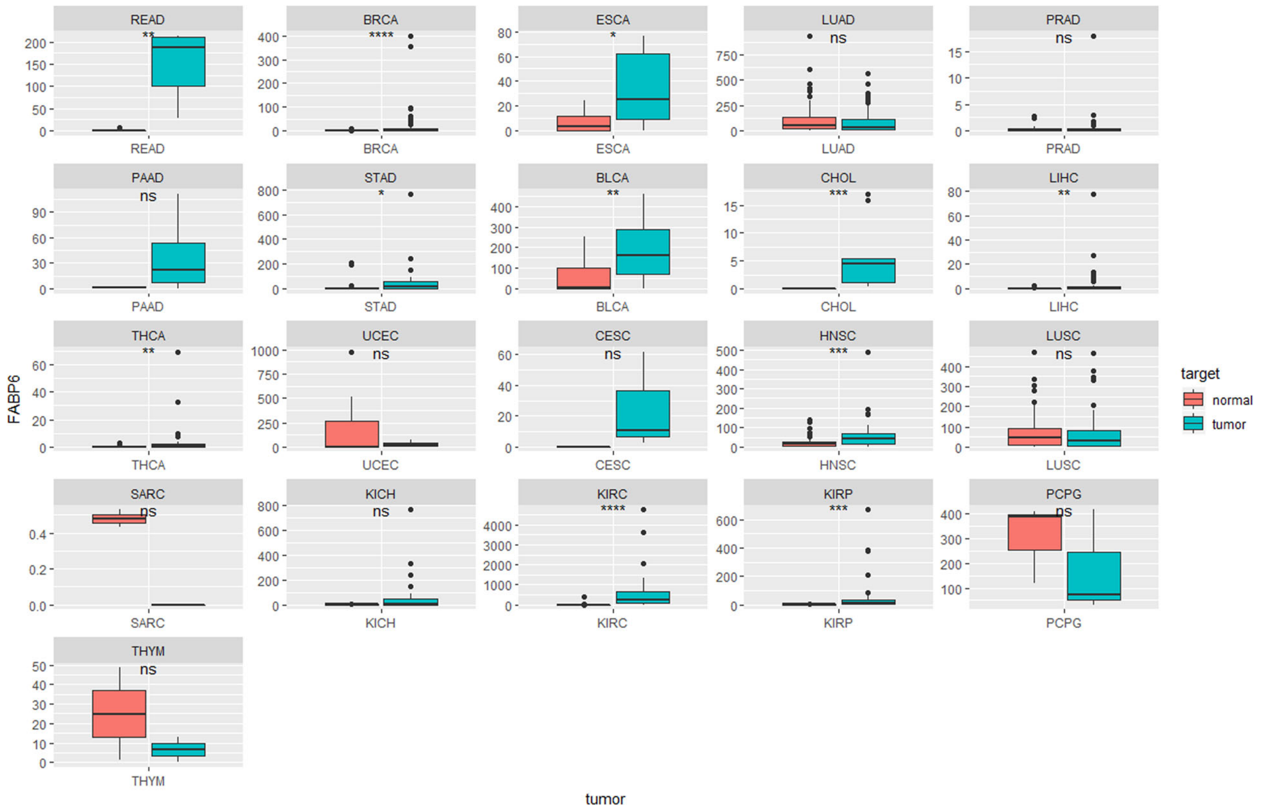


**Fig. 7** Box plot panel with the comparision between tumour and control samples through 21 tumour s types from TCGA

an inhibitory mechanism could lead to positive results in slowing down the growth of the tumour . Furthermore, as mentioned above, FABP6 is an early diagnostic biomarker, which would greatly assist the various possible treatments of this type of cancer.

Regarding its function, this gene intervenes mainly in the signalling peroxisome proliferator-activated receptor (PPAR) pathway. The FABP family activates the PPAR signalling pathway, which acts as transcription factors, regulating the expression of different genes related to lipid metabolism, adipocytic differentiation, adaptive thermogenesis, cell survival, gluconeogenesis and ubiquitination [46]. These functions may be related to the development and differentiation of cancer cells. In addition, previous studies have already linked this pathway to cancer, and specifically to colon cancer [47–50].

Comparisons with other studies and databases, show a significant decrease in the survival of patients with a high copy number of the FABP6 gene, as seen in Fig. 8, obtained from the web and article [51, 52]. From this survival curve and the function of the gene, it can be inferred that when there is a very abrupt change in the coding of the FABP6 gene, there can be serious problems in the survival of the patient. Due to its function of regulating fatty acids and bile acids, and after development of COAD, the patient will present greater deregulation in gastrointestinal homeostasis, which would justify its worst prognosis. At this point, it is interesting to point out that the aberration or inhibition of this gene in tumour cells alone could theoretically provide an advantage when considering this gene as a possible therapeutic target. Due to the need for tumour cells to provide continuous energy, the metabolic pathways related to fatty acids must be expressed in a significant way. Deregulation of these cellular pathways could provide detection of the growth and development of the tumour . This annotation could justify the results, previously commented on in the article [42].

On the other hand, the drug selected as the ligand for this gene, *Abemaciclib*, has reasonable characteristics to be used as a drug against this type of cancer. It is a small molecule specific inhibitor of cyclin-dependent kinase 4/6, so its effect lies in the detection of cell division by acting on the regulation of G1 phase of the cell cycle. It was approved for use in breast cancer patients in 2017. Although there are no results from clinical trials for this type of cancer and this drug, large pharmaceutical companies are already testing it in combination with other drugs, such as *Ramucirumab* for patients with advanced cancer, colon cancer, and mantle cell lymphoma [53], which also supports our hypothesis.

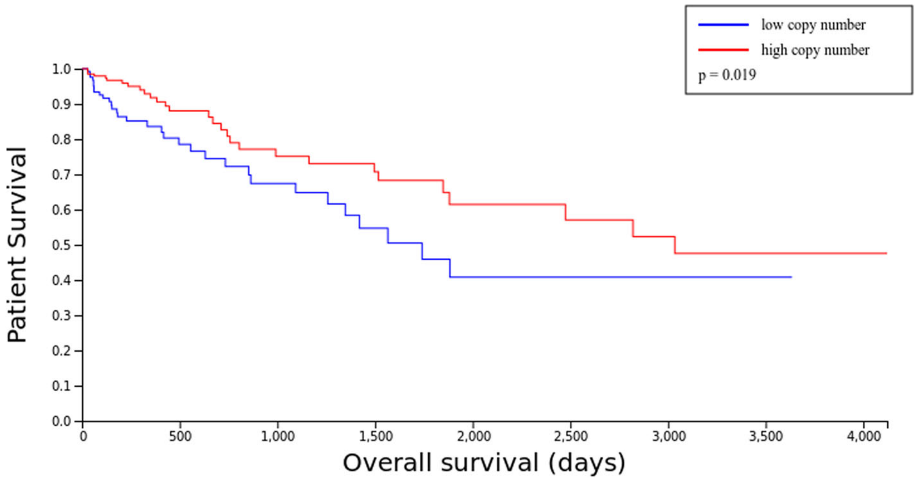


**Fig. 8** Survival curve according to the number of copies of the FABP6 gene. Extracted from [52]

Finally, and after the evidence gathered, both in our own experiments and in previous work, the FABP6 gene and the drug *Abemaciclib* are proposed as a possible targets and treatment, respectively, in colon cancer. The effect of the drug on other types of cancer, as well as the results obtained in this work, support the hypothesis put forward by the present researchers that this drug will join CDK4/6 and FABP6 protein products (specifically in carcinogenic tissue due to its low expression in different tissues), inhibiting both functions and therefore significantly reducing the development of cancer. Although this hypothesis must be validated experimentally, there is sufficient theoretical evidence to think of the gene and the drug as potential anti-cancer therapies.

## Conclusions

The results in silico of this work show how the drug *Abemaciclib*, previously approved for treatment in breast cancer could be used, a priori, in the treatment of colon cancer. In breast cancer, Abemeciclib inhibits CDK4/6, interrupting the cell cycle and the development of the tumour. In this work, we report that this drug could also be used for the treatment of colon cancer, after subsequent experimental validations, due to its strong interaction with all the protein PDB structures of the FABP6 gene. A thorough comparative study was carried out, to observe the evidences that existed after the inhibition of this protein product. All of the evidence indicates that inhibition of the expression of the FABP6 gene, specifically in tumour cells, would reduce the development and growth of the tumour .

This work demonstrates that in silico techniques, such as Machine Learning and Molecular Docking techniques, create added value to the data reported by other initiatives. Owing to the reuse of free access data, it is possible to use computational methods to validate, test, and prove a hypotheses, and thereby considerably reduce research costs.

Finally, in order to obtain new alternatives in the treatment of cancer, the presented hypothesis need to be experimentally validated in the laboratory.

## Methods

### Datasets

RNASeq2 data from the COAD cohort was downloaded from the TCGA repository [3] using the TCGA2STAT package [54]. Patients were filtered according to the type of problem being studied. For classification according to the metastatic status in the lymph nodes, a total of 283 patients were obtained, classifying 166 in stage $N_0$ and 117 between stages $N_1$ and $N_3$.

For the disease stage classification problem, 154 patients were classified between the stages $S_1$ and $S_2$, while 120 were classified between the stages $S_3$ and $S_4$.

Finally, for the problem of classification between healthy and tumour tissues, 26 patients were included in the analysis. These patients presented RNASeq data of their tumour tissue and adjacent normal tissue. For a better understanding of this cohort, some of the clinical data of these patients can be seen Supplementary information file S3.

### Differential expression analysis

A differential expression analysis was performed using the edgeR package. This package assumes that the number of readings in each sample (j) assigned to a gene (i) is modeled through a binomial negative distribution with two parameters, the mean $\mu_i$, j and the overdispersion parameter $\Theta_{ij}$.

$$Y_{ij}\ BN\left(u_{ij}, \Theta_{ij}\right)$$

$Y_{ij}$ corresponds to the non-negative whole number of readings in each sample (j) assigned to a gene (i). The values of the mean and the overdispersion, in practice, are not known so we must estimate them from the data. Finally, using the exact test for the negative binomial distribution, differentially expressed genes are estimated.

### Machine learning

The following algorithms were implemented: random forest (RF) and generalized linear model (glmnet). A nested cross validation was used for training the models. In other words, there were two validation phases. Firstly, a hold-out was used for the selection of the best hyperparameters (2/3 for training and 1/3 for testing) and secondly, a Leave One Out was used for the validation of the model.

### Molecular docking

The strength of the interactions were quantified by the affinity energy (AE, kcal/mol) of ligands for protein targets using the open software AutoDock Vina [55]. The entire processing was done into the BioCAI cluster from the University of A Coruna (Spain). The docking flow had several steps that included the ligand and protein processing, conversion and geometry optimisation before the docking calculations.

Thus, the ligands are presented as a list of commercial drug names. Using PubChem APIs, the compounds for all drugs have been downloaded as SDF 2D. The ligand molecules were converted into PDB by optimising the 3D structure using babel software [56]. The protein targets were only filtered for the first PDB model, the non-protein part was eliminated (water molecules, other ligands, etc.). The PDB of ligands and proteins were converted into PDBQT format using AutoDockTools scripts (prepare_ligand4.py and prepare_receptor4.py) [57]. The protein target was considered rigid in all docking calculations and the interaction searching was considering the entire surface of the targets. The docking flow is based on python and bash scripts, including the reading of the final results. The cut-off for stable interactions is considered AE $< -7.0\frac{kcal}{mol}$ [58]. The results are based on the first docking conformer of the ligands with reference root-mean-square deviation of atomic positions (RMSD) of 0 [59]. We are presenting the top 50 interactions (the most negative AE values). We used 155 protein targets and 151 compounds (24.273 dockings/AE values). The list of interactions and the docking figures are presented for one of the best interaction such as nilotinib – compound 644241 (ligand) with 3s0i (protein target).

In order to understand all the details, a new open Web tool was introduced as COAD-DRD: Colon Adenocarcinoma Drug Repurposing with Docking (https://muntisa.github.io/COAD-DRD/). The tool includes several sections about the best proposed drug for COAD, the top 50 interactions, our selection of interaction and the entire dataset of docking results. All the files and the source of the tool is available as an open GitHub repository at https://github.com/muntisa/muntisa.github.io/tree/master/COAD-DRD. The sections of the web included interactive tables, plots, pivot table and 3D structures widgets (generated with python jupyter notebooks based on HTML, plotly - https://plotly.com, ipywidgets - https://ipywidgets.readthedocs.io/en/latest/, nglview - https://github.com/arose/nglview (DOI:10.5281/zenodo.3700850), pivottablejs - https://pivottable.js.org and datatables - https://datatables.net). Thus, it is possible to zoom into the 3D complex structures between drug binding poses and targets, search for specific results, find details into plots, understand the drug signature on all COAD genes, check the contact atoms and hydrogen bonds of the interactions, and download all docking files.

### Analysis pipeline

In this section we will describe the pipeline followed to obtain the candidates for genes presented in this work. Next, each of the stages carried out in this work will be described step by step.

#### *State of the art review*

The objective of this work consisted of the search and validation of signatures and therapeutic targets for colorectal cancer already reported in the literature.

To this end, a review of published papers that have used TCGA data for the execution of Machine Learning algorithms was carried out. Of all the works found, only those studies in which the dependent variable was related to the prognosis of the disease were chosen. Finally, three papers were identified. Each of these studies reported a signature of genes related to the prognosis of colon cancer patients.

#### *Generation of the meta-signature*

Secondly, a gene signature was built by merging the three previously identified signatures. A total of 34 genes were obtained. The signature was checked for previously defined drivers for colon cancer. For this purpose, the drivers defined in colon cancer were downloaded from the Intogen database, and no coincidence was found between the two lists.

If we assume that the expression of these genes influences the prognosis of individuals, it is interesting, firstly, to know if this signature accurately predicts the prognosis of patients in a cohort such as TCGA and secondly, if any of these genes could be a future protein target, which could be attacked with drugs already approved in the industry.

**Table 5** List of genes, with PDB annotation, used for the Molecular Docking experiment

| PADI4 | VIP | GRIP2 | NCKIPSD |
|---|---|---|---|
| PGLYRP1 | FABP6 | CDH3 | VEGFA |
| NOG | EPB41L3 | IL6R | CITED2 |
| NR3C2 | RHEB | PTPRN | GLTP |

The signature was then validated for two different types of problems. Firstly for the classification of the stage of cancer, and secondly, for the classification of patients between healthy and sick. This experiment was followed by a study of the importance of the variables within the best models.

### *Search for new therapeutic targets*

The next focus in this work was the detection of possible new therapeutic targets using drug repurposing. This experiment presents two well-differentiated parts: the obtaining of targets (proteins) and the obtaining of ligands (drugs).

In order to obtain the targets, the signature of genes (HGNC nomenclature) and all of its possible protein PDB structures were transformed. The transformation was carried out through the biomaRt package. In this step, part of the genes were lost because there is no annotation in PDB for all the protein products of all the genes. In the end, we were left with 16 genes that do have PDB annotation. In the Table 5, the list of genes used for the Molecular Docking experiment is shown. A total of 155 PDB structures derived from these genes were analysed.

To obtain the ligands, anti-cancer drugs that have already been approved for treatment were chosen. The objective of this process was to find a drug, already approved, that has a significant interaction force against a protein target in order to reuse it, in this case, for colon cancer.

The anti-cancer drugs were obtained from the website of the National Cancer Institute [60]. To validate all the names of the drugs, they were downloaded from the repository DRUG REPURPOSING HUB [61]. We made a combination of both lists and only kept those that were already passed the clinical trial and, therefore, are in the market. Finally, after processing, we were left with 81 approved anti-cancer drugs.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12860-020-00295-w.

**Additional file 1:** Additional file S1. List of ligands with their corresponding annotation in ChEMBL. It showed a list of ligand used in the docking experiment. A total of 159 ChEMBL coumpounds are listed from 81 anti-cancer drugs downloaded.

**Additional file 2:** Additional file S2. Top 50 interactions from docking experiment. In this excel file it can be found the interaction force between ligand (drug) and target (protein). The interaction force is measured by kcal/mol.

**Additional file 3:** Additional file S3. Clinical data of patients involved in classification between tumour and health tissue. In this excel file it can be found several clinical variables that correspond to patients involved in tumour and health tissue classification.

### Abbreviations

AE: affinity energy; AECC: Spanish Cancer Association; AIDS: Acquired Immunodeficiency Syndrome; BLCA: Urothelial Bladder Carcinoma; BRCA: Breast Invasive Carcinoma; CCRCC: Clear Renal Cell Carcinoma; CDK4/6: Cyclin-dependent kinase 4/6; CESC: Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma; CHOL: Cholangiocarcinoma; CLL: Chronic Lymphocytic Leukemia; COAD: Colon Adenocarcinoma; ESCA: Esophageal Carcinoma; FDA: Food and Drug Administration; GLMNET: Generalized Linear Model; HIV: Human Immunodeficiency Virus; HNGRI: The National Human Genome Research Institute; HNSC: Head-Neck Squamous Cell Carcinoma; ILBP: Ileal lipid-binding protein; KICH: Kidney Chromophobe; KIRC: Kidney Renal Clear Cell Carcinoma; KIRP: Cervical Kidney renal papillary cell carcinoma; LIHC: Liver Hepatocellular Carcinoma; LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinoma; MD: Molecular Docking; ML: Machine Learning; NCI: National Cancer Institute; PAAD: Pancreatic adenocarcinoma; PCPG: Pheochromocytoma and Paraganglioma; PDB: Protein Data Bank; PPAR: Peroxisome Proliferator-Activated Receptor; PRAD: Prostate Adenocarcinoma; READ: Rectum Adenocarcinoma; RF: Random Forest; RPKM: Reads Per Kilobase Million; SARC: Sarcoma; STAD: Stomach Adenocarcinoma TCGA: The Cancer Genome Atlas; THCA: Thyroid Cancer; THYM:Thymoma; UCEC: Uterine Corpus Endometrial Carcinoma

### Acknowledgements

Not applicable.

