UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA QUÍMICA, BIOTECNOLOGÍA Y MATERIALES

IMPLEMENTACIÓN DE MODELOS DE CLASIFICACIÓN EN CÁNCER BASADOS EN DATOS MUTACIONALES Y CLÍNICOS

TESIS PARA OPTAR AL GRADO DE DOCTOR EN CIENCIAS DE LA INGENIERÍA
MENCIÓN INGENIERÍA QUÍMICA Y BIOTECNOLOGÍA

KAREN YASMINE ORÓSTICA TAPIA

PROFESORES GUÍAS:

ÁLVARO OLIVERA-NAPPA
GONZALO NAVARRO
JUAN A. ASENJO DE LEUZE

MIEMBROS DELA COMISIÓN:

ZIOMARA GERDTZEN HAKIM
RICARDO ARMISEN YAÑEZ
BÁRBARA POBLETE LABRA

Enero 2021
Santiago de Chile

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE: Doctor en Ciencias de la Ingeniería mención en Ingeniería QuÍmica y Biotecnología POR: Karen Yasmine Oróstica Tapia
FECHA: 18 de Enero 2021
PROF. GUÍAS: Dr. ÁLVARO OLIVERA-NAPPA, GONZALO NAVARRO Y JUAN ASENJO DE LEUZE

IMPLEMENTACIÓN DE MODELOS CLASIFICACIÓN EN CÁNCER BASADOS EN DATOS MUTACIONALES Y CLÍNICOS

El cáncer es la segunda causa principal de muerte en el mundo, generando 9.6 millones de decesos durante 2018. Ésta es una enfermedad compleja que combina tanto factores epigenéticos como genéticos. Actualmente, se han identificado varios factores de riesgo en cáncer; uno de los más importantes es el consumo de cigarrillo o tabaco. El consumo de este compuesto está asociado positivamente por lo menos a 13 tipos de cáncer, destacando en cáncer de hígado, riñón, vejiga y páncreas, entre otros. Sin embargo, pocos estudios han profundizado la relación entre generación de mutaciones somáticas y el consumo de cigarrillo, y cómo ésta afecta la propagación del cáncer a otros tejidos u órganos. En el primer capítulo de este trabajo se analizó la relación entre las variables clínicas, como el fumar cigarrillo, el subtipo tumoral, el subtipo histológico y la edad, con el número de mutaciones en pacientes con cáncer de pulmón. Encontramos que fumar cigarrillo aumenta de forma significativa el número de mutaciones con respecto a pacientes que nunca han fumado. Además, encontramos que los pacientes fumadores con Adenocarcinoma de Pulmón poseen una mayor cantidad de mutaciones, a diferencia de los pacientes con Carcinoma de Células Escamosas de Pulmón, en el cual no varía el número de mutaciones con los tipos de fumadores. Finalmente, usamos las variables clínicas junto con el número de mutaciones somáticas, que llamamos Carga Mutacional Total (*Total Mutational Load*) para clasificar el estado metastásico en los pacientes con cáncer de pulmón mediante modelos clasificatorios de Random Forest. Como principal hallazgo destacamos que la carga mutacional, el fumar cigarrillo y el grado tumoral contribuyen de una forma importante la clasificación de metástasis en pacientes con cáncer de pulmón.

Por otro lado, en el segundo capítulo de esta tesis, se aborda la clasificación del cáncer primario, dado que en cerca del 2% de la población con cáncer no se logra definir el órgano de origen del cáncer, lo que interfiere con la aplicación de terapias efectivas para estos pacientes. En este trabajo se analizó el poder predictivo de la acumulación de los distintos tipos de mutaciones somáticas, tales como SNVs, indels y SVs, para clasificar 33 tipos tumorales mediante modelos de clasificación *Random Forest* en 2653 pacientes provenientes del *Pan-Cancer Analysis of Whole Genomes (*PCAWG*) Consortium.* Como principal resultado obtuvimos que los patrones de SNVs permiten clasificar con una alta precisión la mayoría de los tipos tumorales evaluados. Esto indica que existen señales mutacionales específicas en los órganos que originan los tumores.

Para mi amado esposo
y mis increíbles padres

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less”.

Marie Curie

LISTA DE PUBLICACIONES GENERADAS A PARTIR DE ESTE TRABAJO

TOTAL MUTATIONAL LOAD AND CLINICAL FEATURES AS PREDICTORS OF THE METASTATIC STATUS IN LUNG ADENOCARCINOMA AND SQUAMOUS CELL CARCINOMA PATIENTS

Oróstica KY [1] , Juan Saez Hidalgo [1, 2], Pamela R. de Santiago [4], Solange Rivas[5], Gonzalo Navarro [1,2], Juan A. Asenjo [1] , Álvaro Olivera-Nappa [1,*] and Ricardo Armisén [3,*]
[1]Centre for Biotechnology and Bioengineering (CeBiB), Department of Chemical Engineering, Biotechnology and Materials, University of Chile, 8370456, Santiago, Chile, [2]Department of Computer Science, University of Chile, Santiago 8370459, Chile, [3] Centro de Genética y Genómica, Instituto de Ciencias e Innovación en Medicina, Facultad de Medicina Clínica Alemana Universidad del Desarrollo, 7590943, Santiago, Chile and [4]Department of Cell and Molecular Biology, Pontificia Universidad Católica de Chile, [5]Department of Basic Clinical Oncology, Faculty of Medicine, University of Chile. En preparación.

CLASSIFICATION OF PRIMARY CANCER BASED ON MUTATION PATTERNS USING RANDOM FOREST METHOD

Oróstica KY [1], Gonzalo Navarro[1,2], Juan A. Asenjo[1] , Álvaro Olivera-Nappa[1] Ricardo Armisén [3] and Esa Pitkänen[4,*]
[1]Centre for Biotechnology and Bioengineering (CeBiB), Department of Chemical Engineering, Biotechnology and Materials, University of Chile, 8370456, Santiago, Chile, [2]Department of Computer Science, University of Chile, Santiago 8370459, Chile, [3] Centro de Genética y Genómica, Instituto de Ciencias e Innovación en Medicina, Facultad de Medicina Clínica Alemana Universidad del Desarrollo, 7590943, Santiago, Chile, [4] Institute for Molecular Medicine Finland, Finland. En preparación.

AGRADECIMIENTOS

Durante mis estudios de postgrado he logrado reconocerme como una nueva persona, capaz de ver el mundo con otros ojos, con ojos de científica, feminista y amante de la naturaleza. Si bien este camino no ha sido fácil, donde hubo momentos en que pensé que no lo lograría, finalmente pude enfocar mis energías y llevar a cabo mis objetivos. Sin embargo esto no lo hubiera alcanzado sin el apoyo incondicional de muchas personas a mi alrededor. Mi más sentido agradecimiento a las siguientes personas por contribuir de una forma significativa en mi formación científica y personal.