### Authors' contributions

"Conceptualization, JLB, CFL, AP; analysis pipeline, JLB, CFL; molecular docking CRM; Machine Learning, JLB; formal analysis, JLB, CFL; writing–original draft preparation, JLB; writing–review and editing, JLB, CRM, AP, CFL; supervision, CFL. All author(s) have read and approved the manuscript".

### Funding

This work was supported by 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." and the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431G/01, ED431D 2017/16), the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23) and Competitive Reference Groups (Ref. ED431C 2018/49). The calculations were performed on resources provided by the Spanish Ministry of Economy and Competitiveness via funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER). The funding body did not have a role in the experimental design; data collection, analysis and interpretation; and writing of this manuscript.

### Availability of data and materials

The datasets analysed during the current study are available in the TCGA repository.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 26 February 2020 Accepted: 24 June 2020
Published online: 08 July 2020

## References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. Int J Cancer. 2015;136(5):359–86.
2. Observatoria de la Asociación Española contra el Cáncer. http://observatorio.aecc.es. Accessed 19 Aug 2019.
3. The Cancer Genome Atlas. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. Accessed 23 July 2019.
4. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell. 2018;173(2):338–54.
5. Way GP, Sanchez-Vega F, La K, et al. Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas. Cell Rep. 2018;23(1):172–80.
6. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. 2018;23(1):181–93.
7. Salvucci M, Würstle ML, Morgan C, et al. A stepwise integrated approach to personalized risk predictions in stage iii colorectal cancer. Clin Cancer Res. 2017;23(5):1200–12.
8. Ekins S, Godbole AA, Kéri G, Orfi L, Pato J, Bhat RS, Verma R, Bradley EK, Nagaraja V. Machine learning and docking models for mycobacterium tuberculosis topoisomerase i. Tuberculosis. 2017;103:52–60.
9. Li J, Fu A, Zhang L. An overview of scoring functions used for protein–ligand interactions in molecular docking. Interdisc Sci Comput Life Sci. 2019;11(2):1–9.
10. Torjesen I. Drug development: the journey of a medicine from lab to shelf. Pharm J. 2015. https://www.pharmaceutical-journal.com/publications/tomorrows-pharmacist/drug-development-the-journey-of-a-medicine-from-lab-to-shelf/20068196.article.
11. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18(1):41.
12. Nieman KM, Kenny HA, Penicka CV, Ladanyi A, Buell-Gutbrod R, Zillhardt MR, Romero IL, Carey MS, Mills GB, Hotamisligil GS, et al. Adipocytes promote ovarian cancer metastasis and provide energy for rapid tumor growth. Nat Med. 2011;17(11):1498.
13. Jing C, Beesley C, Foster CS, Rudland PS, Fujii H, Ono T, Chen H, Smith PH, Ke Y. Identification of the messenger rna for human cutaneous fatty acid-binding protein as a metastasis inducer. Cancer Res. 2000;60(9):2390–8.
14. Guaita-Esteruelas S, Bosquet A, Saavedra P, Guma J, Girona J, Lam EW-F, Amillano K, Borras J, Masana L. Exogenous fabp4 increases breast cancer cell proliferation and activates the expression of fatty acid transport proteins. Mol Carcinog. 2017;56(1):208–17.
15. Shen X, Yue M, Meng F, Zhu J, Zhu X, Jiang Y. Microarray analysis of differentially-expressed genes and linker genes associated with the molecular mechanism of colorectal cancer. Oncol Lett. 2016;12(5):3250–8.
16. Zhao D, Ma Y, Li X, Lu X. microrna-211 promotes invasion and migration of colorectal cancer cells by targeting fabp4 via ppar$\gamma$. J Cell Physiol. 2019;234(9):15429–37.
17. Das R, Hammamieh R, Neill R, Melhem M, Jett M. Expression pattern of fatty acid-binding proteins in human normal and cancer prostate cells and tissues. Clin Cancer Res. 2001;7(6):1706–15.
18. Hashimoto T, Kusakabe T, Sugino T, Fukuda T, Watanabe K, Sato Y, Nashimoto A, Honma K, Kimura H, Fujii H, et al. Expression of heart-type fatty acid-binding protein in human gastric carcinoma and its association with tumor aggressiveness, metastasis and poor prognosis. Pathobiology. 2004;71(5):267–73.
19. Bao Z, Malki MI, Forootan SS, Adamson J, Forootan FS, Chen D, Foster CS, Rudland PS, Ke Y. A novel cutaneous fatty acid–binding protein-related signaling pathway leading to malignant progression in prostate cancer cells. Genes Cancer. 2013;4(7-8):297–314.
20. Korpela J, Adlercreutz H, Turunen M. Fecal free and conjugated bile acids and neutral sterols in vegetarians, omnivores, and patients with colorectal cancer. Scand J Gastroenterol. 1988;23(3):277–83.
21. Hill M, Lennard-Jones J, Melville D, Neale K, Ritchie J. Faecal bile acids, dysplasia, and carcinoma in ulcerative colitis. Lancet. 1987;330(8552):185–6.
22. Kurtz W, Leuschner U. Bile acids in patients suffering from colorectal carcinoma–a pilot study. Tokai J Exp Clin Med. 1983;8(1):59–69.
23. Ohmachi T, Inoue H, Mimori K, Tanaka F, Sasaki A, Kanda T, Fujii H, Yanaga K, Mori M. Fatty acid binding protein 6 is overexpressed in colorectal cancer. Clin Cancer Res. 2006;12(17):5090–5.
24. Zhang Y, Zhao X, Deng L, Li X, Wang G, Li Y, Chen M. High expression of fabp4 and fabp6 in patients with colorectal cancer. World J Surg Oncol. 2019;17(1):171.
25. Sun D, Chen J, Liu L, Zhao G, Dong P, Wu B, Wang J, Dong L. Establishment of a 12-gene expression signature to predict colon cancer prognosis. PeerJ. 2018;6:4942.
26. Xu G, Zhang M, Zhu H, Xu J. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on svm. Gene. 2017;604:33–40.
27. Wen J-X, Li X-Q, Chang Y. Signature gene identification of cancer occurrence and pattern recognition. J Comput Biol. 2018;25(8):907–16.
28. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney SJ, Lopez-Bigas N. Intogen: integration and data mining of multidimensional oncogenomic data. Nat Methods. 2010;7(2):92.
29. Robinson MD, McCarthy DJ, Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
30. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, et al. The protein data bank. Acta Crystallogr D Biol Crystallogr. 2002;58(6):899–907.
31. Compound: VENETOCLAX. https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL3137309/. Accessed 17 July 2019.
32. Compound: ABEMACICLIB. https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL3301610/. Accessed 17 July 2019.
33. Leung DW, Cachianes G, Kuang W-J, Goeddel DV, Ferrara N. Vascular endothelial growth factor is a secreted angiogenic mitogen. Science. 1989;246(4935):1306–9.
34. Kong D-H, Kim M, Jang J, Na H-J, Lee S. A review of anti-angiogenic targets for monoclonal antibody cancer therapy. Int J Mol Sci. 2017;18(8):1786.
35. Vasudev NS, Reynolds AR. Anti-angiogenic therapy for cancer: current progress, unresolved questions and future directions. Angiogenesis. 2014;17(3):471–94.
36. Blanco JL, Porto-Pazos AB, Pazos A, Fernandez-Lozano C. Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection. Sci Rep. 2018;8(1):15688.
37. Ramaprasad ASE, Singh S, Venkatesan S, et al. Antiangiopred: a server for prediction of anti-angiogenic peptides. PloS ONE. 2015;10(9):0136990.
38. Laengsri V, Nantasenamat C, Schaduangrat N, Nuchnoi P, Prachayasittikul V, Shoombuatong W. Targetantiangio: A sequence-based tool for the prediction and analysis of anti-angiogenic peptides. Int J Mol Sci. 2019;20(12):2950.
39. Agrawal P, Kumar S, Singh A, Raghava GP, Singh IK. Neuropipred: a tool to predict, design and scan insect neuropeptides. Sci Rep. 2019;9(1):5129.
40. Birkenmeier E, Rowe L, Crossman M, Gordon J. Ileal lipid-binding protein (illbp) gene maps to mouse chromosome 11. Mamm Genome. 1994;5(12):805–6.
41. Amiri M, Yousefnia S, Forootan FS, Peymani M, Ghaedi K, Esfahani MHN. Diverse roles of fatty acid binding proteins (fabps) in development and pathogenesis of cancers. Gene. 2018;676:171–83.
42. Nagao K, Shinohara N, Smit F, de Weijert M, Jannink S, Owada Y, Mulders P, Oosterwijk E, Matsuyama H. Fatty acid binding protein 7 may be a marker and therapeutic targets in clear cell renal cell carcinoma. BMC Cancer. 2018;18(1):1114.
43. Hendrick AG, Müller I, Willems H, Leonard PM, Irving S, Davenport R, Ito T, Reeves J, Wright S, Allen V, et al. Identification and investigation of novel binding fragments in the fatty acid binding protein 6 (fabp6). J Med Chem. 2016;59(17):8094–102.
44. Venturi M, Hambly RJ, Glinghammar B, Rafter JJ, Rowland IR. Genotoxic activity in human faecal water and the role of bile acids: a study using the alkaline comet assay. Carcinogenesis. 1997;18(12):2353–9.
45. Genome Browser - FABP6. https://genome-euro.ucsc.edu/cgi-bin/hgGene?hgg_gene=ENST00000393980.8&hgg_prot=ENST00000393980.8&hgg_chrom=chr5&hgg_start=160187366&hgg_end=160238735&hgg_type=knownGene&db=hg38&hgsid=232990685_6FxFtqmb7FlsDDM8hiv1bAV3HmFy. Accessed 05 Aug 2019.

46. KEGG PPAR pathway. https://www.genome.jp/kegg-bin/show_pathway?map03320. Accessed 05 Aug 2019.
47. Fanale D, Amodeo V, Caruso S. The interplay between metabolism, ppar signaling pathway, and cancer. PPAR Research. 2017;2017:1–2.
48. Simula MP, Cannizzaro R, Canzonieri V, Pavan A, Maiero S, Toffoli G, De Re V. Ppar signaling pathway and cancer-related proteins are involved in celiac disease-associated tissue damage. Mol Med. 2010;16(5-6): 199–209.
49. Jansson EÅ, Are A, Greicius G, Kuo I-C, Kelly D, Arulampalam V, Pettersson S. The wnt/$\beta$-catenin signaling pathway targets ppar$\gamma$ activity in colon cancer cells. Proc Natl Acad Sci. 2005;102(5):1460–5.
50. Antonosante A, d'Angelo M, Castelli V, Catanesi M, Iannotta D, Giordano A, Ippoliti R, Benedetti E, Cimini A. The involvement of ppars in the peculiar energetic metabolism of tumor cells. Int J Mol Sci. 2018;19(7): 1907.
51. Genetic determinants of cancer patient survival. http://survival.cshl.edu/. Accessed 23 Aug 2019.
52. Smith JC, Sheltzer JM. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. Elife. 2018;7:39217.
53. Lilly Clinical Trial with Abemaciclib. https://clinicaltrials.gov/ct2/show/NCT02745769?id=%22NCT02745769%22&rank=1. Accessed 05 Aug 2019.
54. Wan Y-W, Allen GI, Liu Z. Tcga2stat: simple tcga data access for integrated statistical analysis in r. Bioinformatics. 2015;32(6):952–4.
55. Trott O, Olson AJ. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–61.
56. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open babel: An open chemical toolbox. J Cheminformatics. 2011;3(1):33.
57. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30(16):2785–91.
58. Chang MW, Lindstrom W, Olson AJ, Belew RK. Analysis of hiv wild-type and mutant structures via in silico docking against diverse ligand libraries. J Chem Inf Model. 2007;47(3):1258–62.
59. Coutsias EA, Seok C, Dill KA. Using quaternions to calculate rmsd. J Comput Chem. 2004;25(15):1849–57.
60. Cancer Drugs. https://www.cancer.gov/about-cancer/treatment/drugs. Accessed 16 July 2019.
61. The drug repurposing hub. https://clue.io/repurposing. Accessed 16 July 2019.

## Publisher's Note

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

## 3.3 Predicción de fármacos para el tratamiento del cáncer mediante técnicas de Machine Learning

La metodología de ML que se utiliza en la sección 3.2, también se aplica para desarrollar el siguiente artículo [3]. El objetivo de este trabajo es el desarrollo de modelos capaces de predecir actividades complejas de péptidos a partir de su secuencia primaria. Esto ofrece la posibilidad del descubrimiento de nuevos fármacos. Específicamente, el trabajo se centra en la construcción de un modelo de *screening* automático de fármacos anti-tumorales, basados en su actividad anti-angiogénica.

A pesar de que diversas moléculas anti-angiogénicas han mostrado eficacia en el laboratorio, la aprobación de una molécula concreta para su uso en la clínica requiere mucho tiempo y esfuerzo económico. En este trabajo se describe un modelo que permitiría acelerar dicho proceso a partir de modelos *in silico* de Machine Learning.

Tomando las secuencias aminoacídicas de los péptidos como datos de entrada, se transforman a matrices numéricas mediante descriptores moleculares, que describen la composición de los aminoácidos en la secuencia.

Finalmente, modelos lineales, específicamente un modelo *glmnet*, entrenado con la combinación de diferentes características pertenecientes a diversos descriptores, obtiene resultados mejorando el estado del arte. El modelo presenta un rendimiento del 0.96 en AUC y 0.86 en Acc. Las técnicas de selección de características fueron críticas en la creación del modelo. Debido al gran número de descriptores, se seleccionaron aquellas variables con una correlación significativa respecto a la variable respuesta, en este caso, la presencia o no de actividad anti-angiogénica.

Finalmente se realiza un análisis de las variables utilizadas por el modelo. De acuerdo con los resultados, secuencias diaminoacídicas formadas por combinaciones de Serina-Prolina, Prolina-Fenilalanina ó Valina-Ác.Aspártico; secuencias triaminoacídicas tales como Leucina-Serina-Leucina ó Prolina-Ác.Aspártico-Prolina; y presencia de Cisteína, Arginina y Valina son importantes para la predicción de la actividad anti-angiogénica.

Fui el autor principal del artículo. Realicé las pruebas, construí las figuras y las tablas y escribí el manuscrito. La información de participación de los otros autores está descrita en el artículo.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 29 March 2018
Accepted: 6 October 2018
Published online: 24 October 2018

# Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection

Jose Liñares Blanco[1], Ana B. Porto-Pazos[1,2], Alejandro Pazos[1,2] & Carlos Fernandez-Lozano[1,2]

**Screening and *in silico* modeling are critical activities for the reduction of experimental costs. They also speed up research notably and strengthen the theoretical framework, thus allowing researchers to numerically quantify the importance of a particular subset of information. For example, in fields such as cancer and other highly prevalent diseases, having a reliable prediction method is crucial. The objective of this paper is to classify peptide sequences according to their anti-angiogenic activity to understand the underlying principles via machine learning. First, the peptide sequences were converted into three types of numerical molecular descriptors based on the amino acid composition. We performed different experiments with the descriptors and merged them to obtain baseline results for the performance of the models, particularly of each molecular descriptor subset. A feature selection process was applied to reduce the dimensionality of the problem and remove noisy features – which are highly present in biological problems. After a robust machine learning experimental design under equal conditions (nested resampling, cross-validation, hyperparameter tuning and different runs), we statistically and significantly outperformed the best previously published anti-angiogenic model with a generalized linear model via coordinate descent (glmnet), achieving a mean AUC value greater than 0.96 and with an accuracy of 0.86 with 200 molecular descriptors, mixed from the three groups. A final analysis with the top-40 discriminative anti-angiogenic activity peptides is presented along with a discussion of the feature selection process and the individual importance of each molecular descriptors According to our findings, anti-angiogenic activity peptides are strongly associated with amino acid sequences SP, LSL, PF, DIT, PC, GH, RQ, QD, TC, SC, AS, CLD, ST, MF, GRE, IQ, CQ and HG.**

The angiogenesis process consists of the growth and development of new blood vessels from existing ones. The continuous interaction between endothelial cells and the cellular environment that surrounds them is fundamental for this process to occur. Under normal conditions, there is constant regulation between inhibitory and promoter molecules in this process, generating a correct vascularization of tissues[1].

The study of this field has grown enormously in recent years due to the discovery of effective anti-angiogenic therapies in numerous fields, including dermatology[2], ophthalmology[3], vascular diseases[4] and oncology[5].

Cancer research is the area in which most studies are being conducted due to the increased incidence of this disease across the population. Recent data from the National Institute of Health (NIH) indicate that between 2012 and 2030, the incidence of cancer is expected to rise by 50%, from 14 to 21 million patients a year. As to the number of deaths, an increase of 60% is expected, from 8 to 13 million deaths a year.

Previous studies have shown that cancer cells induce the growth of the blood vessels around them, providing them with nutrients and molecules vital for their development[6]. In addition, many cancer types have been reported as being dependent on the process of angiogenesis and respond well to anti-angiogenic therapies[7]. The promising results obtained by researchers and the FDA approval of drugs that inhibit this process as a treatment for various types of cancer have led to the development of a therapy focused on the inhibition of cellular angiogenesis from multiple small peptides, each with a specific target on different metabolic pathways[1].

[1]Department of Computer Science, Faculty of Computer Science, University of A Coruña, A Coruña, 15071, Spain. [2]Instituto de Investigación Biomédica de A Coruña (INIBIC). Complexo Hospitalario Universitario de A Coruña, A Coruña, Spain. Correspondence and requests for materials should be addressed to C.F.-L. (email: carlos.fernandez@udc.es)

A peptide is constituted by the union of amino acids by peptide bonds. The main difference from proteins is its size and structure. Peptides are smaller (between 2 and 50 amino acids) and do not have complex, tertiary or quaternary structures. These characteristics have a series of advantages when peptides are used as therapeutic agents. On the one hand, they are small, organic molecules with a very low level of toxicity. They present, on the other hand, great specificity when joining with other molecules, which facilitates a targeted therapy for a variety of tissues. In addition, they can be designed *in vitro*[1]. Given the simple characteristics of these molecules, it is easy to see that an amino acid sequence will be crucial for the presence of anti-angiogenic function. Previous studies reported that the presence of residues such as Cys, Pro or Ser is related to this activity, while residues such as Ala, Asp or Ile have the opposite functionality[8].

A metabolic pathway is a molecular process involving various gene products with the aim of performing a specific function. The interaction among the various proteins can be direct (phosphorylation, acetylation, etc.) or indirect (via second messengers such as cAMP). The coordination of all molecules is crucial for the correct functioning of the route. This is why the inhibition of a molecule can be considered as a therapeutic target to stop a certain activity or molecular pathway.

Because of the high cost and low speed of the experimental techniques used to evaluate the presence of any peptide activity, researchers increasingly rely on *in silico* experiments for the a priori prediction of the possible activities of each peptide. Previously, these methods *in silico* were based solely on searching in the scientific literature and public databases–for example, the search for protein domains[9], the genomic and proteomic study later contrasted by the null hypothesis theory[10] or the sequence homology[11]. In this way, a theoretical framework was presented prior to evaluation by experimental techniques, owing largely to massive projects such as the Human Genome Project, thus achieving great savings in economic, time and human resources.

After the implementation of more powerful statistical and computational techniques in the biomedical field and with the drastic reduction of costs in the acquisition of hardware, the theoretical framework was strengthened, and machine learning (ML) algorithms, among others, began to be used. Thus, with the use of classification algorithms, models with high performances were obtained[8,12–14], achieving great success in the classification of peptide activities[15], cell-penetrating peptides[16] or anti-cancer peptides[17].

The classification model reported in the present paper represents a quantitative structure activity relationship (QSAR) between the protein amino-acid composition and the biological function. Previous studies on other protein functions focused on anti-oxidant[18], transporter[19], cell-penetrating[20], anti-viral[21], enzyme regulator[13], cell death-related[22], cancer-related[23,24], microbiome-related[25] or signaling[12] proteins.

Regarding anti-angiogenic activity, most studies were based only on the experimental part[1,26–28], which greatly increased the cost and time spent in the characterization process. Although there are also studies that have implemented algorithms based on ML[8], the highest prediction performance value is closest to 0.81 in accuracy, and this study did not report any combined measure to control overfitting or type I and II errors.

Due to the need to strengthen the theoretical framework in the prediction of peptides with anti-angiogenic activity, our aim is to obtain the molecular descriptors, select the best variables within the descriptors regardless of the descriptor and look for the machine learning algorithms that better convey the underlying knowledge in the data. With the combination of these different stages, as shown in Fig. 1 new information on anti-angiogenic activity will be obtained, and an effective predictive model will be presented to ensure that one specific peptide will be a potential candidate for evaluation through experimental techniques.

Once the state of the art and the present study have been introduced, we move on to discuss the structure of this paper. First, the results are divided into several subsections: baseline algorithms without feature selection, feature selection and best model determination. This is followed by a discussion and the conclusions of this study. Finally, the materials and methods section contains a brief introduction to the dataset, molecular descriptors, machine learning, feature selection and experimental design used in this work.

## Results

Previous studies have shown that anti-angiogenic peptides have common functionality, structure and composition[1,10]. Regarding the structure, the vast majority of folds are anti-parallel beta sheets and contain a relatively high incidence of hydrophobic and cationic residues[10]. Furthermore, it has been shown that such peptides are more prone to have certain residues and amino acid sequences in their composition, although this feature is not completely defined. Alignment analysis has not indicated significant sequential commonalities among the peptides[10]. Therefore, the present study aims to elucidate fundamental aspects in the study of the aminoacidic composition of these peptides.

**Baseline algorithms without feature selection.** We used four different machine learning algorithms: RF, SVM, k-NN and glmnet. Initially, we considered the three original datasets (AAC, TC, DC) and merged them. As shown in Fig. 2a, the most informative dataset is AAC, and the merging of the datasets (AAC_TC and AAC_DC) significantly improves the performance of the models (DC and TC) while slightly reducing the deviation of the results, as shown in Fig. 2b.

Results do not improve those published before in the literature (0.809 in accuracy) using a SVM and NT15 terminus dataset (contain first fifteen residues from the N-terminal region of the peptide sequence), as shown in Fig. 2c, but to reduce the noise in the datasets, an FS approach should be applied. Figures in this paper were built using the ggplot2 package[29].

In conclusion, in terms of both AUC and accuracy, the best result was obtained by the RF algorithm trained with the AAC dataset, as shown in Fig. 2a,c. The TC and DC datasets, generally, achieved lower performance with all algorithms than AAC and the combination of them. In light of these results, we considered two complementary AAC descriptors: parallel correlation pseudo-amino-acid composition (PC-PseAAC) and series correlation pseudo-amino-acid composition (SC-PseAAC)[30]. As shown in Fig. 3 (violin plot), the two better-performing

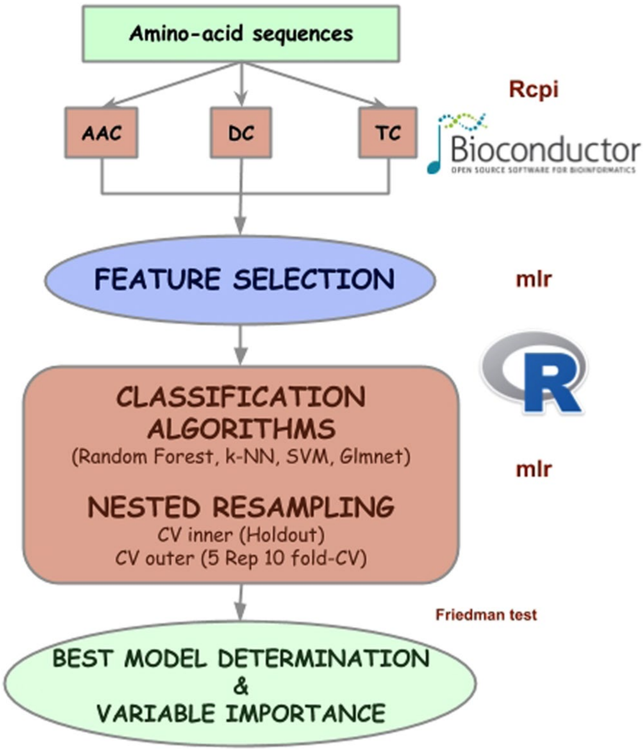


**Figure 1.** Flowchart of this study. The authors thank Bioconductor and R (https://www.r-project.org/logo/) for the provision of the logos under CC-BY and CC-BY-SA open access licenses.

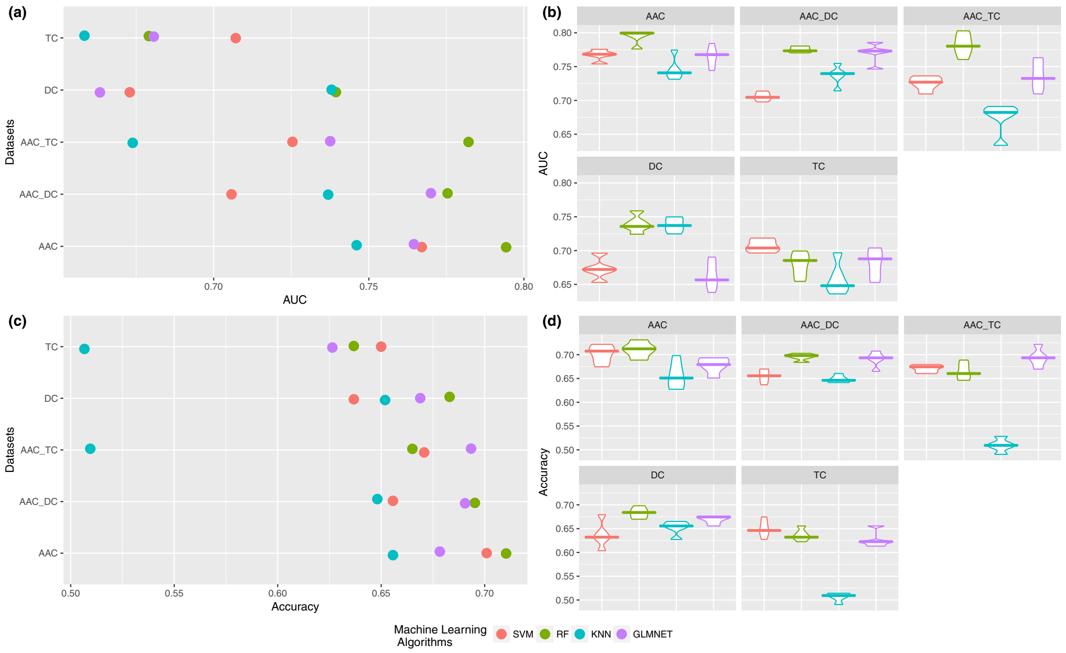


**Figure 2.** Results obtained with the original datasets AAC, TC and DC and their combination. (**a**) Summary of the performance of the four algorithms (AUC), (**b**) boxplot of the behavior of each model across experiments (AUC), (**c**) summary of the performance of the four algorithms (accuracy), and (**d**) boxplot of the behavior of each model across experiments (accuracy).

models in terms of accuracy with the AAC dataset, RF and SVM (Fig. 2c) had opposite behaviors. On the one hand, RF (best model) achieved a comparable result with a similar median value for PC-PseAAC but with outliers for SC-PseAAC in the lower part of the plot; this seems to indicate that it is less stable than AAC[31,32]. However, we

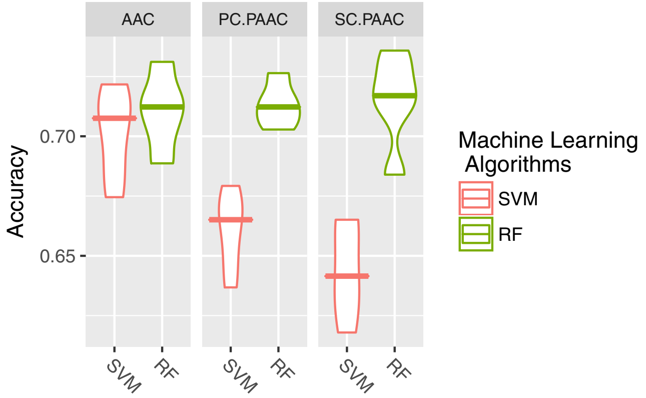


**Figure 3.** Results obtained with the RF and SVM algorithms using AAC and the novel parallel correlation pseudo-amino-acid composition and series correlation pseudo-amino-acid composition.

found that SVM achieved significantly poorer results with a clear decreasing trend in the performance for datasets PC-PseAAC and SC-PseAAC; this is consistent with other works published in the literature[33]. Furthermore, as mentioned before, the aim of this work is to find the aminoacidic composition capable of biologically explaining the differences between anti-angiogenic and non-anti-angiogenic peptides. Due to this instability and to the fact that biologically, AAC is easy to understand, explain and validate by biological lab methods, we decided to use the original AAC dataset for the feature selection experiments.

More precisely, on a closer inspection of the boxplots in Fig. 2b,d, all models show a high variance in their results across experiments, especially the models trained with the descriptors that present a greater number of variables (DC and TC).

**Feature selection.** At this point, we performed an FS approach to reduce the noise in the three original datasets (AAC, DC, TC) and merged them. We ranked the features for each dataset and explored the sizes of different subsets, (5, 10 and 15) for AAC, (25, 50, 75 and 100) for DC, (75, 100, 125 and 150) for TC and (50, 100, 150 and 200) for the union of the three.

The results for each model obtained after feature selection are shown in Fig. 4. Therefore, at this point, the feature selection process shows the most relevant features–in this case, the amino acid residues and sequences of two and three amino acids–that better discriminate between the group of anti-angiogenic peptides and non-anti-angiogenic peptides. These results may have a great impact on later wet studies. This screening process hopefully minimizes the search for important sequences in anti-angiogenic peptides.

The best model in terms of both AUC (see Fig. 4a) and accuracy (see Fig. 4c) has been the glmnet algorithm. The algorithm was trained with the union of the three datasets (AAC, DC and TC) and only the 200 features with the highest rank. In this context, it seems that the glmnet and RF algorithms work better than the others after the feature selection process. In addition, the results seem to indicate a dramatic improvement in the behavior of the algorithms, as seen in Fig. 4b,c. All models show a low variance in their results across experiments. Furthermore, in Fig. 5, the percentage of features of each of the datasets in the combination is shown. An increase in the importance of the features from AAC and DC is observed along with a decrease in TC. Remarkably, the percentage of important features (best model) versus useless features in the datasets is shown in Fig. 6.

Moreover, after the feature selection process, a significant improvement in the performance of the models was obtained, with an average of approximately 15% in accuracy and AUC for all models. The feature selection process works to reduce the noisy features.

The red line in Fig. 4c indicates the best result from the literature for anti-angiogenic peptides by Ramaprasad *et al.*[8] (*accuracy* = 0.809). We statistically outperformed the literature with our experiments and experimental design with more than ten combinations of algorithms and descriptors.

**Best model determination.** The final step in our experimental design[34] is the statistical significance comparison of the performance (AUC) of the machine learning models. As shown in Fig. 4c, seven models outperform the state-of-the-art approach. We used these models to evaluate the statistical significance. Parametric tests have more power than non-parametric tests but, unfortunately, can be used only under certain circumstances. Thus, we checked the normality with a Shapiro-Wilk test, with a level of confidence $\alpha = 0.05$ and the null hypothesis that the data follow a normal distribution; this was rejected with values W = 0.9302 and *p-value* = 0.02836. We performed a Bartlett test with the null hypothesis that our results are heteroscedastic, and we could not reject the null hypothesis with a value for Barlett's K squared of 3.2445 with 6 degrees of freedom and *p-value* = 0.7776. In this case, one of the conditions does not hold, so following the tests, we performed a non-parametric Friedman test with the Iman-Davenport extension assuming the null hypothesis that all models have the same performance; this was rejected with *p-value* = $1.8665 \times 10^{-9}$. A Finner post hoc procedure must be used to correct and adjust p-values for multiple comparisons. Hence, after this test and multiple comparison corrections, the null hypothesis was rejected for all models except the best models using datasets TC_125 and AAC_TC_DC_150, which performed statistically equally to the winning glmnet model with dataset AAC_TC_DC_200.

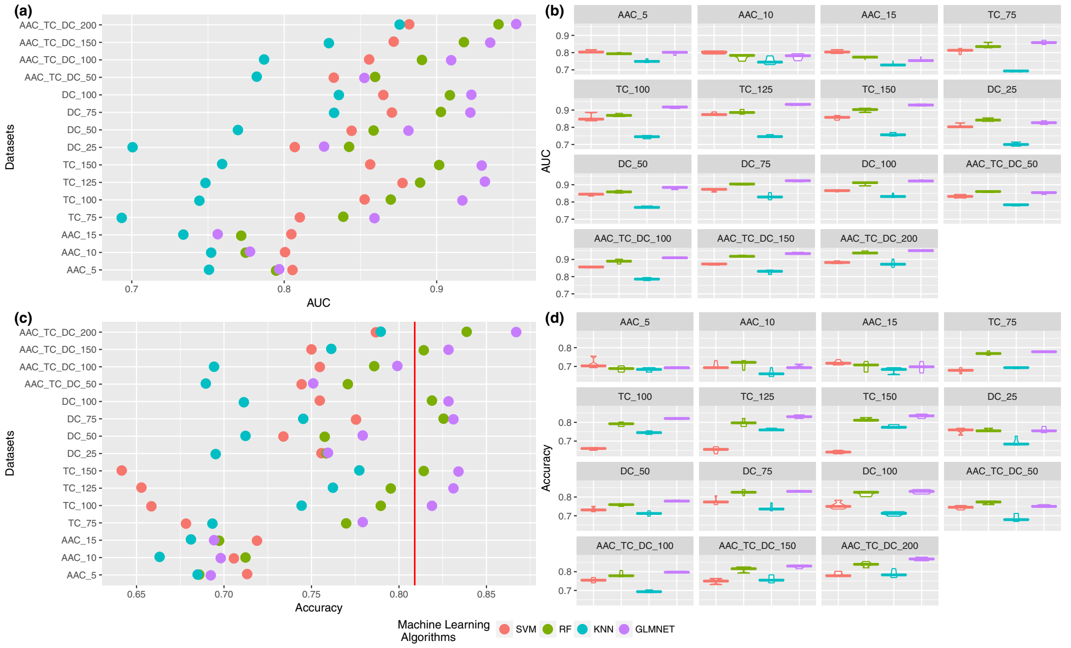


**Figure 4.** Results obtained in the feature selection process. (**a**) Summary of the performance of the four algorithms (AUC), (**b**) boxplot of the behavior of each model across experiments (AUC), (**c**) summary of the performance of the four algorithms (accuracy), and (**d**) boxplot of the behavior of each model across experiments (accuracy). The red line represents the best previously published value in the literature by Ramaprasad *et al.*[8].