En primer lugar, me gustaría expresar mi profunda gratitud a mis supervisores, Dr. Álvaro Olivera-Nappa, Dr. Gonzalo Navarro y el Dr. Juan A. Asenjo, por su constante asistencia y orientación, sobre todo por su confianza al apoyarme cuando no logré al comienzo optar a la Beca Conicyt. Además, gracias a ellos, en particular al Dr. Gonzalo, pude realizar dos pasantías a Finlandia donde pude conocer y acceder a nuevas ideas y proyectos que se tradujeron a un capítulo de mi tesis. Estas pasantías fueron apoyadas y financiadas por el proyecto BIOINFORMATICS AND INFORMATION RETRIEVAL DATA STRUCTURES ANALYSIS AND DESIGN, financiado por el programa de investigación e innovación Horizonte 2020 de la Unión Europea en el marco de Marie Sklodowska-Curie (acuerdo de subvención nº 690941).

También agradecer al Dr. Esa Pitkänen del Institute for Molecular Medicine Finland (FIMM) en Finlandia, quien me recibió en su laboratorio en Finlandia. Esa me dió excelentes consejos y orientación durante mi pasantía. Además, siempre estuvo dispuesto a ayudar y contribuyó mucho al desarrollo de mi tesis.

No quiero dejar de mencionar a todas las personas que trabajan en el CeBiB, a los profesores, secretarias, compañeros de trabajo y al personal de aseo, que sin ellos mi estadía en el CeBiB no hubiera sido igual, en especial a los “K”, los llevo en el corazón.

Quisiera agradecerle especialmente a mi marido, Ricardo A. Verdugo, por ser mi pilar, mi compañero; y brindarme su amor y apoyo de una forma incondicional, el cual fue esencial durante este tiempo, y sobre todo cuando estuvimos lejos por las pasantías. También, debo mencionar a mis padres quienes fomentaron mi curiosidad e interés por la ciencia, demostrando que no hay límites o barreras que me impidan seguir el camino que he elegido.

Finalmente, agradecer a la Agencia Nacional de Investigación y Desarrollo (ANID) por la Beca de doctorado Nacional (#21182123) y al CeBiB por el financiamiento de mi trabajo de investigación.

TABLA DE CONTENIDO

Listas de tablas

Listas de Figuras

INTRODUCCIÓN

El enfoque genómico para el diagnóstico y tratamiento dirigido del cáncer

El cáncer es una enfermedad compleja que surge de los efectos combinados de múltiples variaciones genéticas y epigenéticas (F. Zhang et al. 2016). Estos cambios pueden influir en la vulnerabilidad celular, provocando la transformación de una célula normal a una cancerosa (Dimitrakopoulos and Beerenwinkel 2017). Estudios recientes han determinado que los cánceres son entidades con un alto número de mutaciones, donde los genes afectados están implicados principalmente en la proliferación, regeneración y apoptosis celular, impulsando el crecimiento maligno de de los tumores (Kou et al. 2016).

Con el advenimiento de Secuenciación de Nueva Generación (*Next Generation Sequencing*, NGS) y el desarrollo de la genómica se ha revolucionado el entendimiento de la biología del cáncer, mejorando el diagnóstico y las terapias, dando paso a la era de la medicina de precisión y la genómica del cáncer (Garraway and Lander 2013). La medicina de precisión es definida como un enfoque clínico que busca seleccionar el método terapéutico más adecuado para cada paciente considerando diferentes parámetros clínicos y biomarcadores. En el último tiempo, se han realizando estudios sistemáticos del genoma humano para la identificación de alteraciones genéticas recurrentes en tipos específicos de cáncer, contribuyendo en el entendimiento del cáncer a nivel molecular. Estas alteraciones genéticas consisten principalmente en Variaciones de nucleótido único (*Single Nucleotide Variants*, SNVs), pequeñas inserciones y deleciones (Indels), fusión de genes, Variaciones en el número de copias (*Copy-Number Variations*, CNVs) y grandes re-arreglos cromosomales, también llamados Variaciones Estructurales (*Structural Variants*, SV) (Cheng, Zhao, and Zhao 2016).

Actualmente, la investigación se ha enfocado en la identificación de mutaciones *drivers*, que confieren una ventaja proliferativa selectiva a la célula con cáncer con respecto a células normales. Este tipo de mutación estarían causalmente implicadas en la oncogénesis, siendo los principales blancos para el desarrollo de nuevas terapias (Nik-Zainal 2014). Además, uno de los desafíos planteados es la discriminación de mutaciones *drivers* de mutaciones aleatorias o *passenger* que no desempeñan, hasta ahora, un papel significativo en el desarrollo del cáncer. Así mismo, éstas y otras mutaciones presentes en el tumor podrían estar asociadas con el diagnóstico y/o pronóstico del cáncer, confiriendoles un valor predictivo independientemente de la histología del tumor (Zutter et al. 2014). Mediante estos estudios genómicos se ha podido determinar que el cáncer no es sólo complejo, si no que también es altamente heterogéneo, debido a que los mecanismos genéticos pueden variar entre pacientes del mismo tipo patológico (W. Yang et al. 2014).

De esta forma, se ha podido identificar variantes genéticas predisponentes y subtipos tumorales basándose en la firmas moleculares características de cada paciente (Riazalhosseini and Lathrop 2016).
Las técnicas de NGS han permitido a los investigadores identificar variaciones genómicas de los cánceres humanos con altas tasas de mortalidad e incidencia, como el cáncer de pulmón y mama, mediante diferentes técnicas de secuenciación como: genoma completo (*Whole-Genome Sequencing*, WGS), exoma completo (*Whole-Exome Sequencing*, WES) y subsets de genes de interés (*Targeted Panel Sequencing*). Estos avances han facilitado la identificación de rutas genéticas alteradas importantes y entregan una visión genómica amplia de los cánceres.

Tecnologías de secuenciación aplicadas en genómica del cáncer

La llegada de los nuevos enfoques de secuenciación ha dado lugar a un importante aumento en la velocidad y cobertura de secuenciación del genoma humano. Por ejemplo, WGS es una técnica de resecuenciación capaz de identificar mutaciones somáticas en todo el genoma, incluyendo regiones reguladoras como promotores. Por otro lado, WES sólo analiza el 2% del genoma que corresponde a regiones codificantes, es decir, todos los genes en el genoma. Por último, Targeted Panel Sequencing permite la identificación de variaciones genéticas más comunes en un subset de genes que tienen asociaciones conocidas o sospechosas con una enfermedad o fenotipo. Esta técnica permite evaluar paneles desde 20 a 500 genes, y está basada en la captura de amplicones, siendo efectiva en la detección de SNV e indels. Este enfoque ofrece la ventaja de una alta profundidad y cobertura del exón (mayor de 99%). La profundidad representa el número de veces en que una base ha sido secuenciada y alineada a un genoma de referencia, mientras que la cobertura indica el porcentaje de genes (exones) generados por al menos un read secuenciado (Horak, Fröhling, and Glimm 2016).
Si bien, se sabe que la secuenciación del genoma completo es la estrategia más robusta para la caracterización genómica de tumores, Targeted Panel Sequencing y WES representan una alternativa práctica para ensayos clínicos, debido a que el almacenamiento y análisis computacional es más manejable, a diferencia de WGS (Damodaran et al. n.d.).
En esencia, para analizar el genoma del cáncer mediante estas técnicas de resecuenciación es necesario contar con dos muestras por paciente. Una muestra de DNA tumoral, extraída del tumor, y una muestra de DNA de sangre, proveniente generalmente de linfocitos circulantes (muestra normal germinal). Estas dos muestras son sometidas a fragmentación independiente generando millones de reads o lecturas. Esta colección de reads de DNA debe ser alineada contra un genoma de referencia, para así, armar el genoma del paciente del cual fueron