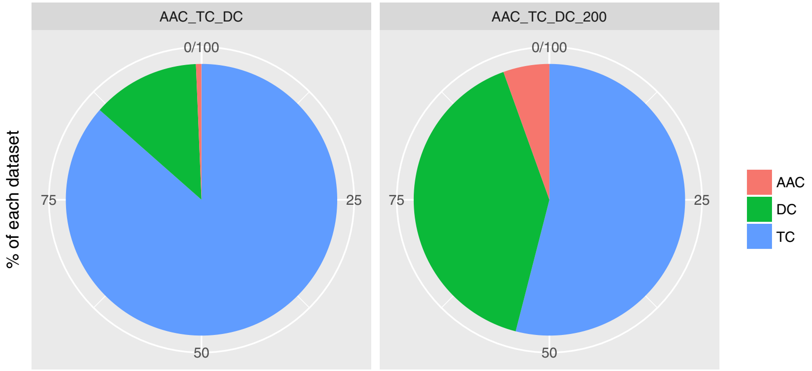


**Figure 5.** Percentage of the variables of each descriptor in the best-performing dataset before (3058 variables) and after (200 variables) the feature selection approach. An increase in the relative quantity of the AAC and DC descriptors is observed.

## Discussion

The top-40 features of the winning model are shown, for clarity reasons, in Fig. 7. We add the beta value (importance) of the glmnet model for each fold and experiment to understand the global importance of each feature. We plotted in different colors to show features belonging to a particular original dataset to clarify the feature selection process. We mentioned before in Fig. 2a that the use of AAC, TC or DC features alone or in groups of two is not enough because the datasets, in general, are noisy. In fact, our results show that a feature selection process with a combination of them can outperform our previous results and, more importantly, those in the literature. Furthermore, the best individual subgroup results were found with the AAC dataset followed by DC and TC, as shown in Fig. 2a. The discriminatory power found in the feature selection process is shown in Fig. 7, where the combination of the most informative features from different datasets, mostly from DC, allowed us to increase the knowledge about peptides.

The variables shown in Fig. 7 represent the proportion of residues, the sequences formed by two and three amino acids.

The three residues with a negative sum of betas obtained in our model (C, S and P) have also been reported by Ramaprasad *et al.*[8]. Furthermore, Karagiannis *et al.*[9] reported that the CXC domain is prevalent in anti-angiogenic activity, which supports the presence of C residues in our model. However, it has been reported that residues such

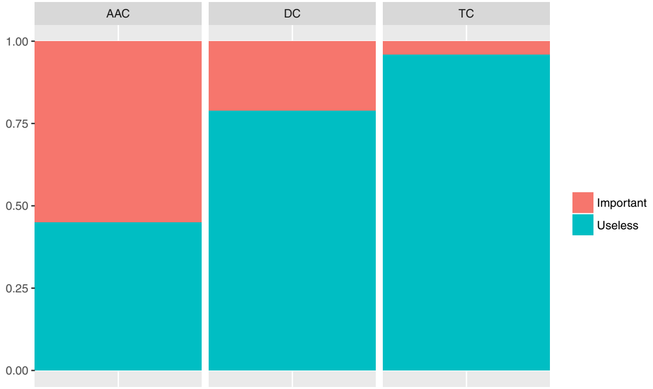


**Figure 6.** Relative proportion of the discarded variables (in blue) of the descriptor after applying the FS approach in the best-performing dataset.

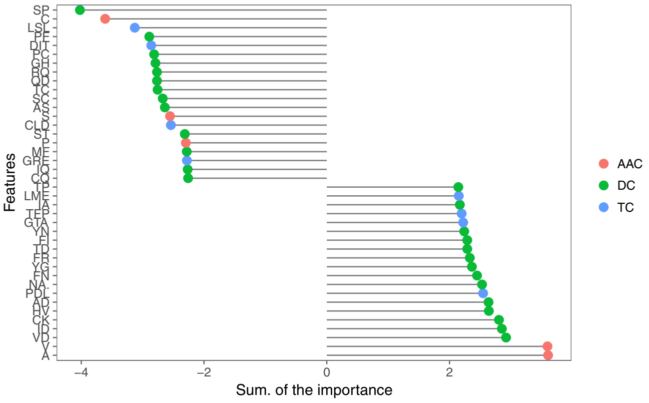


**Figure 7.** Variable importance of 200 features of the glmnet algorithm.

as Val and Ala are prevalent in non-anti-angiogenic peptides[8], and in this study, these two residues obtain the major score of betas in this peptide activity.

In addition, the model has associated the presence of sequences formed by two or three amino acids with great power of discrimination for anti-angiogenic activity. In previous studies, the analysis of motifs in anti-angiogenic peptides has shown that motifs such as CG-G, TC, SC, SP-S, W-S-C, WS-C are most predominant in this type of peptide[8]. In our model, sequences such as SP, TC and SC also have great importance. In addition, the work reported by Vazquez RodrÃguezÂguez, G. *et al.*[35] shows a list of peptide sequences with vasoinhibins activity. After a thorough analysis of their sequences, it has been observed that there is a high prevalence of sequences such as SP, SC, MF, IQ, CQ and HG, all of which have been reported in this work.

Corresponding to sequences formed by three amino acids, LSL and GRE sequences have been found in the Anginex peptide, an artificial peptide with high anti-angiogenic activity[36]. Although these sequences do not belong to the functional part of the protein, as Dings *et al.*[28] reported, they have a role in the creation of bonds that generate and stabilize the secondary structure of the protein. In addition, LSL sequences have great prevalence in aminoacidic sequences of vasoinhibins, as reported by Vazquez Rodríguezguez, G. *et al.*[35].

From a review of the state of the art, it seems that the top 20 variables with a negative sum of betas are related to anti-angiogenic activity, while a positive sum of betas is related to non-anti-angiogenic activity. Therefore, sequences with a positive sum of betas, such as PF, DIT, PC, GH, RQ, QD, AS, CLD and ST, which have been reported in this work but without any reference in the literature, can generate new knowledge on amino acid composition in anti-angiogenic activity.

The information and knowledge derived from this study are not based on any biological assumption. This study has therefore generated a new approach regarding the amino acid composition of anti-angiogenic peptides. This is a poorly known factor to which few scientific studies have paid attention. In addition, this work adds several levels of complexity in the study of this matter. First is the presence of molecular descriptors that refer to sequences formed by three amino acids, which has never been reported. Second is the union of variables from different descriptors, increasing the molecular information in a unique dataset. Finally, the implementation of FS approaches has significantly helped improve the performance of the models, exceeding on several occasions the model reported by Ramaprasad *et al.*[8], which has been the one with the highest accuracy.

## Conclusions

This paper presents classification models of anti-angiogenic peptides with the best performance reported to date using four different machine learning models and molecular descriptors obtained through the Rcpi package of Bioconductor.

The results obtained in this work support those obtained in the literature related to anti-angiogenic activity such as Ramaprasad *et al.*[8]. After analyzing the importance of the variables, the model considers that the presence of the amino acid sequences SP, LSL, PF, DIT, PC, GH, RQ, QD, TC, SC, AS, CLD, ST, MF, GRE, IQ, CQ and HG is critical to distinguish where a peptide exhibits anti-angiogenic activity.

Because the model shown in this article did not work at any time under biological assumptions, it provides a more comprehensive approach than biological studies that failed to decipher their results. This is why further studies are needed to demonstrate the existing biological relationship between the variables most related to the anti-angiogenic activity presented in this work and their possible biological interactions. In addition to this approach, more *in silico* studies are necessary that examine the interaction between these peptides and the target proteins of our organism.

Since user-friendly and publicly accessible web servers represent the future direction for practically developing more useful models[37–40], we shall endeavor in our future work to provide a web server for the method presented in this paper.

## Materials and Methods

**Dataset.** The data were obtained from Ramaprasad *et al.*[8]. This dataset represents a list of peptide sequences classified according to their activity into two classes: anti-angiogenic and non-anti-angiogenic. The number of sequences in each class is 107. None of the peptides have an identity equal to or greater than 70% with any other of the positive peptides. The understanding of the biological process of angiogenesis is critical to understand how malignant tumors are formed in the body. The peptides were collected from various research articles and patents (https://doi.org/10.1371/journal.pone.0136990.s007). As there is no source of experimentally proven non-anti-angiogenic peptides, the authors extracted a similar number of random peptide regions from proteins from the Swiss-Prot database[41] and treated them as non-anti-angiogenic peptides (https://doi.org/10.1371/journal.pone.0136990.s003). Though some of these randomly selected peptides could be anti-angiogenic in nature, the probability is very low.

The dataset consists of 107 peptide sequences classified as anti-angiogenic and 107 as non-anti-angiogenic. Following an initial check to ensure that all sequences presented a correct nomenclature, two of them were found to be erroneous and were eliminated from the database. The study therefore consisted of a total of 107 anti-angiogenic and 105 non-anti-angiogenic sequences. This type of balanced data is most suitable for use as inputs to the algorithms, as we ensure that there is no probabilistic tendency to classify a peptide in a specific class. The set of sequences were converted into three types of physicochemical descriptors (see the materials and methods) based on the primary sequence of the peptides. This way, we converted these sequences into a mathematical description of the aminoacidic composition of each peptide. After removing the variables of each descriptor that presented zero value in all observations, 2645 variables were obtained for TC, 20 variables for AAC and 393 variables for DC.

To conduct this study, amino acid sequences of peptides classified according to anti-angiogenic or non-anti-angiogenic activity were used. These sequences were converted into molecular descriptors, from the Rcpi[42] package present in the Bioconductor project[43]. Subsequently, the datasets were subjected to multivariate analysis and classification methods based on machine learning algorithms to determine the model that presents the best performance in the classification of these peptides.

**Obtaining molecular descriptors.** For the comparison and classification of peptides according to their activity, additional information from their sequence must be extracted. The package Rcpi[42], presented in the Bioconductor project[43], offers the possibility of obtaining structural and physicochemical characteristics of peptides from their amino acid sequences. In addition, it is highly interesting to gather information on characteristics as heterogeneously as possible to try to best determine all molecular information. Thus, machine learning algorithms, underlying knowledge from the data, will be able to obtain information covering as much solution space as possible.

In this study, we have worked with three types of molecular descriptors that have been calculated from different amino acid sequences: amino acid composition (AAC), dipeptide composition (DC) and tripeptide composition (TC)[44]. These descriptors are characterized by describing the composition of the amino acid sequence by easily interpretable variables. Below is a brief description of each set of descriptors.

*AAC.* This descriptor calculates the composition of each amino acid within the sequence. It reports an output with a total of 20 features/dimensions, each corresponding to an amino acid. The composition of each amino acid is obtained with the fraction of each type of amino acid within the peptide sequence[44]. Thus, Equation 1 is used for calculating the fraction of the 20 natural amino acids:

$$\text{Fraction of aai} = \frac{\text{total number of amino acids of type i}}{\text{total number of amino acids in protein}} \tag{1}$$

where i is a specific type of amino acid.

*DC.* This type of protein descriptor calculates the percentage present in each sequence of all possible combinations of the 20 amino acids pairs. Because there exist in nature 20 different amino acids, there exist 400 possible pairs ($20^2$). Therefore, a count is made of the pairs of adjacent amino acids found in each sequence. We adopted the same dipeptide composition-based approach as[44], which involves calculation of the following equation, as a fraction of amino acids considering their local order (Equation 2):

$$\text{Fraction of DC(i)} = \frac{\text{total number of DC(i)}}{\text{total number of all possible dipeptides}} \tag{2}$$

where DC(i) is one dipeptide i of the 400 possible dipeptides.

*TC.* similar to DC, TC calculates the percentage of all possible tripeptides that can be found in a sequence. The tripeptide composition was used to transform the variable length of proteins to fixed-length feature vectors. The tripeptide composition gave a fixed pattern with length equal to $20^3$.

**Machine learning.** Machine learning is the field of study interested in the development of computational algorithms capable of transforming data into intelligent actions. This field is extensive in several areas, as it helps explain and extract specific knowledge from a set of data that humans would not be able to achieve. The algorithms used are designed to perform a probabilistic search working in large spaces that involve states that can be represented by datasets. There are two main types of learning: supervised and unsupervised. The main difference between them is that in the former, learning occurs via labeled observations, while in the latter, the examples are not labeled, and the algorithm seeks to cluster the data into different groups. In this study, we will work with supervised classification algorithms from a set of labeled examples; these algorithms try to assign a label to a second set of examples.

We used four different implementations of the following machine learning algorithms: random forest (RF)[45], K-nearest neighbors (k-NN)[46], support vector machine (SVM)[47] and a generalized linear model (glmnet)[48].

Each of these machine learning algorithms has a particular set of hyperparameters that should be tuned to find the best possible combination and, consequently, the best prediction of and solution to the problem. Machine learning algorithms are very powerful techniques, but the training process is critical. This kind of algorithm learns through samples, so the same samples should not be used for learning, validation or hyperparameter tuning. We explain in further detail our robust experimental design in the experimental design section.

Random forest (RF) was developed by Breiman[45] and consists of an ensemble of independent decision trees based on random resampling of the variables for the construction of each tree. A majority vote of the trees in classification is taken as the prediction. Thus, RF adds an additional layer of randomness to a conventional bagging approach.

A search was made of the appropriate values for the parameters mtry (number of variables randomly sampled in each division of the data) and nodesize (minimal size of the terminal nodes). The range for the number of variables was established between 1 and, as the upper limit, the square root of the number of variables with the largest dataset. The minimal size of the terminal nodes ranged between 1 and 3. Low values for this parameter provide great growth and depth of each tree, improving the accuracy of predictions. In addition, the number of trees was 1000. A large number of trees ensures that each observation is predicted at least several times.

The K-nearest neighbor (k-NN) algorithm is a technique based on cluster theory. It is a very basic algorithm, but it has been reported to yield excellent results for classification. In this case, we used a variant called weighted k-NN[49]. It is based on the fact that a new observation that is particularly close to an observation within the learning set should have a great weight in the decision and, conversely, an observation that is at a farther distance will have a much smaller weight[50]. The observations are mapped following the Minkowski distance.

For this algorithm, only the hyperparameter k has been tuned, which represents the number of neighbor data points that are considered closest. Because a very high k can cause over-training of the model, the decision was made to maintain intermediate levels. The range of values used was from 1 to 5.

The objective of support vector machines (SVMs) in binary classification problems is to obtain the best hyperplane that separates the two classes, thus minimizing the error. The hyperplane is defined through support vectors. Since most real problems do not have a linear relationship, the SVM algorithm offers the possibility of calculating a kernel function to map the data in a greater number of dimensions, making it possible to linearly separate the data[51]. There are different kernel functions. For this study, the kernel function RBF (Gaussian radial basis) was used.

The values for the hyperparameters C and sigma were searched, both with a range between $2^{-12}$ and $2^{12}$ with a step size of one–i.e., $2^{-12}$, $2^{-11}$, $2^{-10}\ldots$. The modification of C implies an adjustment of the penalty of the misclassified observations. Sigma represents the standard deviation of the Gaussian distribution.

Logistic regressions are popular classification algorithms in machine learning problems when the response variable is categorical. The logistic regression algorithm represents the class-conditional probabilities through a linear function of the predictors. In this study, we use a fast regularization algorithm that fits a generalized linear model with elastic-net penalties, called glmnet. The algorithm was developed by Tibshirani *et al.*[48]. The elastic-net penalty can tend towards the lasso penalty[52] to the ridge penalty[53]. The ridge penalty is known to shrink the coefficients of correlated predictors towards each other, while the lasso tends to pick one of them and discard the others. Therefore, the elastic-net penalty mixes these two.

The grids of alpha and lambda for tuning are (0.0001, 0.001, 0.01, 0.1, 1) and (0, 0.15, 0.25, 0.35, 0.5, 0.65, 0.75, 0.85, 1), respectively. Alpha controls the elastic-net penalty, from lasso ($\alpha = 1$) to ridge ($\alpha = 0$). The lambda parameter controls the total force of the penalty.