extraídas las muestras. Una vez alineado el genoma se procede a identificar diferencias genómicas entre la muestra del tumor, la muestra normal y el genoma de referencia. Este proceso es conocido como Llamado de Variantes Somáticas (*Somatic Variant Call*). Por medio de este proceso es posible identificar mutaciones somáticas adquiridas en el tumor y las mutaciones provenientes de la línea germinal del paciente. Posteriormente, las mutaciones germinales son removidas, manteniendo exclusivamente las mutaciones somáticas presentes en el tumor del paciente, formando un listado de mutaciones para cáncer de los pacientes estudiados.

Hoy en día existen varias bases de datos que tienen como principal objetivo tener este tipo de información a disposición de la comunidad científica, lo más rápido y con bajas restricciones, para acelerar la investigación acerca de las causas y el desarrollo del cáncer.

Fuentes de información

Datos de NGS y desarrollo de herramientas de análisis

Los proyectos *The Cancer Genome Atlas* (TCGA) y International Cancer Genomics Consortium (ICGC) han podido recopilar datos genómicos, epigenéticos y proteómicos a partir de estudios genómicos que realizan experimentos de secuenciación a un gran número de individuos con diferentes tipos de cáncer. TCGA es un proyecto de investigación que tiene como objetivo facilitar la comprensión de la genética del cáncer usando las nuevas herramientas de secuenciación y análisis para la identificación de los *drivers* del cáncer a nivel genético. De esta forma, este proyecto ha liderado la caracterización y recopilación de más de 10.000 muestras derivadas de 33 tipos de cáncer, proporcionando una importante oportunidad para el estudio de la relevancia biológica de los recientes descubrimientos en el cáncer a nivel genómico (J.-S. Lee 2016).

En cuanto al proyecto ICGC, este cuenta con 81,782,588 mutaciones somáticas y 57.658 genes mutados provenientes de 19.305 muestras donadas de acuerdo con la última actualización en el 2016 (release 23) (*Welcome \textbar ICGC Data Portal* n.d.). Otro ejemplo, es el repositorio *Cancer Genomics Hub* (CGHub), que reúne información de otras fuentes como *National Cancer Institute* (NCI) de Estados Unidos, TCGA, *Cancer Cell Line Encyclopedia* (CCLE) y el proyecto *Therapeutically Applicable Research to Generate Effective Treatments* (TARGET). El propósito de CGHub es el intercambio de datos de cáncer, para el desarrollo de la medicina personalizada, facilitando el acceso a los datos. Hasta la fecha CGHub cuenta con 42 tipos de cánceres y controles (Y. Yang et al. 2015).

Existen otros importantes repositorios que incorporan parte de la información contenida en TCGA y ICGC, con la ventaja de implementar herramientas de visualización, lo que permite un mejor entendimiento de los datos, dado que proporciona aplicaciones interactivas para la consulta de asociaciones entre

fenotipos, información clínica y los genes identificados como *drivers*. Por ejemplo, se encuentra cBioportal for Cancer Genomics, el cual es un portal interactivo para la exploración y visualización de datos genómicos de cáncer. Estos datos son obtenidos de TCGA y CCLE. En particular, este portal permite explorar, por medio de su interfaz gráfica, alteraciones genómicas en muestras que corresponden a un estudio, o comparar frecuencias de mutaciones en múltiples proyectos (Gao et al. 2013).
Otro ejemplo es The *UCSC Cancer Genomics Browser*, el cual es una herramienta web implementada para la visualización y análisis de datos genómicos y clínicos del cáncer. Este browser alberga 71.870 proyectos genómicos provenientes de TCGA, CCLE, entre otros. Además, esta herramienta permite acceder a más de un proyecto o conjunto de datos, para así, poder compararlos a nivel de su expresión génica y CNVs en distintos tipos de cáncer (Y. Yang et al. 2015).

Catálogos de mutaciones asociadas a cáncer

Actualmente, existen grandes proyectos internacionales que tienen como objetivo crear un catálogo de todos los genes y mutaciones responsables de la iniciación y progresión del cáncer. Un ejemplo es Catalogue of Somatic Mutations in Cancer (COSMIC) (Horak, Fröhling, and Glimm 2016). COSMIC es una de las bases de datos más completas y abarca 572 genes en los cuales se han encontrado mutaciones causalmente implicadas en el cáncer. Además, hasta la fecha se han recopilado 15.047 genomas obtenidos por WGS, 2.710.499 mutaciones codificantes, 10.567 fusiones de genes, 61.232 rearreglos genómicos y 702.652 CNVs (Y. Yang et al. 2015). Otra base de datos es The Atlas of Genetics and Cytogenetics in Oncology and Haematogy (AGCOH), que almacena datos de anomalías cromosómica y genes implicados en el cáncer. Esta base de datos contiene 1452 genes, y un set de genes que no han sido asociados con cáncer antes, pudiendo ser utilizados como un set de control negativo en futuros trabajos (Dimitrakopoulos and Beerenwinkel 2017; Huret, Dessen, and Bernheim 2001).

Enfoques computacionales para la identificación de genes y pathway asociados al cáncer

La caracterización genómica del cáncer entrega una visión amplia de cómo las variaciones genómicas pueden definir tipos de cánceres, expandiendo la compresión de las vías moleculares disfuncionales que conducen el proceso oncogénico. Por tales motivos numerosos estudios se han enfocado en la identificación de mutaciones y genes *drivers* que, además, apoyados por la masiva generación de datos provenientes de NGS, han impulsado el desarrollo de nuevas estrategias computacionales para la identificación de los drivers del cáncer. En la literatura se han planteado enfoques computacionales para la identificación de mutaciones, genes, y rutas biológicas asociados al cáncer. Estos enfoques corresponden a Métodos basados en rutas conocidas y en redes, en Métodos *de novo* de aprendizaje de rutas biológicas y en la frecuencia mutacional.