**Feature selection.** The number of high-dimensional datasets is skyrocketing, and some of the features are redundant or noisy. To better explore the space of solutions, the number of useless features should be reduced as much as possible. The ultimate goal of the FS approaches is to find a subset of features from the original that contains as much information as possible without altering the original representation of the data. Furthermore, this subset should increase or at least not decrease the performance of the models. At the same time, it should prevent over-fitting and allow the fastest generation of better models[54]. Therefore, the redundant, noisy variables[12,54,55] are eliminated along with, generally, the variables that are more correlated without providing new information.

In machine learning, there are three main approaches for FS, known as filter, wrapper and embedded[55]. The main difference between the filter approach and the other two is that the filter approach searches for the features selected independently of the classification algorithm, while the wrapped and embedded approaches search for the feature selection depending on the classification algorithm.

Filter approaches obtain a score that measures the relevance of the features against the class vector by observing only the intrinsic properties of the data without taking any assumptions from the classifiers. In addition, this approach is computationally simple and fast. It is especially relevant for high-dimensional data. As these approaches are independent from the classification algorithm, the subset of selected features is used as the input to any algorithm. There are two different filter approaches: univariate and multivariate. Univariate filter approaches are fast, scalable and independent of the classifier but ignore feature dependencies and interaction with the classifier. Multivariate models feature dependencies with independence of the classifier and are thus computationally better than wrapper methods. The reasons for using filter feature selection (univariate) in peptide prediction are its easy-to-understand output feature ranking, higher speed than multivariate approaches and ease of validation by biological lab methods and the fact that experts usually do not need to consider descriptor interactions[55–57]. Therefore, this approach allows us to perform a better comparison among the different classification models[55] and particularity of each feature, independently of the particular behavior of each technique.

Therefore, we followed a univariate filter FS approach, and for the calculation of the relevance of the variables, a T-test was used. The T-test is one of the most robust parametric univariate statistical tests and one of the most widely used in the literature. Several sizes of features have been extracted from each of the three sets of descriptors under study—in a growing approximation, the minimal number of features most suitable to solving the problem.

**Experimental design.** The experimental design of this work is based on the classification of peptides into two different classes: anti-angiogenic and non-anti-angiogenic. The dataset consists of primary amino acid sequences of two classes of peptides, represented by the nomenclature of a letter (A, R, N, etc.).

We used the Rcpi[42] package from the Bioconductor project[43] to calculate different descriptors for each sequence (AAC, DC and TC). In addition, the set of descriptors was merged to obtain the subgroup of descriptive variables coming from each descriptor with a greater prediction capability for the peptide activity, which was saved in a unique database. Finally, the data were standardized so that the distribution of the sample has an average equal to zero and a standard deviation equal to one. The output obtained from the FS approach was used in the training and evaluation of the different classification algorithms.

A nested resampling was used for the training of the models. The characteristic of this process is the presence of an independent internal cross-validation (2/3 for training and 1/3 for validation) for the selection of the best hyperparameters of each algorithm and an independent external cross-validation (5 repetitions of a 10-fold-CV) to evaluate the model in a general way. For each 10-fold-CV experiment, the peptide sequences were randomly divided into ten sets. Nine sets were used for training the model, and the remaining set was used for testing. The process was then repeated ten times such that each set was used once as a test set. The average performance of all ten sets was reported as the final performance of the method. We repeated this process 5 times for each ML algorithm, and we presented the mean average of the 5 runs in the figures of the paper.

The performance of the different experiments was determined through the package mlr[58]. This package facilitates the design of machine-learning-based experiments, reducing the amount of scripting needed and providing a simpler and more manageable platform for development while facilitating reproducibility and replicability. Moreover, this package ensures that the execution of the machine learning algorithms follows the experimental design under the same conditions, thus allowing the comparison under equality of conditions. For the evaluation of the models, we used accuracy (to compare our findings with the state of the art) and the area under the ROC (AUC) to control for type I and II errors.

Finally, the finding of the best results and the analysis of the statistical significance of the results were carried out by means of null hypothesis tests. Furthermore, the importance of each particular descriptor in the best final model was analyzed and compared to previous findings in the literature.

## Data Availability

The R script code for dataset generation is available for download at https://doi.org/10.6084/m9.figshare.6016994.

## References

1. Rosca, E. V. *et al.* Anti-angiogenic peptides for cancer therapeutics. *Current pharmaceutical biotechnology* **12**, 1101–16 (2011).
2. Coras, B. *et al.* Antiangiogenic therapy with pioglitazone, rofecoxib, and trofosfamide in a patient with endemic Kaposi sarcoma. *Archives of dermatology* **140**, 1504–1507 (2004).
3. Quiroz-Mercado, H., Martinez-Castellanos, M. A., Hernandez-Rojas, M. L., Salazar-Teran, N. & Chan, R. V. P. Antiangiogenic therapy with intravitreal bevacizumab for retinopathy of prematurity. *Retina* **28**, S19–S25 (2008).
4. Carmeliet, P. & Jain, R. K. Angiogenesis in cancer and other diseases. *Nature* **407**, 249–257 (2000).
5. Ucuzian, A. A., Gassman, A. A., East, A. T. & Greisler, H. P. Molecular mediators of angiogenesis. *Journal of burn care & research: official publication of the American Burn Association* **31**, 158 (2010).
6. Vasudev, N. S. & Reynolds, A. R. Anti-angiogenic therapy for cancer: Current progress, unresolved questions and future directions (2014).
7. Al-Husein, B., Abdalla, M., Trepte, M., DeRemer, D. L. & Somanath, P. R. Antiangiogenic therapy for cancer: An update (2012).
8. Ramaprasad, A. S. E. *et al.* Antiangiopred: a server for prediction of anti-angiogenic peptides. *PloS one* **10**, e0136990 (2015).

9. Karagiannis, E. D. & Popel, A. S. A systematic methodology for proteome-wide identification of peptides inhibiting the proliferation and migration of endothelial cells. *Proceedings of the National Academy of Sciences* **105**, 13775–13780 (2008).
10. Dings, R. P., Nesmelova, I., Griffioen, A. W. & Mayo, K. H. Discovery and development of anti-angiogenic peptides: A structural link. *Angiogenesis* **6**, 83–91 (2003).
11. Koskimaki, J. E. *et al*. Serpin-derived peptides are antiangiogenic and suppress breast tumor xenograft growth. *Translational oncology* **5**, 92–97 (2012).
12. Fernandez-Lozano, C. *et al*. Classification of signaling proteins based on molecular star graph descriptors using Machine Learning models. *Journal of Theoretical Biology* **384**, 50–58 (2015).
13. Fernandez-Lozano, C. *et al*. Improving enzyme regulatory protein classification by means of SVM-RFE feature selection. *Molecular BioSystems* **10**, 1063 (2014).
14. Tang, H., Su, Z.-D., Wei, H.-H., Chen, W. & Lin, H. Prediction of cell-penetrating peptides with feature selection techniques. *Biochemical and biophysical research communications* **477**, 150–154 (2016).
15. Kandemir Çavaş, Ç. & Yildirim, S. Classifying ordered-disordered proteins using linear and kernel support vector machines. *Turkish Journal of Biochemistry* **41**, 431–436 (2016).
16. Wei, L. *et al*. Cppred-rf: A sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *Journal of Proteome Research* **16**, 2044–2053, PMID: 28436664 (2017).
17. 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* bty451 (2018).
18. Fernáandez-Blanco, E., Aguiar-Pulido, V., Munteanu, C. R. & Dorado, J. Random forest classification based on star graph topological indices for antioxidant proteins. *Journal of theoretical biology* **317**, 331–337 (2013).
19. Fernandez-Lozano, C. *et al*. Kernel-based feature selection techniques for transport proteins based on star graph topological indices. *Current topics in medicinal chemistry* **13**, 1681–1691 (2013).
20. Chen, L., Chu, C., Huang, T., Kong, X. & Cai, Y.-D. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. *Amino acids* **47**, 1485–1493 (2015).
21. Qureshi, A., Tandon, H. & Kumar, M. Avp-ic50pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (ic50). *Peptide Science* **104**, 753–763 (2015).
22. Fernandez-Lozano, C. *et al*. Markov mean properties for cell death-related protein classification. *Journal of theoretical biology* **349**, 12–21 (2014).
23. Aguiar-Pulido, V. *et al*. Naïve bayes qsdr classification based on spiral-graph shannon entropies for protein biomarkers in human colon cancer. *Molecular BioSystems* **8**, 1716–1722 (2012).
24. 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. *Journal of theoretical biology* **257**, 303–311 (2009).
25. Liu, Y. *et al*. Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. *Expert Systems with Applications* **72**, 306–316 (2017).
26. Rosca, E. V., Lal, B., Koskimaki, J. E., Popel, A. S. & Laterra, J. Collagen iv and cxc chemokine derived anti-angiogenic peptides suppress glioma xenograft growth. *Anti-cancer drugs* **23**, 706 (2012).
27. Xu, Y. *et al*. A novel antiangiogenic peptide derived from hepatocyte growth factor inhibits neovascularization *in vitro* and *in vivo* (2010).
28. Dings, R. P. & Mayo, K. H. A journey in structure-based drug discovery: from designed peptides to protein surface topomimetics as antibiotic and antiangiogenic agents. *Accounts of chemical research* **40**, 1057–1065 (2007).
29. Wickham, H. ggplot2: Elegant Graphics for Data Analysis, http://ggplot2.org (Springer-Verlag New York, 2009).
30. Liu, B. *et al*. Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences. *Nucleic Acids Research* **43**, W65–W71 (2015).
31. Kumar, R., Kumari, B. & Kumar, M. Prediction of endoplasmic reticulum resident proteins using fragmented amino acid composition and support vector machine. *Peer J* **5**, e3561 (2017).
32. Zhang, W. *et al*. Accurate prediction of immunogenic t-cell epitopes from epitope sequences using the genetic algorithmbased ensemble learning. *Plos One* **10**, 1–14 (2015).
33. Zubek, J. *et al*. Multi-level machine learning prediction of protein–protein interactions in Saccharomyces cerevisiae. *Peer J* **3**, e1041 (2015).
34. Fernandez-Lozano, C., Gestal, M., Munteanu, C. R., Dorado, J. & Pazos, A. A methodology for the design of experiments in computational intelligence with multiple regression models. *Peer J* **4**, e2721 (2016).
35. Rodriguez, G. V., Gonzalez, C. & Rodriguez, A. D. L. Novel fusion protein derived from vasostatin 30 and vasoinhibin ii-14.1 potently inhibits coronary endothelial cell proliferation. *Molecular biotechnology* **54**, 920–929 (2013).
36. Griffioen, A. W. *et al*. Anginex, a designed peptide that inhibits angiogenesis. *The Biochemical journal* **354**, 233–242 (2001).
37. Wei, L., Xing, P., Shi, G., Ji, Z. L. & Zou, Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1–1 (2018).
38. Wei, L., Xing, P., Tang, J. & Zou, Q. Phospred-rf: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE transactions on nanobioscience* **16**, 240–247 (2017).
39. Wei, L., Wan, S., Guo, J. & Wong, K. K. A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* **83**, 82–90 (2017).
40. Xing, P., Su, R., Guo, F. & Wei, L. Identifying n 6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Scientific reports* **7**, 46757 (2017).
41. Consortium, T. U. Activities at the universal protein resource (uniprot). *Nucleic Acids Research* **42**, D191–D198 (2014).
42. 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).
43. Gentleman, R. C. *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).
44. Bhasin, M. & Raghava, G. P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *Journal of Biological Chemistry* **279**, 23262–23266 (2004).
45. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).
46. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory* **13**, 21–27 (1967).
47. Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273–297 (1995).
48. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1 (2010).
49. Hechenbichler, K. & Schliep, K. Weighted k-nearest-neighbor techniques and ordinal classification (2004).
50. Liu, W. & Chawla, S. Class confidence weighted knn algorithms for imbalanced data sets. *Advances in Knowledge Discovery and Data Mining* 345–356 (2011).
51. Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* **2**, 121–167 (1998).
52. Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 267–288 (1996).
53. Saunders, C., Gammerman, A. & Vovk, V. Ridge regression learning algorithm in dual variables. *In ICML* **98**, 515–521 (1998).

54. Yu, L. & Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. *In ICML* **3**, 856–863 (2003).
55. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in. *Bioinformatics* **23**, 2507–2517 (2007).
56. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
57. Estevez, P. A., Tesmer, M., Perez, C. A. & Zurada, J. M. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks* **20**, 189–201 (2009).
58. Bischl, B. *et al.* Machine Learning in R. *Journal of Machine Learning Research* **17**(170), 1–5 http://jmlr.org/papers/v17/15-066.html (2016).

## Acknowledgements

This work is supported by 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 (Ref. ED431G/01, ED431D 2017/16), the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23), and the Spanish Ministry of Economy and Competitiveness via funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union and the "Juan de la Cierva" fellowship program supported by the Spanish Ministry of Economy and Competitiveness (Carlos Fernandez-Lozano, Ref. FJCI-2015-26071).

## Author Contributions

J.L.B., A.B.P.-P., A.P. and C.F.-L. conceived the experiment(s), J.L.B. and C.F.-L. conducted the experiment(s), and J.L.B., A.B.P.-P., A.P. and C.F.-L. analyzed the results. J.L.B. and C.F.-L. wrote the paper. J.L.B., A.B.P.-P., A.P. and C.F.-L. reviewed the manuscript. All authors read and approved the manuscript.

## Additional Information

**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.

# 4

# Discusión

En conjunto, los tres artículos que componen esta tesis estudian la aplicación de técnicas de Machine Learning para el diagnóstico y tratamiento oncológico de precisión mediante el análisis de datos ómicos. Siguiendo el hilo argumental de los tres artículos y los resultados obtenidos, esta sección aborda aspectos específicos y transversales generados a lo largo de la investigación.

La identificación de alteraciones en algún tipo concreto de cáncer se ha convertido en una tarea altamente compleja debido a la cantidad de modificaciones posibles en los procesos cancerígenos. Para ello es imprescindible un abordaje multidisciplinar. Las alteraciones que presentan un rol principal en el desarrollo de tumores pueden encontrarse a cualquier nivel ómico, incrementando la complejidad del problema. En este escenario, también aumentan las oportunidades de desarrollo de nuevas técnicas de diagnóstico y nuevos tratamientos oncológicos de precisión. La motivación para el desarrollo de la presente tesis reside en el abordaje de este problema mediante técnicas computacionales. El análisis de datos ómicos procedentes de tumores, el *screening* automático y el reposicionamiento de fármacos anti-tumorales presentados en esta tesis ofrecen nuevas soluciones con el fin de alcanzar un diagnóstico temprano y un tratamiento personalizado para cada paciente.

Gracias al desarrollo de las nuevas tecnologías de secuenciación, y principalmente a su abaratamiento, se han podido caracterizar molecularmente muchos tipos de tumores que se desarrollan en los seres humanos. Desgraciadamente, esto aún no ha sido suficiente para disminuir la tasa de mortalidad de esta enfermedad en ciertos tipos de tumores.

Actualmente, la forma más eficiente de sobrevivir a un cáncer sigue siendo detectarlo en etapas tempranas de su desarrollo. La dificultad reside en que el diagnóstico y la posterior búsqueda del tratamiento ha de ser personalizada para cada paciente.

Afortunadamente, dos campos de estudio, aparentemente independientes, como son la biomedicina y la informática, han experimentado un crecimiento en la última década. Por un lado, el campo de la biomedicina está aportando una inmensa cantidad de datos. Por el otro lado, en el campo de la informática se están desarrollando una serie de herramientas capaces de manejar y extraer información de grandes y complejos conjuntos de datos. La fusión de estos dos campos ha sido imprescindible en el avance de la medicina del siglo XXI, sentando las bases para el desarrollo de la biomedicina computacional.

Grandes consorcios internacionales como el TCGA, han puesto de forma pública datos multi-ómicos generados a partir de un elevado número de pacientes. Este enfoque ofrece la posibilidad de desarrollar nuevas aproximaciones que sean capaces de extraer nuevo conocimiento biológico en grandes cohortes de pacientes.

Gracias al libre acceso de estos datos, se han desarrollado numerosas ramas computacionales. Entre ellas, los modelos basados en Machine Learning son los que más potencial tienen para generar un punto de inflexión en el campo de la biomedicina. El contexto generado por la actual investigación biomédica ha propiciado que los desarrollos realizados en el campo de la Inteligencia Artificial se utilicen para resolver los problemas de manejo y procesado de grandes volúmenes de datos masivos en biomedicina. Como ejemplos de grandes hitos alcanzados en este campo se encuentra la identificación de la molécula *halicin* con función antibacteriana [107] y el desarrollo de AlphaFold [108] para la predicción del plegamiento de proteínas. Para el descubrimiento de la *halicin* como agente antibacteriano, entrenaron un modelo de ML basado en grafos moleculares para predecir dicha actividad en múltiples moléculas. Este modelo fue utilizado en grandes repositorios de datos públicos con el fin de encontrar moléculas candidatas. Inesperadamente, la molécula *halicin* presentó valores significativos en la predicción y tras su validación experimental, se observó su actividad antibacteriana. En segundo lugar, en 2021, el equipo de DeepMind [109], desarrolló AlphaFold [108]. AlphaFold es un algoritmo de ML desarrollado con el fin de predecir la estructura 3D de una proteína basada únicamente en su secuencia aminoacídica. Obtuvo el primer puesto del ranking de la competición CASP14 [110] con grandes diferencias sobre el segundo puesto.