Métodos basados en pathways conocidas

Para interpretar mutaciones somáticas se han desarrollado herramientas que se basan en la comparación de estas mutaciones y la información pública de las rutas biológicas. Básicamente, se evalúa la probabilidad de solapamiento entre un set de genes mutados de interés y un set de genes con anotación funcional conocida, utilizando un Test Exacto de Fischer. Sí la probabilidad de observar solapamiento entre las listas de genes es lo suficientemente pequeña, bajo la hipótesis nula, el set es considerado enriquecido para la función respectiva. La anotación funcional se encuentra disponible en bases de datos como KEGG, *Gene Ontology* (GO) y Reactome, entre otras. Uno de los programas que se destaca dentro de este enfoque, es PathScan, el cual es capaz de detectar rutas enriquecidas a través de todos los pacientes evaluados. A pesar que estos enfoques son utilizados para interpretar listas de genes mutados, estos no pueden ser utilizados para predecir nuevas rutas, ya que están basados en rutas reportadas en estudios anteriores. Además, estos enfoques no consideran la dependencia entre genes de la misma ruta biológica (Wendl et al. 2011).

Métodos basados en redes

En cuanto a la estructura de una red biológica, variaciones genéticas pueden provocar cambios estructurales en la red, desencadenando cambios en las propiedades bioquímicas del sistema.

Estas redes han sido usadas ampliamente en estudios de cáncer para detectar patrones entre diferentes tipos de esta enfermedad. La mayoría de los métodos basados en redes integran mutaciones somáticas y redes de interacción con el objetivo de detectar grupos o módulos de interacción de genes mutados. La idea detrás de estos métodos, es que las mutaciones que sufren los pacientes, del mismo cáncer, afectan diferentes genes, pero estos participan en el mismo proceso

biológico, siendo representados por subgrafos densamente conectados en la red de interacción. En el trabajo de (Hofree et al. 2013) se propone un método, llamado *Network-based Stratification* (NBS), el cual agrupa pacientes con mutaciones en sectores parecidos en la red usando una técnica de *data mining* llamada *Clustering.*

Los métodos basados en red han sido efectivos en la interpretación de las consecuencias biológicas de aberraciones genómicas en el cáncer. Sin embargo, estos métodos sufren serias limitaciones. Primero, las redes de Interacción Proteína-Proteína, obtenidas de experimentos de alto rendimiento, cubren sólo entre el 20 y 30% de todas las potenciales interacciones en la célula humana, es decir, el actual interactoma está aproximadamente un 80% incompleto. Segundo, las redes han sido construidas con información obtenida de experimentos de gran escala y/o por medio de algoritmos de predicción para un tejido o condición en particular.

Métodos *de novo* de aprendizaje de rutas biológicas

Estos métodos también son capaz de detectar genes y rutas biológicas asociados a cánceres, mediante la identificación de patrones combinados de mutaciones. Existen dos tipos de enfoques de patrones combinados. El primero, llamado Mutación Mutuamente Excluyente (*Mutually Exclusive Mutation*), se basa en la probabilidad de ocurrencia de dos mutaciones, donde existe una correlación negativa entre estas dos mutaciones presentes en genes de la misma ruta. Entonces, la primera mutación confiere la ventaja selectiva que aumenta la proliferación, lo cual promueve la expansión clonal. Esto reduce la probabilidad que ocurra una segunda mutación en otro gen del mismo proceso biológico, debido a que ya existe la ventaja selectiva otorgada por la primera mutación. El segundo, llamado Mutación Co-ocurrente (*Co-ocurring mutation*), se caracteriza por la presencia simultánea de mutaciones en dos o más genes para que se genere la ventaja proliferativa, es decir, detecta genes mutados correlacionados positivamente. Un ejemplo es el programa MEMcover, el cual detecta patrones de alteraciones genéticas mutuamente excluyentes. Este método utiliza un test de permutaciones aleatorias para identificar patrones mutuamente excluyentes que se encuentran en un tipo específico de tejido o en varios (Kim et al. 2015).

Métodos basados en la frecuencia mutacional

Este enfoque computacional ha sido ampliamente usado en la determinación de genes *drivers* y se basa en la identificación de genes que poseen una frecuencia mutacional significativamente más alta que el escenario determinado por el tipo de cáncer. Por ejemplo en el trabajo de (Malouf et al. 2016) se identificaron alteraciones genómicas en 26 pacientes con Carcinoma de Células Renales con Sarcomatoide (sRCC, *Renal Cell Carcinoma with Sarcomatoid*). En este trabajo se logró estimar y comparar la frecuencia mutacional de los genes involucrados en este tipo de cáncer mediante el uso de un Test Exacto de Fisher.

Avances y desafíos en genómica del cáncer

Se han realizado varios estudios que buscan caracterizar el escenario genómico para los principales cánceres, como por ejemplo cáncer de pulmón, mama, próstata, estómago entre otros, los cuales tienen una alta incidencia y mortalidad. Esta caracterización genómica consiste principalmente en la identificación de genes conductores (*drivers*) y rutas biológicas (pathways), los cuales poseen un rol clave en la generación y desarrollo del cáncer.

En el trabajo de (Li et al. 2016) se identificó los principales genes *drivers* en Adenocarcinoma de Pulmón (LUAD, Lung Adenocarcinoma) en población asiática. Otro estudio en LUAD propuso caracterizar el escenario genómico de esta enfermedad mediante la búsqueda de señales de mutación asociados con la progresión del tumor, logrando identificar genes *drivers* por medio de la integración de datos epigenómicos, genómicos, la evolución clonal e información de las características clínicas. En otro reciente estudio, se describió la prevalencia y ubicación de mutaciones somáticas y re-ordenamientos cromosómicos en muestras de tumor en pacientes con cáncer de próstata. Además, se identificó mutaciones propias de pacientes afroamericanos, reportando una nueva fusión de genes presente en el 17% de los pacientes del estudio (Lindquist et al. 2016).

En la investigación de (Heo et al. 2017) se identificaron mutaciones somáticas usando WES en pacientes coreanos con Leucemia Mieloide Aguda (LMA, Acute Myeloid Leukemia), concluyendo que los subtipos morfológicos de la LMA pueden ser reflejados por patrones específicos de alteraciones genómicas. Además, exponen la necesidad de estudios de este tipo, los cuales son muy útiles para el diagnóstico temprano de estas enfermedades complejas.

En el trabajo de (Pereira et al. 2016), no sólo identificaron 40 genes con mutaciones drivers en 2.433 muestras de tumor en cáncer primario de mama, sino que buscaron patrones de asociación entre eventos mutacionales somáticos evaluando el nivel de relación con la sobrevida y patrones clínicos.

En otro estudio se realizó un análisis a gran escala de los patrones mutacionales presentes en el cáncer metastásico de mama, logrando descubrir alteraciones genómicas y señales mutacionales implicadas en la resistencia a terapias actuales (Lefebvre et al. 2016).

En el trabajo de (Choi et al. 2017) se secuenció el exoma completo de pacientes con Carcinoma de Células Escamosas de Pulmón (LUSC, Lung Squamous Cell Carcinoma) considerando en su análisis la anotación clínica de los pacientes. Además, pudieron encontrar genes mutados que están asociados a la respuesta clínica, pudiendo ampliar el entendimiento biológico de esta enfermedad y aplicar terapias personalizadas en pacientes con estadios temprano de LUSC.

Si bien la identificación de genes drivers es un importante avance para lograr comprender la biología del cáncer, identificar y comprender los factores que

contribuyen de forma significativa en la generación y desarrollo del cáncer, es el siguiente desafío en el área de investigación biomédica y de medicina de precisión.

Nuevos enfoques de estudio

El espectro de mutaciones que se encuentra en el genoma del cáncer puede ser explicado por procesos mutacionales específicos causados por exposiciones mutagénicas, errores en la reparación del DNA, modificación enzimática, entre otros. Estos procesos han sido denominados Mutational Signatures (Nik-Zainal et al. 2012). Este tipo de estudio se ha destacado y ganado notoriedad debido a que la presencia de un proceso mutacional podría señalar una mutación driver del cáncer, siendo un importante paso hacia el entendimiento de los mecanismos causantes del cáncer (Fischer et al. 2013).

Por otro lado, también se han estudiado ampliamente alteraciones genómicas específicas que están en co-ocurrencia y exclusión mutua. Estos estudios han sido basados en la idea de que mutaciones presentes en genes de diferentes pathways pueden ocurrir en el mismo cáncer, mientras que es muy raro encontrar mutaciones en genes que participan en la misma pathways. En el trabajo de (W. Zhang et al. 2017) se reportaron patrones mutacionales que no siguen esta idea, como por ejemplo los genes TP53 y CDKN2A, que participan en la pathway CDK, están en coocurrencia y no en exclusión mutua en LUAD y Head and Neck Squamous Cell Carcinoma (HNSC). Sin embargo, el trabajo de (W. Zhang et al. 2017) se estudió los patrones de co-ocurrencia y exclusión mutua para un solo gen, TP53, con un listado de genes drivers. Por lo tanto, realizar estudios integrativos de datos genómicos y clínicos podría ayudar en la identificación de patrones significativos de mutaciones somáticas, pudiendo asignarles una interpretación o sentido biológico relacionado con la respuesta a tratamientos tales como Sobrevida general (*Overall Survival, OS)*, Tiempo libre de la enfermedad (*Disease-free survival,* DFS), entre otros.

Otra propuesta que ha tomado fuerza en el último tiempo, tiene que ver con el rol de las mutaciones passenger y su relación con la resistencia del cáncer al tratamiento aplicado. En el trabajo de (Birkbak et al. 2013) se propone que la carga mutacional somática, también conocida como Total Mutational Burden (TMB), refleja los problemas de reparación del DNA y está asociado con la respuesta a los tratamientos aplicados en cáncer de ovario. Por lo tanto, estudiar la asociación entre el número total de mutaciones somáticas, con las características clínicas, puede ser útil en el desarrollo de nuevas estrategias terapéuticas.

Dado lo anterior, más el reciente aumento de datos genómicos multidimensionales y el rápido desarrollo de la ciencia de la computación, ha impulsado la búsqueda y extracción de señales ocultas dentro de los datos. En particular, en el área de la medicina personalizada, el uso de técnicas de data mining podrían apoyar el estudio de la respuesta clínica de enfermedades con alta tasa de incidencia y mortalidad, como el cáncer.

Integración de técnicas de minería de datos en la Medicina de precisión

Dada la complejidad de los datos genómicos provenientes de NGS, en términos de volumen y diversidad, surge la necesidad de contar con nuevas arquitecturas y herramientas computacionales para implementar análisis genómicos en la práctica clínica.

Recientemente, esta necesidad ha sido abordada por el concepto de Big data. Este concepto incluye diferentes aspectos como la colección, procesamiento y análisis de datos masivos provenientes de diferentes fuentes, con el objetivo de describir los datos, revelar patrones y correlaciones entre ellos, y predecir algunas respuestas clínicas de interés. Estos aspectos son llevados a cabo por distintas técnicas, entre ellas las de minería de datos o data mining. Estas técnicas buscan convertir la información bruta en activos de gran valor, es decir, se basa en la identificación de patrones ocultos en los datos (Bramer 2007; Alzahani et al. 2015).

Un enfoque que permite identificar patrones ocultos es *Machine Learning* (ML), el cual explora métodos analíticos de clasificación, clustering y regresiones lineales, con énfasis en la predicción de variables de interés. Estos algoritmos están clasificados en aprendizaje supervisado (*Supervised Learning*), aprendizaje no supervisado (*Unsupervised Learning*) y aprendizaje semi-supervisado (*Semi-supervised Learning*). El aprendizaje supervisado es un método de predicción de una salida (*output*) basándose en un conjunto de datos de entrenamiento, que corresponde a información ya clasificada o etiquetada. Un ejemplo, son las técnicas de clasificación, las cuales se encargan de construir reglas para asignar objetos a un set de clases pre-establecidas, también llamada variable predictiva, basándose en un vector de mediciones sobre estos objetos. Las técnicas de clasificación incluyen regresión logística (*Logistic regression*), métodos bayesianos (*Naive Bayesian methods*), árboles de decisión (decision trees), redes neuronales (*Neural networks*), redes bayesianas (*Bayesian network*) y *Support Vector Machine* (SVM), entre otros (Loyola-González et al. 2013).

En cuanto al aprendizaje no supervisado, este trata de encontrar patrones dentro de los datos sin información acerca de la salida que agrupa los datos, un ejemplo, es Clustering. Este enfoque se usa para encontrar grupos en los datos por medio de métricas de distancias. Existen diferentes técnicas de *Clustering* tales como *k-means* y *clustering* basado en componentes principales (*Principal*

Components-based Clustering). Estas técnicas han sido utilizadas ampliamente en el análisis de *microarray*, análisis filogenético y, recientemente, en el estudio de enfermedades complejas. El aprendizaje semi-supervisado intenta equilibrar el desempeño y precisión de los dos enfoques anteriores mediante el uso de pequeños set de datos anotados (etiquetados) y grandes set de datos no clasificados (Libbrecht and Noble 2015; Lee and Yoon 2017).

En recientes estudios se plantea la posibilidad de aplicar estas técnicas de *Machine Learning* para evaluar clasificar tipos tumorales, predecir la respuesta clínica en diferentes enfermedades, identificar patrones mutacionales asociados a fenotipos clínicos, entre otras cosas (Lynch et al. 2017; Vural, Wang, and Guda 2016). Donde el principal objetivo es poder determinar cuales son los factores que tienen más relevancia en la aparición y progresión del cáncer, contribuyendo de esta forma a la selección de la mejor estrategia terapéutica.

La memoria de esta tesis está organizada de la siguiente manera:

Capítulo 1: Carga mutacional total y características clínicas como predictores del estado metastásico en pacientes con adenocarcinoma de pulmón y carcinoma de células escamosas.