Estos logros ponen de manifiesto la importancia y el papel principal que desempeñan estas técnicas computacionales de ML en el abordaje de problemas biológicos, antes inalcanzables utilizando técnicas convencionales. Las técnicas de ML ya han obtenido muy buenos resultados en problemas en la búsqueda de fármacos [107] y en la predicción de la estructura 3D de una proteína [108], como se comentó previamente. Lo común en estas dos estrategias ha sido la utilización de vastas cantidades de datos para la resolución del problema. En dominios basados en el análisis de datos ómicos se están haciendo grandes esfuerzos en la identificación de pacientes que responden a un tratamiento, en la búsqueda de biomarcadores para diagnóstico temprano, o por ejemplo, en la reclasificación de subtipos tumorales. Uno de los artículos incluido en esta tesis aborda el problema comentado mediante una extensa revisión de trabajos previos publicados [1].

El uso de técnicas de ML en este tipo de problemas presenta una limitación principal, el tamaño muestral. Así como para los anteriores dos artículos mencionados existe un volumen de observaciones (en este caso moléculas) muy amplio, los trabajos basados en cohortes de pacientes oscilan entorno al millar de muestras en el mejor de los casos. Este factor es limitante a la hora de obtener modelos de ML que puedan ser capaces de generalizar en conjuntos de pacientes que no se utilizaron para entrenar el modelo. Por otro lado, los datos ómicos presentan una gran dimensionalidad. Si se atiende por ejemplo a datos RNASeq, este tipo de datos presentan alrededor de 50.000 variables. Si se quiere añadir una capa de complejidad incluyendo datos de metilación, se añaden, según el tipo de secuenciación, entre 27.000 y 450.000 variables. Además, a medida que se mejoran las tecnologías de secuenciación, el número de variables que se generan también va en aumento.

Estos problemas donde el número de observaciones es mucho menor que el número de características que las definen ($n << p$), deben de ser tratados con una rigurosa metodología para evitar un sobreentrenamiento de los modelos y sesgos en los resultados. Por lo tanto, es necesario seguir un diseño experimental muy estricto, con numerosos procesos de validación. Basado en el diseño experimental presentado por Fernández-Lozano et al. [68] y en la comparación metodológica realizada en [1], esta tesis aplica una metodología robusta y reproducible siguiendo las directrices de ambos trabajos. En este contexto, es esencial la selección de características relevantes en los conjuntos de datos para la eliminación de características redundantes e inútiles en el análisis y evitar la denominada *Curse of dimensionality* [111].

Este concepto se basa en que a medida que exista un mayor número de variables en los datos de entrenamiento, los modelos de ML van a tender a sobreentrenarse. Si pensamos en un problema de clasificación binaria, por ejemplo, al modelo le será relativamente fácil encontrar diferencias entre las dos clases, ya que la búsqueda se realiza en todas las dimensiones disponibles. Las diferencias encontradas van a ser intrínsecas al conjunto de datos de entrenamiento (que siempre va a ser una muestra pequeña del total). La validación de dicho modelo en cohortes externas presentará un rendimiento pobre con respecto al entrenamiento. Por lo tanto, es necesario simplificar los problemas a un menor número de dimensiones para poder obtener modelos capaces de generalizar.

Las técnicas de selección de características tuvieron un papel principal en uno de los artículos presentados en el compendio de esta tesis [3] (ver sección 3.3). En este trabajo se utilizan técnicas de filtrado univariado para identificar las características más relevantes capaces de diferenciar las muestras por su actividad anti-angiogénica. Los resultados muestran un aumento significativo en el rendimiento de los modelos tras realizar la selección de características. La estrategia seguida fue la generación de descriptores moleculares a partir de las secuencias aminoacídicas de los péptidos. De esta manera, se es capaz de identificar secuencias de aminoácidos que tengan un rol importante en la actividad proteica. Al existir un total de 20 aminoácidos en la naturaleza, la cantidad de descriptores se veía incrementada del orden de $20^n$. Por ejemplo, para generar descriptores de cada posible tripéptido, el número de variables asciende a 8.000. La selección de características en este escenario permite obtener las variables más significativas para el problema y eliminar aquellas que no presenten información alguna.

En biomedicina, es crucial obtener resultados donde se pueda aportar una explicación biológica. Las técnicas *filter* permiten identificar variables relacionadas directamente con la variable respuesta. Al ser independientes del modelo, al contrario que las técnicas *wrapper* y *embedded*, las variables identificadas por las técnicas *filter* van a poder ser validadas por diferentes modelos. En caso de que diferentes modelos presenten rendimientos altos con el mismo subgrupo de variables, se puede concluir que dicho subgrupo alberga información importante para la resolución del problema. En cuanto a los otros dos tipos de técnicas mencionadas, el subgrupo de variables seleccionadas y en consecuencia el rendimiento alcanzado, estará directamente relacionado con el algoritmo utilizado. Este hecho dificulta en gran medida la explicabilidad biológica del modelo.

Por otra parte, en [2] se sigue otra estrategia para búsqueda de nuevos fármacos. Basándose en los artículos identificados en [1], se seleccionaron tres firmas genéticas previamente publicadas (ver sección 3.2). Estas firmas se obtuvieron a partir de diferentes técnicas de selección de características, presentando todas ellas un valor prognóstico para muestras de cáncer de colon. En este caso, las técnicas de ML se utilizan para validar el poder predictivo de las firmas genéticas. De esta manera, se puede inferir qué información biológica alberga el conjunto de estos genes para el problema analizado. El objetivo final es la identificación y/o estratificación de pacientes en diferentes subgrupos mediante datos moleculares, con el fin de definir tratamientos personalizados.

Los tratamientos personalizados requieren de gran inversión y tiempo. La metodología de ML y las técnicas de *docking* que se utilizan en esta investigación aportan nuevas formas de búsqueda rápida y a bajo coste de nuevos tratamientos personalizados *in silico*. Los resultados de los artículos [2, 3] son un claro ejemplo. Estos modelos reducen notablemente el coste temporal y económico en los procesos de descubrimiento [3] y validación de fármacos [2]. A diferencia de lo que se hacía anteriormente, ahora existe la posibilidad de generar modelos que añadan un filtro previo en la validación de fármacos.

En [3] se utilizan descriptores moleculares capaces de modelar la composición de los péptidos matemáticamente. Este hecho ofrece la posibilidad de crear un modelo, basado en ML, capaz de predecir la función anti-angiogénica de péptidos. El uso de fármacos anti-angiogénicos en cáncer y otras enfermedades es conocido desde hace bastante tiempo [112]. En 2004 la FDA aprobó el bevacizumab, el primer fármaco anti-angiogénico, para el tratamiento de cáncer de colon metastático. Actualmente existen más de diez fármacos anti-angiogénicos para tratar tipos específicos de tumores. Aunque los resultados tras la aplicación de estos tipos de drogas son realmente fascinantes [113, 114, 115], no existe un catálogo amplio de este tipo de fármacos. El factor limitante es el tiempo necesario en la fase de descubrimiento de fármacos. Poder reducir el tiempo gracias a *screening* computacional, es crucial para conseguir un mayor número de candidatos. En [3] se consigue este objetivo, ya que se crea el modelo que presenta el mayor rendimiento en el estado del arte para la predicción de péptidos anti-angiogénicos. Uno de factores clave de esta investigación, es que la predicción se realiza a partir de la secuencia aminoacídica de los péptidos.

Otro abordaje en el *screening* computacional de fármacos, se realizó en [2]. En este caso se aplicaron técnicas de reposicionamiento de fármacos. El objetivo de estas técnicas es buscar fármacos candidatos que puedan ser utilizados para una enfermedad diferente de la que fueron aprobados. En esta investigación se realiza la búsqueda a partir de fármacos anti-cancerígenos aprobados por la FDA. Estos fármacos fueron cruzados con los productos de una lista de genes candidatos a ser dianas terapéuticas, identificados en [1] y validados mediante experimentos de ML en [2]. Gracias a técnicas de *Docking Molecular* se estudió, de forma *in silico*, la posible interacción entre ligando (fármaco) y diana (proteína). El resultado muestra una interacción interesante entre los producto génicos del gen FABP6 y el fármaco

Abemaclicib. Dicho fármaco, aprobado previamente para cáncer de mama, podría ser ahora un posible fármaco en pacientes con cáncer de colon que presentasen niveles elevados de expresión del gen FABP6. Debido a que este fármaco ya ha sido aprobado, por lo que ha pasado las fases I y II, únicamente tendría que probarse que es eficaz para solucionar el problema propuesto. Este enfoque reduce enormemente el tiempo total de los ensayos clínicos, y abarata los costes asociados al proceso de desarrollo y aprobación del fármaco.

En conclusión, los resultados presentados en esta tesis ofrecen nuevas formas de aplicación de modelos de Machine Learning para el análisis de datos ómicos. Los resultados aportan nuevas estrategias en el diagnóstico temprano y tratamientos oncológicos de precisión. Debido a que el campo está todavía en desarrollo, se encontraron un conjunto de limitaciones del trabajo.

En primer lugar, hay que tener en cuenta que los modelos de ML han sido optimizados en campos de investigación alejados de la biomedicina. En esos casos, los problemas a resolver presentan unas características muy diferentes. Por ejemplo, el campo donde más se está utilizando estas técnicas es en las grandes compañías tecnológicas. Los datos a analizar en estas compañías, podrían definirse como datos *verticales*. Es decir, tienen un gran número de observaciones y/o muestras y un conjunto relativamente pequeño de variables. En estos casos, el entrenamiento de los algoritmos será óptimo, ya que parten de un conjunto muy gran de observaciones, y su poder de generalización va a ser mayor. Además, no se tendrá que hacer tanto énfasis en procesos de validación externa. Como se ha podido ver a lo largo de esta tesis, los problemas que propone la biomedicina tienen que abordarse partiendo de unos datos, podríamos llamarlos, *horizontales*. Bajo estas características, será mucho más difícil la capacidad de generalización de los modelos. Es aquí donde cobran especial importancia las técnicas de selección de características. Realizar la búsqueda de las características es un proceso vital para el desarrollo de modelos predictivos generales. Además, en biomedicina, los modelos deben de ser fácilmente explicables, por lo que el tipo de técnica de selección de características influye en este proceso.

Aunque es cierto que la biomedicina está generando una gran cantidad de datos, y cada vez, es más el número de pacientes involucrados en los estudios, desgraciadamente, este hecho también presenta una limitación, que es la falta de estandarización de las metodologías. Existen muchos tipos de plataformas y estrategias de secuenciación. Por lo tanto, existe una gran heterogeneidad entre cohortes de pacientes, lo que dificulta las formas de integración y la creación de meta-cohortes, con el fin de obtener un mayor tamaño muestral. Es necesario por lo tanto, evitar y corregir el denominado *batch effect*. Este concepto se refiere a las ocasiones cuando factores no biológicos (como puede ser el tipo de plataforma) causan cambios en los datos producidos por el experimento. Aunque existen a día de hoy herramientas estadísticas capaces de reducir en cierta medida este fenómeno, inevitablemente se añade un pequeño sesgo en el análisis.

Esta serie de limitaciones están presentes en prácticamente todas las aplicaciones de ML para el análisis de datos ómicos. En general, esta tesis presenta, siguiendo unas directrices basadas en trabajos previos [1], una aplicación de modelos de ML para la resolución de dos tipos de problemas biológicos: 1) identificación de biomarcadores en cáncer de colon [2] y 2) desarrollo de un modelo automático para la búsqueda de nuevos fármacos anti-tumorales [3].

# 5 Conclusiones

Se presentan en este capítulo las conclusiones extraídas tras el desarrollo de la presente tesis. Las conclusiones fueron sacadas de los tres artículos presentados así como de la investigación transversal realizada.

El objetivo de esta tesis es la aplicación de técnicas de ML para el diagnóstico y tratamiento oncológico de precisión mediante el análisis de datos ómicos. Esta investigación ha permitido formular las siguientes conclusiones:

1. En el análisis de datos ómicos con técnicas de ML se identifican patrones y variables de interés que las técnicas convencionales (tales como inferencia estadística, contraste de hipótesis o análisis de expresión diferencial) no son capaces de encontrar. El estudio bibliográfico realizado en [1] muestra como el aprendizaje supervisado es el tipo de aprendizaje más utilizado. Este aprendizaje mayoritariamente se utiliza con datos de expresión (RNASeq), siendo los modelos basados en *kernel*, específicamente el algoritmo SVM, los más utilizados. No existe un modelo significativamente mejor a otros para todos los casos. Las redes de neuronas, por su parte, están teniendo un papel importante en estos análisis, sobre todo con topologías de Deep Learning, siendo su principal factor limitante el tamaño muestral de las cohortes. La gran parte de los avances de estos algoritmos se ha realizado en datos de imágenes. Debido a la heterogeneidad de los perfiles moleculares que caracterizan los subtipos de cáncer, son necesarias técnicas integrativas para estratificar correctamente los pacientes. En general, las técnicas de ML aún están en un grado muy reciente de aplicación en estos tipos de problemas.

2. Los modelos de Machine Learning pueden ser utilizados para predecir actividades complejas de péptidos utilizando datos públicos de sus secuencias. El *screening* automático de grandes conjuntos de péptidos permite un abaratamiento significativos en el proceso de desarrollo de fármacos.

3. En concreto, el modelo desarrollado en [3] predice la actividad anti-angiogénica de péptidos con una precisión mayor que el resto de los modelos del estado

del arte hasta la fecha. Se descubrieron secuencias importantes tales como Serina-Prolina, Valina-Ac. Aspártico, Lisina-Serina-Lisina y una densidad de Cisteína, Valina ó Alanina a la hora de discernir entre la presencia o no de actividad anti-angiogénica.

4. Las metodologías de Machine Learning también pueden ser utilizadas para validar subconjuntos de genes con el fin de predecir ciertos rasgos fenotípicos. Los resultados de estos experimentos dan una aproximación de la capacidad predictiva de dichos conjuntos para determinadas condiciones clínicas.

5. Específicamente, en [2] mediante modelo de Machine Learning, se ha llegado a la conclusión de que el gen FABP6 podría ser un potencial biomarcador para cáncer de colon.

6. Así como las técnicas automáticas de *screening* abaratan los costes de producción, el reposicionamiento de fármacos anti-tumorales anula los costes tanto en experimentación como en desarrollo. En este trabajo, un estudio de reposicionamiento de fármacos propone al fármaco Abemaciclib (aprobado para cáncer de mama) como candidato dirigido hacia los productos génicos de FABP6 en pacientes con cáncer de colon.

7. Finalmente, la investigación desarrollada en esta tesis muestra la potencialidad del uso de las técnicas de Machine Learning para la resolución de problemas biológicos complejos. El manejo y análisis de datos ómicos, para el entrenamiento de modelos de ML necesita de una metodología estándar y reproducible para futuras comprobaciones. Esta tesis aplica una metodología ML robusta y reproducible para la resolución de diferentes problemas en el campo de la biomedicina.

# 6 Conclusions

This chapter presents the conclusions reached as a result of the development of this thesis. The conclusions were extracted from the three papers as well as from the transversal research conducted.

The goal of this thesis is the application of ML techniques for precision cancer diagnosis and treatment using omics data analysis. This research has allowed the formulation of the following conclusions:

1. ML techniques in omic data analysis identifies patterns and variables of interest that conventional techniques (such as statistical inference, hypothesis testing or differential expression analysis) are not able to detect. The bibliographic study carried out in [1] shows how supervised learning is the most widely used type of learning. This kind of learning is mostly used with expression data (RNASeq), with kernel based models. Specifically, the SVM algorithm is the most widely used. No single model is significantly better than others in all scenarios. Neural networks are playing an important role in these analyses, especially with deep learning topologies, with the main limiting factor being the sample size of the cohorts. Most of the progress in these algorithms has been made on image data. Due to the heterogeneity of molecular profiles characterising cancer subtypes, integrative techniques are needed to correctly stratify patients. In general, ML techniques are still at a very early stage of application in these kind of problems.

2. Machine Learning models can be used to predict complex peptide activities using publicly available sequence data. The automatic screening of large sets of peptides enables significant reduction of costs in the drug development process.

3. In particular, the model developed in [3] predicts the anti-angiogenic activity of peptides with better accuracy than all other state-of-the-art models to date. It has been shown that important sequences such as Serine-Proline, Valine-Aspartic Acid, Lysine-Serine-Lysine and a density of Cysteine, Valine or Alanine are important to discerning the anti-angiogenic activity.

4. Machine Learning methodologies can also be used to validate subsets of genes that predict certain phenotypic traits. The results of these experiments provide an approximation of the predictive capacity of these sets for certain clinical conditions.

5. Specifically, in [2], it has been concluded that the FABP6 gene could be a potential biomarker for colon cancer using Machine Learning modelling.

6. In the same manner as automated screening techniques reduce production costs, the repurposing of anti-tumour drugs eliminates both experimental and development costs. In this paper, a drug repurposing study proposes the drug Abemaciclib (approved for breast cancer) as a candidate targeting FABP6 gene products in colon cancer patients.