Los resultados del capítulo 1 responden a los objetivos 1 y 2. En este capítulo se exhibe la clasificación del estado metastásico en 1144 pacientes con Adenocarcinoma de pulmón y Carcinoma de células escamosas utilizando la carga mutacional y variables clínicas como predictores mediante el uso de modelos de Random Forest. Para lograr esto propusimos un benchmarking para seleccionar la mejor estrategia de clasificación del estado metastásico. En este trabajo encontramos que la carga mutacional está asociada a fumar cigarrillo, donde los pacientes que son fumadores o han fumado alguna vez presentan un elevado número de mutaciones a diferencia de los pacientes que nunca han fumado. Además, encontramos que las variables clínicas como el tipo de fumador, el subtipo de cáncer y la edad, junto con la carga mutacional, resultaron ser buenos predictores del estado metastásico con un F1 score de 0.64.

Capítulo 2: clasificación del cáncer primario basada en patrones de mutación utilizando modelos de Random Forest

Los resultados de este capítulo responden al tercer objetivo de esta tesis. En este trabajo se realizó un estudio de clasificación del tipo tumoral primario utilizando las mutaciones passenger como predictores en 33 tipos de cánceres provenientes del *Pan-Cancer Analysis of Whole Genomes* (PCAWG) *consortium*. Para esto, se identificaron patrones mutacionales basándose en el número de SNV, SV, indels y en el %GC de las secuencias. Luego, se evaluó el poder predictivo de cada uno de los patrones identificados, donde obtuvimos que para algunos cánceres los SNV correspondian a uno de los mejores predictores, logrando un F1 score mayor a 0.7. Por lo tanto, el uso de los SNVs permite clasificar el tipo tumoral con una alta precisión.

Hipótesis

1. La carga mutacional junto con el subtipo tumoral y estado de fumador son buenos predictores del estado metastásico en pacientes con cáncer de pulmón.
2. Los patrones en las mutaciones somáticas passenger son informativas para la clasificación de tumor de origen.

Objetivos

Objetivo General

Implementar nuevas estrategias computacionales para la identificación de patrones y relaciones entre los datos genómicos y/o clínicos en el estudio del cáncer.

Objetivos Específicos

1. Determinar la asociación entre carga mutacional tumoral y las características clínicas en pacientes con cáncer de pulmón.
2. Clasificar el estado metastásico de pacientes con cáncer de pulmón basándose en la relación de la carga mutacional total y las características clínicas.
3. Identificar patrones somáticos que permitan clasificar el tipo tumoral de origen en muestras provenientes del *Pan-Cancer Analysis of Whole Genomes* (PCAWG) *consortium*.

CAPÍTULO 1

TOTAL MUTATIONAL LOAD AND CLINICAL FEATURES AS PREDICTORS OF THE METASTATIC STATUS IN LUNG ADENOCARCINOMA AND SQUAMOUS CELL CARCINOMA PATIENTS

1.1 Abstract

Over the years, numerous studies have made public genomic and clinical data available from cancer patients.The large volume of accumulated data creates the opportunity of using Machine Learning methods to identify relevant factors in metastasis development, which is the greatest contributor to deaths from cancer. Therefore, being able to identify previously patients at high risk of developing metastasis using these relevant factors could contribute to the development of new therapeutic strategies. In this work, we propose the number of somatic mutations, defined as Total Mutational Load, and clinical features to classify metastasis in 972 Lung Adenocarcinoma and Lung Squamous Cell Carcinoma (LSCC) patients using the Random Forest method. We found that the Total Mutational Load, type of cancer and smoking history were good predictors of the spread of metastasis with an F1 score of 0.64.

1.2 Introduction

Lung cancer is the most common cause of cancer-related mortality worldwide for both men and women, being responsible for more than 1.4 million deaths a year [1]. The most frequent type of lung cancer is non-small-cell lung carcinoma (NSCLC), reaching up to 85% of all lung cancer cases (Siegel, Miller, and Jemal 2018). The main histologic subtypes of NSCLC are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LSCC), accounting for around 40% and 25-30% of all lung cancer, respectively (Zappa and Mousa 2016). LUAD is a highly heterogeneous disease with different genomic and histological patterns, presenting substantial differences in the mutational load of patients. In previous works (Shi et al. 2019; Sun, Schiller, and Gazdar 2007), the smoking condition has been recognized as the leading risk factor for lung cancer, especially for the LUAD subtype. Moreover, specific genes are affected depending on whether the patient is a smoker or not (Inamura et al. 2017). Recent advances in next-generation sequencing (NGS) have allowed the genomic characterization of this disease and the identification of frequently mutated genes such as TP53, EGFR, KRAS, ALK, BRAF, MET, RET, and ROS1 (Song et al. 2019). Despite these advances in genomic characterization, LUAD remains a challenge when it comes to associating genomic information with the clinical response, given its complexity and heterogeneity. Following LUAD, squamous cell carcinoma of the lung (LSCC) is the second most common histological subtype of NSCLC [2]. Differential gene expression profiles are found between LUAD and LSCC compared to normal lung tissue, and different cellular pathway are enriched in each case, emphasizing the divergent underlying mechanisms found between these subtypes (Lu et al. 2016). Regarding mutational load, LSCC patients showed a large number of DNA alterations, including mutations in critical genes as TP53, CDKN2A, PTEN, PIK3CA, KEAP1, MLL2, HLA-A, NFE2L2, NOTCH1, and RB1(TCGA network, 2012). Nevertheless, molecular therapies had not been effective for LSCC (Denisov et al. 2019).

Nowadays, researchers are using machine learning (ML) to find hidden insights inside their data in order to be able to resolve classification or prediction problems. In cancer research, identifying the best treatment strategy is the priority. For this purpose, researchers have started using ML models in their studies to classify or predict relevant clinical outcomes, such as overall survival (OS) and distant metastasis (DM), among others, since they deliver valuable information to better select therapeutic strategies (Adam et al. 2020). mRNA expression data, somatic mutational features (Yu et al. 2019), mutation of driver genes (Cho et al. 2018), and other features have been actively studied intending to obtain a prognostic value from this information, specifically for predicting overall survival as a variable of interest. However, unlike in survival studies, most of the factors determining clinical progression for distant metastasis are unknown or understudied (Wu et al. 2015), impacting clinical and survival prognosis. In the work of (Wang et al. 2019), the

authors used machine learning to identify lymph node metastasis from LUAD, based on DNA methylation signatures. Despite the advances in this area, more studies are needed to understand metastasis development and to improve diagnostic and prognostic accuracy in patients with Lung cancer, due to metastasis is the main cause of patient morbidity and mortality con cancer (Steeg 2016).

Therefore, finding which factors, whether clinical or genomic, contribute the most to the development of metastasis is crucial for predicting it in early stages of cancer. This information would allow the medical team to make the best decision when choosing a therapeutic strategy in patients with LUAD. However, studying this type of problem is quite complex, due to there are few data sets with information on the metastatic state of cancer patients, and they are variables with high medical interest. Besides, it is not only enough to have this type of information, but it must also be considered that there is a significant imbalance between the number of patients with and without metastasis, which represents a major challenge for precision medicine.