7. Finally, the research developed in this thesis demonstrates the potential of using Machine Learning techniques to solve complex biological problems. The management and analysis of omics data for the training of ML models requires a standard and reproducible methodology for future testing. This thesis applies a robust and reproducible ML methodology to solve different problems in the field of biomedicine.

# 7 Futuros desarrollos

El resultado de esta tesis abre una serie de cuestiones que deben de ser abordadas en futuras investigaciones. En general, se ve una tendencia claramente marcada hacia la utilización de modelos de ML como soporte a la práctica clínica. Actualmente, el campo se encuentra en fase de investigación y desarrollo, y esta tesis es una prueba de ello. En pocos años presenciaremos cómo estas tecnologías ayudarán en la toma de decisiones clínicas y automatizará procesos de diagnóstico y tratamiento. Además, el campo de inserción de las técnicas es cada vez más amplio gracias al abaratamiento de los costes [116]. Desgraciadamente, para conseguir este objetivo aún existen una serie de limitaciones y/o desarrollos que deben de ser abordados.

De forma individual, los artículos presentados en el compendio generan una serie de futuros desarrollos que se comentan a continuación.

La revisión realizada en [1] da a conocer las fortalezas y limitaciones que tiene el uso del ML para el análisis de datos ómicos. En cuanto a las limitaciones, se identifica una gran cantidad de trabajos que no presentan una estandarización metodológica y que no pueden ser reproducibles. Observando la cantidad de trabajos publicados, prácticamente ninguno es utilizado actualmente en el apoyo de decisiones clínicas. Esto denota una falta de robustez en los resultados obtenidos mediante modelos de ML. Aunque el objetivo de esta tesis es ofrecer soluciones para ello, aún quedan muchas cuestiones abiertas en el campo.

Por ejemplo, el uso de diferentes tipos de datos a la hora del diagnóstico es limitado. La mayoría de los trabajos utilizan datos de expresión genómica, pudiendo no albergar la suficiente información biológica para el fenómeno que se pretende estudiar. Es decir, la información subyacente al problema no se encuentra en los datos de expresión sino en otro tipo de dato ómico. Por otra parte, muchos datos transcriptómicos y/o proteómicos son datos dinámicos, por lo que su secuenciación en un determinado momento es análogo a una fotografía. Para un mayor entendimiento del proceso cancerígeno, es necesario un análisis desde un punto de vista temporal. Esto podría realizarse reclutando cohortes de forma longitudinal y de manera prospectiva, aunque los costes también aumentarían significativamente. Este tipo de cohortes podrían beneficiarse del empleo de algoritmos de ML, ya que existe un gran desarrollo de estas técnicas para el análisis de series temporales.

Además, se deberán tomar decisiones acerca del tipo de tejido a secuenciar. Por ejemplo, sería necesario secuenciar tejido normal no adyacente, para definir molecularmente lo que es *normal*. Por otra parte, sería ideal generar una cohorte como la del TCGA para procesos metastásicos, y otra para respuesta a tratamientos. En este contexto, los algoritmos de ML serían muy útiles para extraer patrones de tales conjuntos de datos longitudinales.

Otro aspecto que se identifica en esta tesis es la falta de modelos de ML que integren de forma correcta datos multi-ómicos, y específicamente una integración temporal de diferentes datos ómicos generados en diferentes momentos (por ejemplo, tejido normal, tejido tumoral, post-tratamiento y metástasis). Aunque existe la base teórica de cómo integrar estos datos [15] su aplicación aún es compleja, debido principalmente a la complejidad para simular las interacciones entre los diferentes niveles biológicos. Por lo tanto, es necesario un mayor desarrollo de modelos que sean capaces integrar diferentes datos ómicos.

Los resultados presentados en [2] dejan claro los futuros desarrollos a corto plazo. La interacción identificada entre Abemaciclib-FABP6 debe de ser validada experimentalmente. En primer lugar en líneas celulares tumorales y posteriormente en ratones. Además, también existe la posibilidad de ejecutar más experimentos computacionales al respecto. Por un lado, el análisis a través de técnicas de *docking* molecular que tengan en cuenta la dinámica de la molécula diana para conocer más profundamente cómo será la interacción. Por otro lado, también sería posible realizar una búsqueda a todos los niveles ómicos, con el objetivo de encontrar una firma capaz de estratificar a los pacientes según una posible respuesta al fármaco. Firmas prognósticas nos permitirían clasificar a los pacientes y enfocar aún más la búsqueda a un diagnóstico y un tratamiento de precisión. Además, sería interesante el estudio de la posible resistencia de estos fármacos según las características ómicas de los pacientes, para conocer el motivo biológico de la resistencia. Con acceso a datos multi-ómicos de pacientes se podría estudiar si esta resistencia es predecible mediante modelos de ML.

Por último, el trabajo desarrollado en [3] también propone una serie de futuros desarrollos. Siguiendo la metodología utilizada para el desarrollo de AlphaFold, se podría descargar datos de grandes repositorios y utilizar el modelo para predecir la actividad de diferentes péptidos, que a priori no tienen actividad anti-angiogénica para descubrir nuevos péptidos anti-tumorales.

De forma general a la metodología utilizada, un futuro desarrollo es el uso de diferentes tipos de técnicas para la selección de características. Se conoce que las variables ómicas presentan mucha correlación entre sí. Por lo tanto, sería interesante aplicar técnicas que impidiesen la selección de variables correlacionadas. El uso de técnicas de filtrado multivariado son idóneas para realizar dicha tarea. En este contexto, se está desarrollando un proyecto con el fin de extraer características con alta correlación con la variable dependiente pero baja correlación entre ellas. El objetivo es generar una firma de expresión genética que sea capaz de estratificar a los pacientes según su probabilidad de supervivencia y posteriormente asociar el valor predictivo de la firma con el efecto de fármacos ya aprobados. Estos resultados ofrecerán información para identificar pacientes que presenten unas características moleculares óptimas para un determinado tratamiento.

# Referencias

[1] Jose Liñares-Blanco, Alejandro Pazos y Carlos Fernandez-Lozano. «Machine learning analysis of TCGA cancer data». En: *PeerJ Computer Science* 7 (2021), e584 (vid. págs. XIX, 12-14, 94, 129-131, 133, 135, 137, 139).

[2] Jose Liñares-Blanco, Cristian R Munteanu, Alejandro Pazos y Carlos Fernandez-Lozano. «Molecular docking and machine learning analysis of Abemaciclib in colon cancer». En: *BMC Molecular and Cell Biology* 21.1 (2020), págs. 1-18 (vid. págs. XX, 12, 14, 39, 41, 130, 131, 133, 136, 138, 140).

[3] Jose Liñares Blanco, Ana B Porto-Pazos, Alejandro Pazos y Carlos Fernandez-Lozano. «Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection». En: *Scientific reports* 8.1 (2018), págs. 1-11 (vid. págs. XX, 14, 36, 114, 130, 131, 133, 135, 137, 140).

[4] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal y Freddie Bray. «Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries». En: *CA: a cancer journal for clinicians* 71.3 (2021), págs. 209-249 (vid. pág. 1).

[5] Michael R Stratton. «Exploring the genomes of cancer cells: progress and promise». En: *science* 331.6024 (2011), págs. 1553-1558 (vid. pág. 1).

[6] Douglas Hanahan y Robert A Weinberg. «The hallmarks of cancer». En: *cell* 100.1 (2000), págs. 57-70 (vid. pág. 2).

[7] Douglas Hanahan y Robert A Weinberg. «Hallmarks of cancer: the next generation». En: *cell* 144.5 (2011), págs. 646-674 (vid. págs. 2, 3).

[8] Michael R Stratton, Peter J Campbell y P Andrew Futreal. «The cancer genome». En: *Nature* 458.7239 (2009), págs. 719-724 (vid. págs. 2, 3, 5).

[9] Sushant Kumar, Jonathan Warrell, Shantao Li, Patrick D McGillivray, William Meyerson, Leonidas Salichos, Arif Harmanci, Alexander Martinez-Fundichely, Calvin WY Chan, Morten Muhlig Nielsen y col. «Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences». En: *Cell* 180.5 (2020), págs. 915-927 (vid. pág. 3).

[10] Marina Salvadores, David Mas-Ponte y Fran Supek. «Passenger mutations accurately classify human tumors». En: *PLoS computational biology* 15.4 (2019), e1006953 (vid. pág. 3).

[11] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon y col. «Comprehensive characterization of cancer driver genes and mutations». En: *Cell* 173.2 (2018), págs. 371-385 (vid. pág. 3).

[12] Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos y Nuria Lopez-Bigas. «IntOGen-mutations identifies cancer drivers across tumor types». En: *Nature methods* 10.11 (2013), págs. 1081-1082 (vid. pág. 3).

[13] David Tamborero, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S Lawrence, Gad Getz, Gary D Bader, Li Ding y col. «Comprehensive identification of mutational cancer driver genes across 12 tumor types». En: *Scientific reports* 3.1 (2013), págs. 1-10 (vid. pág. 3).

[14] Francis Crick. «Central dogma of molecular biology». En: *Nature* 227.5258 (1970), págs. 561-563 (vid. pág. 4).

[15] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass y Dokyoon Kim. «Methods of integrating data to uncover genotype–phenotype interactions». En: *Nature Reviews Genetics* 16.2 (2015), págs. 85-97 (vid. págs. 5, 6, 140).

[16] Mark A Dawson y Tony Kouzarides. «Cancer epigenetics: from mechanism to therapy». En: *cell* 150.1 (2012), págs. 12-27 (vid. pág. 5).

[17] Claudia Calabrese, Natalie R Davidson, Deniz Demircioğlu, Nuno A Fonseca, Yao He, André Kahles, Kjong-Van Lehmann, Fenglin Liu, Yuichi Shiraishi, Cameron M Soulette y col. «Genomic basis for RNA alterations in cancer». En: *Nature* 578.7793 (2020), págs. 129-136 (vid. pág. 5).

[18] Samir M Hanash, Sharon J Pitteri y Vitor M Faca. «Mining the plasma proteome for cancer biomarkers». En: *Nature* 452.7187 (2008), págs. 571-579 (vid. pág. 5).

[19] Karin Ortmayr, Sébastien Dubuis y Mattia Zampieri. «Metabolic profiling of cancer cells reveals genome-wide crosstalk between transcriptional regulators and metabolism». En: *Nature communications* 10.1 (2019), págs. 1-13 (vid. pág. 5).

[20] Gregory D Sepich-Poore, Laurence Zitvogel, Ravid Straussman, Jeff Hasty, Jennifer A Wargo y Rob Knight. «The microbiome and human cancer». En: *Science* 371.6536 (2021) (vid. pág. 5).

[21] Frederick Sanger, Steven Nicklen y Alan R Coulson. «DNA sequencing with chain-terminating inhibitors». En: *Proceedings of the national academy of sciences* 74.12 (1977), págs. 5463-5467 (vid. pág. 7).

[22] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh y col. «Initial sequencing and analysis of the human genome». En: (2001) (vid. pág. 7).

[23] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer y Barbara Wold. «Mapping and quantifying mammalian transcriptomes by RNA-Seq». En: *Nature methods* 5.7 (2008), págs. 621-628 (vid. pág. 9).

[24] Joshua Z Levin, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke y Aviv Regev. «Comprehensive comparative analysis of strand-specific RNA sequencing methods». En: *Nature methods* 7.9 (2010), págs. 709-715 (vid. pág. 9).

[25] *The Cancer Genome Atlas Program*. `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`. Accessed: 2021-05-27 (vid. págs. 11, 12).

[26] Cancer Genome Atlas (TCGA) Research Network y col. «Comprehensive genomic characterization defines human glioblastoma genes and core pathways». En: *Nature* 455.7216 (2008), pág. 1061 (vid. pág. 11).

[27] P. S. Hammerman, M. S. Lawrence, D. Voet y col. «Comprehensive genomic characterization of squamous cell lung cancers». En: *Nature* 489.7417 (2012), págs. 519-525 (vid. pág. 11).

[28] D. Bell, A. Berchuck, M. Birrer y col. «Integrated genomic analyses of ovarian carcinoma». En: *Nature* 474.7353 (2011), págs. 609-615 (vid. pág. 11).

[29] *NCI GDC Data Portal*. `https://portal.gdc.cancer.gov/`. Accessed: 2021-08-10 (vid. pág. 12).

[30] Amy Blum, Peggy Wang y Jean C Zenklusen. «SnapShot: TCGA-analyzed tumors.» En: *Cell* 173.2 (2018), págs. 530-530 (vid. pág. 12).

[31] Jerome Friedman, Trevor Hastie, Robert Tibshirani y col. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001 (vid. págs. 14, 15, 17).

[32] Anthony WF Edwards y Luigi Luca Cavalli-Sforza. «A method for cluster analysis». En: *Biometrics* (1965), págs. 362-375 (vid. pág. 14).

[33] Cecil C Bridges Jr. «Hierarchical cluster analysis». En: *Psychological reports* 18.3 (1966), págs. 851-854 (vid. pág. 14).

[34] James Bergstra y Yoshua Bengio. «Random search for hyper-parameter optimization.» En: *Journal of machine learning research* 13.2 (2012) (vid. pág. 16).

[35] Raymond E Wright. «Logistic regression.» En: (1995) (vid. pág. 17).

[36] Arthur E Hoerl y Robert W Kennard. «Ridge regression: Biased estimation for nonorthogonal problems». En: *Technometrics* 12.1 (1970), págs. 55-67 (vid. pág. 18).

[37] Robert Tibshirani. «Regression shrinkage and selection via the lasso». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), págs. 267-288 (vid. págs. 18, 35).

[38] Jerome Friedman, Trevor Hastie y Rob Tibshirani. «Regularization paths for generalized linear models via coordinate descent». En: *Journal of statistical software* 33.1 (2010), pág. 1 (vid. pág. 18).

[39] Naomi S Altman. «An introduction to kernel and nearest-neighbor nonparametric regression». En: *The American Statistician* 46.3 (1992), págs. 175-185 (vid. pág. 20).

[40] Leo Breiman. «Random forests». En: *Machine learning* 45.1 (2001), págs. 5-32 (vid. págs. 22, 23).

[41] Colin Campbell y Yiming Ying. «Learning with support vector machines». En: *Synthesis lectures on artificial intelligence and machine learning* 5.1 (2011), págs. 1-95 (vid. pág. 24).

[42] Nello Cristianini, John Shawe-Taylor y col. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000 (vid. pág. 24).

[43] Corinna Cortes y Vladimir Vapnik. «Support vector machine». En: *Machine learning* 20.3 (1995), págs. 273-297 (vid. pág. 24).

[44] Bernhard Scholkopf, Sebastian Mika, Chris JC Burges, Philipp Knirsch, K-R Muller, Gunnar Ratsch y Alexander J Smola. «Input space versus feature space in kernel-based methods». En: *IEEE transactions on neural networks* 10.5 (1999), págs. 1000-1017 (vid. pág. 25).

[45] J Mercer. «Functions ofpositive and negativetypeand theircommection with the theory ofintegral equations». En: *Philos. Trinsdictions Rogyal Soc* 209 (1909), págs. 4-415 (vid. pág. 25).

[46] AJ Smola. «Learning with Kernels PhD Thesis». En: *Technische Universitat Berlin* (1998) (vid. pág. 25).

[47] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin y col. *A practical guide to support vector classification*. 2003 (vid. pág. 26).

[48] César Ferri, José Hernández-Orallo y R Modroiu. «An experimental comparison of performance measures for classification». En: *Pattern Recognition Letters* 30.1 (2009), págs. 27-38 (vid. págs. 26, 28).

[49] Tom Fawcett. «An introduction to ROC analysis». En: *Pattern recognition letters* 27.8 (2006), págs. 861-874 (vid. pág. 28).

[50] Ian Jolliffe. «Principal component analysis». En: *Encyclopedia of statistics in behavioral science* (2005) (vid. pág. 29).

[51] Dmitry Kobak y Philipp Berens. «The art of using t-SNE for single-cell transcriptomics». En: *Nature communications* 10.1 (2019), págs. 1-14 (vid. pág. 29).

[52] Fei Tony Liu, Kai Ming Ting y Zhi-Hua Zhou. «Isolation forest». En: *2008 eighth ieee international conference on data mining*. IEEE. 2008, págs. 413-422 (vid. pág. 29).

[53] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall y W Philip Kegelmeyer. «SMOTE: synthetic minority over-sampling technique». En: *Journal of artificial intelligence research* 16 (2002), págs. 321-357 (vid. pág. 29).

[54] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse y Amri Napolitano. «RUSBoost: A hybrid approach to alleviating class imbalance». En: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40.1 (2009), págs. 185-197 (vid. pág. 29).

[55] Terrance DeVries y Graham W Taylor. «Dataset augmentation in feature space». En: *arXiv preprint arXiv:1702.05538* (2017) (vid. pág. 29).

[56] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw y Daniel Rueckert. «Gan augmentation: Augmenting training data using generative adversarial networks». En: *arXiv preprint arXiv:1810.10863* (2018) (vid. pág. 29).

[57] Shuangtao Li, Yuanke Chen, Yanlin Peng y Lin Bai. «Learning more robust features with adversarial training». En: *arXiv preprint arXiv:1804.07757* (2018) (vid. pág. 29).