In the present study, we show a benchmarking to predict metastasis status in Pan-Lung Cancer samples using a Random Forest classifier model, where the predictors were the mutational load and relevant clinical variables. This approach allowed us to select the best processing, training, and validation strategy for the random forest models evaluated considering the characteristics of the studied data set. Therefore, this benchmarking will allow to improve the understanding of the relationships between genomic and clinical features for cancer prognosis.

1.3 Methods

1.3.1 Dataset and data preprocessing

Clinical and somatic mutational data from the Pan-Lung Cancer 2016 dataset was obtained from The Cancer Genome Atlas (TCGA) (Campbell et al. 2016). This dataset contains 1144 patients with Lung adenocarcinoma and Lung Squamous Cell Carcinoma (LSCC), which is an adequate number of samples that allow robust results applying machine learning techniques. The Pan-Lung Cancer 2016 dataset is one of the few data sets that contain information on the metastatic status of the patients, which is the variable of interest in this study. For Age, we classified patients in two groups: <= 60 years and >60 years. In Table 1, the characteristics of the entire cohort are depicted.

Table 1. Distribution of the clinical features in 1144 Pan-Lung Cancer samples.

Clinical features	Entire cohort	%
	N=1144	
Class metastasis		
M0	731	63.9%
M1	22	1.9%
MX	210	18.4%
Cancer type		
Lung Adenocarcinoma	660	57.7%
Lung Squamous Cell Carcinoma	484	42.3%
Tumour stage		
Stage I	8	0.7%
Stage IA	246	21.5%
Stage IB	321	(28.1%
Stage II	4	0.3%
Stage IIA	129	11.3%
Stage IIB	174	15.2%
Stage III	3	0.3%
Stage IIIA	155	13.5%
Stage IIIB	34	3.0%
Stage IV	8	3.3%
Age range		
<=60 years	61	29.5%
>60 years	146	70.5%
Gender		
Female	468	40.9%
Male	673	58.8%
N stage		
N0	624	54.5%

N1	221	19.3%
N2	113	9.9%
Nx	16	1.4%
Smoking status		
Current reformed smoker for < or = 15 years	407	35.6%
Current reformed smoker for > 15 years	212	18.5%
Current Reformed Smoker, Duration Not Specified	86	7.5%
Current Smoker	271	23.7%
Lifelong non-smoker	111	9.7%

1.3.2 Determination of Total Mutational load

From the mutational data, we implemented an *m* x *n* mutation count matrix only by considering missense mutations. These mutations are the most frequent in this dataset. This type of mutation could cause tumour suppressor proteins to be non-functional, or proto-oncogenes gain of function, thus granting a selective growth advantage to cancer cells (Zhao et al. 2018). In the matrix, *m* is the number of samples (1144 patients), and *n* is the number of genes (17305 genes). Therefore, the value in entry$V_{i,j}$ indicates the number of missense mutations of gene *j* in the patient *i*. To obtain the number of missense mutations we used the *mutCountMatrix()* function from the Maftools R package (Mayakonda et al. 2018). Due to the number of genes being much larger than the number of patients, we filtered out genes with a near-zero variance using the *VarianceThreshold()* function of sci-kit learn python package with a threshold of 0.05. Later, we computed the Total Mutational Load (TML), equivalent to the total number of missense mutations, for each patient considering only higher-variance genes (see Equation 1).

$$TML_i = \sum_{j}^{m} V_{i,j}$$

Equation 1. Total Mutational Load (TML) determination based on the number of missense mutations for each patient.

1.3.3 Statistical analysis

Association between TML and clinical features

We evaluated clinical data features corresponding to Sex, Tumour stage, Age range, M stage, Smoking pack by year and smoking history with the TML. For this, we modelled TML with a negative binomial regression (NBR) explanatory model since TML is an over-dispersed count variable. Then, we adjusted the model for the clinical features mentioned before. Next, we applied a backward stepwise model selection over the NBR to determine the effect of each clinical variable over the mutational load. For this, we used the *drop1()* function with a *likelihood-ratio test* (LRT). We next selected predictors using a statistical significance of 0.05. Besides, we apply the *t* student test to determine whether the means of mutations are equal for clinical variables.

1.3.4 Reclassification of patients using the regional lymph node feature

The metastatic status in our data indicates whether cancer has spread from the primary tumour to other parts of the body. The metastatic status is classified into three classes. Thus, the M0 class indicates that cancer has not spread to other parts of the body, M1 symbolises that cancer has spread out, and MX describes that metastasis could not be determined by the pathologist. We reclassify to the patients in the MX class using the regional lymph nodes (N) feature, which indicates whether cancer has spread to surrounding lymph nodes: N0 indicates a negative spread to lymph nodes while N1, N2 and N3 show that cancer has spread to lymph nodes. We used this approach based on the assumption the spread to the lymph nodes is considered as a first stage previous to the development of metastases. Therefore, it is unlikely that a patient in the N0 class could have any metastasis. Given this, we reclassified MX patients as M0 if they were labelled as N0, and M1 in the opposite case. We used M1 as a positive metastasis class andM0 as a negative metastasis class in our benchmarking.

1.3.5 Random Forest models

To identify the optimal predictors of the metastasis status (MS) in patients with lung adenocarcinoma and LSCC samples, we built four Random Forest (RF) classification models where the dependent variable corresponds to metastasis status described by the TNM staging system (https://www.cancer.gov/about-cancer/diagnosis-staging/staging). We do not consider the tumor stage in this analysis, since this variable is highly correlated with the metastatic state. This is because when pathologists assess metastatic status, they consider stage information to determine whether or not there is metastasis.
In the models, we used different combinations of TML and clinical features as follows:

1) model 1: MS ~ Clinical features + TML
2) model 2: MS ~ Clinical features
3) model 3: MS ~ TML

Each model was benchmarked as described below.

1.3.6 Benchmarking of classification process

The proposed benchmarking consists of 4 main steps. The first corresponds to the pre-processing, where we analyzed the clinical and mutational data (See Figure 4). Specifically, we converted every categorical variable to numerical in our clinical data using the One-Hot Encoding method [18]. The number of mutations for each gene and the TML was centred and scaled before building the models, and then we merged the clinical and mutational data into a single dataset. In the classification stage, we apply some strategies in the training process. Due to the high number of predictors in the 1 and 2 models, we decided to include Principal Component Analysis (PCA) to reduce the dimensionality and use the 100 first Principal Components (PCs) to train the models. Alternatively, we did not apply PCA to the data and we train the models without dimensionality reduction. Then, we used 5 Fold−Cross Validation to split the dataset into training a testing dataset, to then apply different sampling methods, such as over and under-sampling (Loyola-González et al. 2013). These resampling methods are designed to remove and add data from the training dataset, to modify the distribution of the unbalanced classes. In parallel, we applied the hold−out method that consists in randomly splitting only one time the dataset into training and test sets so that the training set comprised 70% of the full data set, while the test set was composed of the remaining data, to prevent

overfitting. We applied this process either to the dataset with PCA and without PCA. In the Meta-analysis stage, we evaluated the training models for each method already described with the validation set, being able to obtain performance measures such as F1-score, precision and recall. These measures were utilized to assess the predictive performances of the models. Finally, we compared the models according to performance measures and selected models with the lowest false positive rate and the highest true positive rate.