[58] Jianping Hu, Yijing Zhao, Mengcheng Li, Jianyi Liu, Feng Wang, Qiang Weng, Xingfu Wang y Dairong Cao. «Machine learning-based radiomics analysis in predicting the meningioma grade using multiparametric MRI». En: *European Journal of Radiology* 131 (2020), pág. 109251 (vid. pág. 29).

[59] Peipei Li, Yongjun Piao, Ho Sun Shon y Keun Ho Ryu. «Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data». En: *BMC bioinformatics* 16.1 (2015), págs. 1-9 (vid. pág. 30).

[60] Mark D Robinson y Alicia Oshlack. «A scaling normalization method for differential expression analysis of RNA-seq data». En: *Genome biology* 11.3 (2010), págs. 1-9 (vid. pág. 30).

[61] Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks y col. «Visualizing and interpreting cancer genomics data via the Xena platform». En: *Nature biotechnology* 38.6 (2020), págs. 675-678 (vid. pág. 30).

[62] Simon Anders y Wolfgang Huber. «Differential expression analysis for sequence count data». En: *Nature Precedings* (2010), págs. 1-1 (vid. pág. 30).

[63] Frederick Mosteller y John W Tukey. «Data analysis, including statistics». En: *Handbook of social psychology* 2 (1968), págs. 80-203 (vid. pág. 30).

[64] *Nested Cross-Validation. mlr package*. `https://mlr.mlr-org.com/articles/tutorial/nested_resampling.html`. Accessed: 2021-08-17 (vid. pág. 31).

[65] Salvador García, Alberto Fernández, Julián Luengo y Francisco Herrera. «Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power». En: *Information sciences* 180.10 (2010), págs. 2044-2064 (vid. pág. 31).

[66] Samuel Sanford Shapiro y Martin B Wilk. «An analysis of variance test for normality (complete samples)». En: *Biometrika* 52.3/4 (1965), págs. 591-611 (vid. pág. 32).

[67] Maurice Stevenson Bartlett. «Properties of sufficiency and statistical tests». En: *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* 160.901 (1937), págs. 268-282 (vid. pág. 32).

[68] Carlos Fernandez-Lozano, Marcos Gestal, Cristian R Munteanu, Julian Dorado y Alejandro Pazos. «A methodology for the design of experiments in computational intelligence with multiple regression models». En: *PeerJ* 4 (2016), e2721 (vid. págs. 32, 129).

[69] Yvan Saeys, Inaki Inza y Pedro Larranaga. «A review of feature selection techniques in bioinformatics». En: *bioinformatics* 23.19 (2007), págs. 2507-2517 (vid. pág. 33).

[70] Isabelle Guyon y André Elisseeff. «An introduction to variable and feature selection». En: *Journal of machine learning research* 3.Mar (2003), págs. 1157-1182 (vid. pág. 33).

[71] Isabelle Guyon, Jason Weston, Stephen Barnhill y Vladimir Vapnik. «Gene selection for cancer classification using support vector machines». En: *Machine learning* 46.1 (2002), págs. 389-422 (vid. pág. 33).

[72] Chris Ding y Hanchuan Peng. «Minimum redundancy feature selection from microarray gene expression data». En: *Journal of bioinformatics and computational biology* 3.02 (2005), págs. 185-205 (vid. pág. 33).

[73] Li-Yeh Chuang, Hsueh-Wei Chang, Chung-Jui Tu y Cheng-Hong Yang. «Improved binary PSO for feature selection using gene expression data». En: *Computational Biology and Chemistry* 32.1 (2008), págs. 29-38 (vid. pág. 33).

[74] Pat Langley y col. «Selection of relevant features in machine learning». En: *Proceedings of the AAAI Fall symposium on relevance*. Vol. 184. 1994, págs. 245-271 (vid. pág. 33).

[75] William H Kruskal y W Allen Wallis. «Use of ranks in one-criterion variance analysis». En: *Journal of the American statistical Association* 47.260 (1952), págs. 583-621 (vid. pág. 34).

[76] Milton Friedman. «The use of ranks to avoid the assumption of normality implicit in the analysis of variance». En: *Journal of the american statistical association* 32.200 (1937), págs. 675-701 (vid. pág. 34).

[77] Paruchuri R Krishnaiah. *A Hand Book of Statistics*. Vol. 1. Motilal Banarsidass Publishe, 1980 (vid. pág. 34).

[78] Mark Andrew Hall. «Correlation-based feature selection for machine learning». En: (1999) (vid. pág. 34).

[79] Daphne Koller y Mehran Sahami. *Toward optimal feature selection*. Inf. téc. Stanford InfoLab, 1996 (vid. pág. 34).

[80] Lei Yu y Huan Liu. «Efficient feature selection via analysis of relevance and redundancy». En: *The Journal of Machine Learning Research* 5 (2004), págs. 1205-1224 (vid. pág. 34).

[81] Josef Kittler y col. «Pattern recognition and signal processing». En: *Chapter Feature Set Search Algorithms Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands* (1978), págs. 41-60 (vid. pág. 35).

[82] JH Holand. *Adaptation in Natural and Artificial Systems*. Vol. 1. University of Michigan Press, 1975 (vid. pág. 35).

[83] Iñaki Inza, Pedro Larrañaga, Ramón Etxeberria y Basilio Sierra. «Feature subset selection by Bayesian network-based optimization». En: *Artificial intelligence* 123.1-2 (2000), págs. 157-184 (vid. pág. 35).

[84] Richard O Duda, Peter E Hart y col. *Pattern classification and scene analysis*. Vol. 3. Wiley New York, 1973 (vid. pág. 35).

[85] Ulf Norinder y Vesna Munic Kos. «QSAR models for predicting five levels of cellular accumulation of lysosomotropic macrocycles». En: *International journal of molecular sciences* 20.23 (2019), pág. 5938 (vid. pág. 35).

[86] Carlos Fernandez-Lozano, Marcos Gestal, Nieves Pedreira-Souto, Lucian Postelnicu, Julián Dorado y Cristian Robert Munteanu. «Kernel-based feature selection techniques for transport proteins based on star graph topological indices». En: *Current topics in medicinal chemistry* 13.14 (2013), págs. 1681-1691 (vid. pág. 35).

[87] Carlos Fernandez-Lozano, Marcos Gestal, Humberto González-Díaz, Julián Dorado, Alejandro Pazos y Cristian R Munteanu. «Markov mean properties for cell death-related protein classification». En: *Journal of theoretical biology* 349 (2014), págs. 12-21 (vid. pág. 35).

[88] Abid Qureshi, Himani Tandon y Manoj Kumar. «AVP-IC50Pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC50)». En: *Peptide Science* 104.6 (2015), págs. 753-763 (vid. pág. 35).

[89] Lei Chen, Chen Chu, Tao Huang, Xiangyin Kong y Yu-Dong Cai. «Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models». En: *Amino acids* 47.7 (2015), págs. 1485-1493 (vid. pág. 35).

[90] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu y col. «PubChem 2019 update: improved access to chemical data». En: *Nucleic acids research* 47.D1 (2019), págs. D1102-D1109 (vid. págs. 35, 39).

[91] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda y col. «DrugBank 5.0: a major update to the DrugBank database for 2018». En: *Nucleic acids research* 46.D1 (2018), págs. D1074-D1082 (vid. págs. 35, 39).

[92] Teague Sterling y John J Irwin. «ZINC 15–ligand discovery for everyone». En: *Journal of chemical information and modeling* 55.11 (2015), págs. 2324-2337 (vid. págs. 35, 39).

[93] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani y col. «ChEMBL: a large-scale bioactivity database for drug discovery». En: *Nucleic acids research* 40.D1 (2012), págs. D1100-D1107 (vid. págs. 35, 39).

[94] Joseph L Durant, Burton A Leland, Douglas R Henry y James G Nourse. «Reoptimization of MDL keys for use in drug discovery». En: *Journal of chemical information and computer sciences* 42.6 (2002), págs. 1273-1280 (vid. pág. 36).

[95] David Rogers y Mathew Hahn. «Extended-connectivity fingerprints». En: *Journal of chemical information and modeling* 50.5 (2010), págs. 742-754 (vid. pág. 36).

[96] Paula Carracedo-Reboredo, Jose Liñares-Blanco, Nereida Rodríguez-Fernández, Francisco Cedrón, Francisco J Novoa, Adrian Carballal, Victor Maojo, Alejandro Pazos y Carlos Fernandez-Lozano. «A review on machine learning approaches and trends in drug discovery». En: *Computational and Structural Biotechnology Journal* 19 (2021), pág. 4538 (vid. págs. 37, 40).

[97] Michael Hay, David W Thomas, John L Craighead, Celia Economides y Jesse Rosenthal. «Clinical development success rates for investigational drugs». En: *Nature biotechnology* 32.1 (2014), págs. 40-51 (vid. pág. 41).

[98] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg y Aaron L Schacht. «How to improve R&D productivity: the pharmaceutical industry's grand challenge». En: *Nature reviews Drug discovery* 9.3 (2010), págs. 203-214 (vid. pág. 41).

[99] Fabio Pammolli, Laura Magazzini y Massimo Riccaboni. «The productivity crisis in pharmaceutical R&D». En: *Nature reviews Drug discovery* 10.6 (2011), págs. 428-438 (vid. pág. 41).

[100] Oleg Trott y Arthur J Olson. «AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading». En: *Journal of computational chemistry* 31.2 (2010), págs. 455-461 (vid. pág. 41).

[101] Jerome Eberhardt, Diogo Santos-Martins, Andreas Tillack y Stefano Forli. «AutoDock Vina 1.2. 0: new docking methods, expanded force field, and Python bindings». En: (2021) (vid. pág. 41).

[102] *COAD-DRD: Colon Adenocarcinoma Drug Repurposing with Docking*. `https://muntisa.github.io/COAD-DRD/`. Accessed: 2021-09-21 (vid. pág. 41).

[103] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee y col. «Drug repurposing: progress, challenges and recommendations». En: *Nature reviews Drug discovery* 18.1 (2019), págs. 41-58 (vid. pág. 42).

[104] *NIH: Drugs approved for different types of cancer*. `https://www.cancer.gov/about-cancer/treatment/drugs/cancer-type`. Accessed: 2021-09-22 (vid. pág. 42).

[105] Zhe Zhang, Li Zhou, Na Xie, Edouard C Nice, Tao Zhang, Yongping Cui y Canhua Huang. «Overcoming cancer therapeutic bottleneck by drug repurposing». En: *Signal transduction and targeted therapy* 5.1 (2020), págs. 1-25 (vid. pág. 42).

[106] Audrey A Tran y Vinay Prasad. «Drug repurposing for cancer treatments: A well-intentioned, but misguided strategy». En: *The Lancet Oncology* 21.9 (2020), págs. 1134-1136 (vid. pág. 42).

[107] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann y col. «A deep learning approach to antibiotic discovery». En: *Cell* 180.4 (2020), págs. 688-702 (vid. pág. 128).

[108] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko y col. «Highly accurate protein structure prediction with AlphaFold». En: *Nature* 596.7873 (2021), págs. 583-589 (vid. pág. 128).

[109] *DeepMind*. `https://deepmind.com/`. Accessed: 2021-09-10 (vid. pág. 128).

[110] *14th Community Wide Experiment on the Critial Assessment of Techniques for Protein Structure Prediction - CASP14*. `https://predictioncenter.org/casp14/`. Accessed: 2021-09-10 (vid. pág. 128).

[111] Claude Sammut y Geoffrey I Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011 (vid. pág. 129).

[112] Peter Carmeliet y Rakesh K Jain. «Angiogenesis in cancer and other diseases». En: *nature* 407.6801 (2000), págs. 249-257 (vid. pág. 131).

[113] Krishnansu S Tewari, Michael W Sill, Harry J Long III, Richard T Penson, Helen Huang, Lois M Ramondetta, Lisa M Landrum, Ana Oaknin, Thomas J Reid, Mario M Leitao y col. «Improved survival with bevacizumab in advanced cervical cancer». En: *New England Journal of Medicine* 370.8 (2014), págs. 734-743 (vid. pág. 131).

[114] Robert A Burger, Mark F Brady, Michael A Bookman, Gini F Fleming, Bradley J Monk, Helen Huang, Robert S Mannel, Howard D Homesley, Jeffrey Fowler, Benjamin E Greer y col. «Incorporation of bevacizumab in the primary treatment of ovarian cancer». En: *New England Journal of Medicine* 365.26 (2011), págs. 2473-2483 (vid. pág. 131).

[115] David F McDermott, Mahrukh A Huseni, Michael B Atkins, Robert J Motzer, Brian I Rini, Bernard Escudier, Lawrence Fong, Richard W Joseph, Sumanta K Pal, James A Reeves y col. «Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma». En: *Nature medicine* 24.6 (2018), págs. 749-757 (vid. pág. 131).

[116] *NIH: The Cost of Sequencing a Human Genome*. `https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost`. Accessed: 2021-10-05 (vid. pág. 139).

# A Producción científica

En esta sección se recopila la producción científica generada durante el período de tesis. Se marca la diferencia entre los tres artículos que componen esta tesis y entre los artículos y/o participaciones en congresos generados a raíz del desarrollo de la tesis.

## A.1 Compendio de tres artículos JCR (WOS)

- **Liñares-Blanco, J.**, Pazos, A. and Fernandez-Lozano, C. *Machine learning analysis of TCGA cancer data*. PeerJ Computer Science 7 (2021). Q1, 3.091 IF. https://doi.org/10.7717/peerj-cs.584

- **Liñares-Blanco, J.**, Munteanu, C.R., Pazos, A. and Fernandez-Lozano, C. *Molecular docking and machine learning analysis of Abemaciclib in colon cancer.* BMC Mol and Cell Biol 21, 52 (2020). Q3, 3.066 IF. 3 citas (Google Scholar). https://doi.org/10.1186/s12860-020-00295-w

- **Liñares-Blanco, J.**, Porto-Pazos, A.B., Pazos, A. and 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). Q1, 4.011 IF. 21 citas (Google Scholar). https://doi.org/10.1038/s41598-018-33911-z

## A.2 Otros artículos JCR (WOS) (2)

- Carracedo-Reboredo, P.[1], **Liñares-Blanco, J.**[1], Rodríguez-Fernández, N., Cedrón, F., et al. (2021). *A review on machine learning approaches and trends in drug discovery*. Computational and Structural Biotechnology Journal, 19, 4538. Q1, 7.271 IF. https://doi.org/10.1016/j.csbj.2021.08.011

[1] Estos autores han contribuido lo mismo en el artículo.

- Fernández-Edreira, D., **Liñares-Blanco, J.**, and Fernandez-Lozano, C. (2021). *Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes*. Expert Systems with Applications, 185, 115648. Q1, D1, 6.954 IF. https://doi.org/10.1016/j.eswa.2021.115648

## A.3 Participación en congresos nacionales e internacionales (7)

- **Liñares-Blanco, J.**, Fernandez-Lozano, C., Seoane, JA., and Lopez-Campos, G. *Machine Learning Algorithms Reveals Country-Specific Metagenomic Taxa from American Gut Project Data*. 31st Medical Informatics Europe Conference. May 2021. https://doi.org/10.3233/shti210185

- **Liñares-Blanco, J.**, Fernandez-Lozano, C., Seoane, JA. *A gene expression signature of 37 genes improves survival prognosis in colon cancer patients*. AACR Annual Meeting 2021. April 2021. https://doi.org/10.1158/1538-7445.AM2021-763

- Fernández-Edreira, D., **Liñares-Blanco, J.**, Fernandez-Lozano, C. *Identification of Prevotella, Anaerotruncus and Eubacterium Genera by Machine Learning Analysis of Metagenomic Profiles for Stratification of Patients Affected by Type I Diabetes*. III XoveTIC congress. A Coruña, Spain. September 2020. https://doi.org/10.3390/proceedings2020054050

- **Liñares-Blanco, J.**, Fernandez-Lozano, C. *Prediction of peptide vascularization inhibitory activity in tumor tissue as a possible target for cancer treatment*. II XoveTIC congress. A Coruña, Spain. September 2019. **Best paper award**. https://doi.org/10.3390/proceedings2019021015

- **Liñares-Blanco, J.**, Fernandez-Lozano, C. *Gene Signatures Research Involved in Cancer Using Machine Learning*. II XoveTIC congress. A Coruña, Spain. September 2019. https://doi.org/10.3390/proceedings2019021019

- Fernandez-Lozano, C., **Liñares-Blanco, J.**, Gestal, M., Dorado, J., et al. *Integrative multi-omics data-driven for metastasis prediction in cancer*. Data´18.

Madrid, Spain. October 2018. **Best paper award** https://doi.org/10.1145/3279996.3280040

- **Liñares-Blanco, J.**. *Methodology for differential genetic expression analysis by machine learning*. MOL2NET 2018, International Conference on Multidisciplinary Sciences, 4th edition. September 2018. https://doi.org/10.3390/mol2net-04-05503

## A.4 Capítulos de libro (1)

- **Liñares-Blanco, J.**, Gestal, M., Dorado, J., et al. (2019). *Differential gene expression analysis of RNA-seq data using machine learning for Cancer research*. In Machine Learning Paradigms (pp. 27-65). Springer, Cham. https://doi.org/10.1007/978-3-030-15628-2_3