Also, with these performance measures, we can choose the best strategies, considering if it is necessary to apply dimensionality reduction or not, type of training method and re-sampling ways for this data set.

Accuracy = (TP + TN) / (TP + FP + TN + FN)
Recall = TP / (TP + FN)
Precision = TP / (TP + FP)
F1 = 2 * ((Precision * Recall) / (Precision + Recall)

Where:
TP: True positive rate
TN: True negative rate
FP: False positive rate
FN: False negative rate

For the best model, we will obtain the predictors that most contribute to the classification of the metastatic status.

1.4 Results

1.4.1 Relationship between Total Mutational Load and clinical features

To understand the relationship between TML and clinical variables in the Pan-Lung Cancer cohort, we fitted an NBR model using TML as the outcome variable and clinical features as predictors. We next applied backward stepwise model selection to the NBR model for identifying predictors with a strong association to TML. We found that TML is significantly associated with the number of cigarette packs smoked per year, age-range and Cancer type with a p-value <0.05 (see Table 2). Figure 1 shows the average number of mutations for each age range. We found that there is a significant difference in the number of mutations for the age range, where patients younger than 60 years have more mutations than those older than 60 years. When we compare the mean of mutations for the types of smokers, we find that there is a significant difference between them (see Figure 2). Besides, we see that current smokers are those with the highest number of mutations, unlike patients who have never smoked. Then, when making the same comparison by type of cancer, we found that only for current smokers there is a significant difference in the number of mutations. When we compare the mean of mutations for the types of smokers, we find that there is a significant difference between them. Besides, we see that current smokers are those with the highest number of mutations, unlike patients who have never smoked. Then, when making the same comparison by type of cancer, we found that only for current smokers there is a significant difference in the number of mutations. Interestingly, we found that for patients with LSCC the number of mutations does not vary concerning the smoking history, unlike Lung adenocarcinoma.

Table 2. Association between clinical data and TML

Clinical features	LRT	P-value
Sex	0.2670	0.6053
Tumour stage	5.5828	0.6938
Age range	5.6112	0.0178 *
M stage	5.6824	0.6827
Cigarette packs per year	3.8632	0.0493 *
Smoking history	2.4781	0.4792
Cancer type	9.4596	0.0021 **

*Note: * p < 0.1; ** p < 0.05; *** p < 0.01*

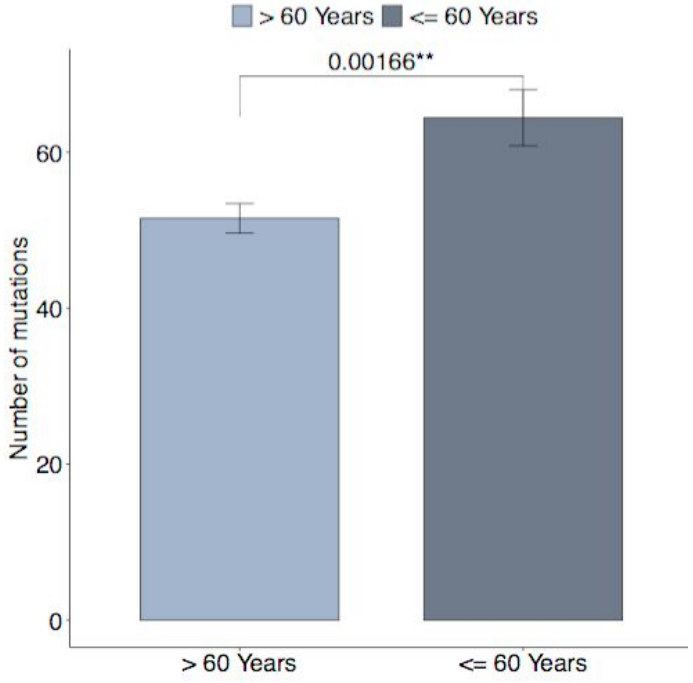


Figure 1. *Number of mutations by Age-range.*

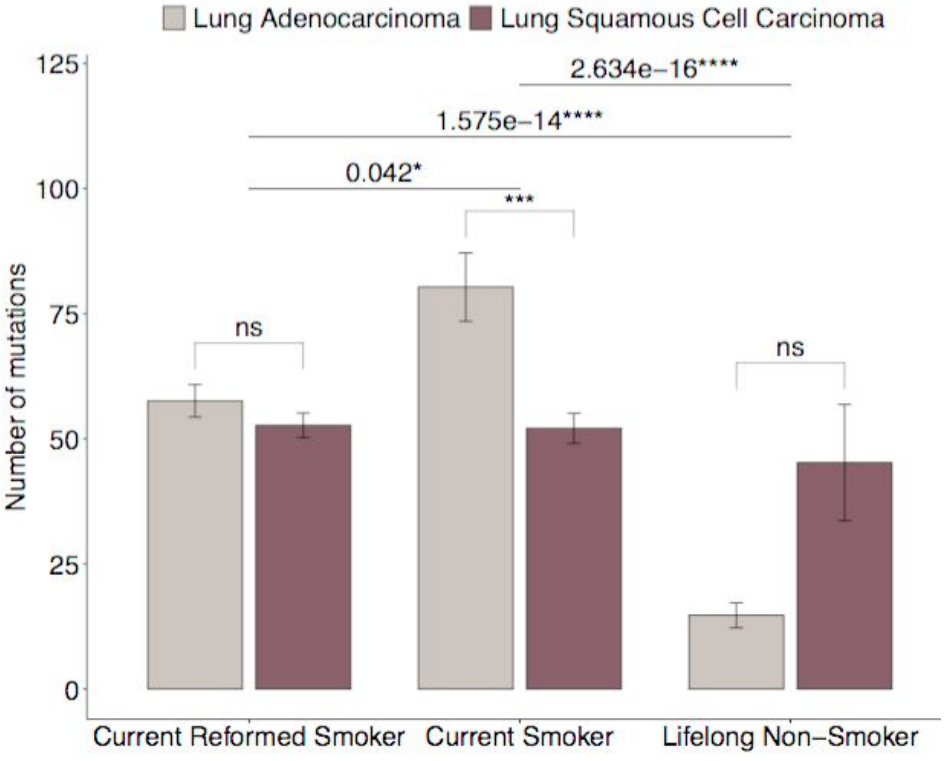


Figure 2. *Number of mutations considering the Cancer type and Smoking history.*

1.4.2 Reclassification of patients using cancer spread to lymph nodes

We reclassified our MX patients based on the N stage. With this procedure, our dataset comprised 881 M0 and 87 M1 patients. Figure 3 shows the results of our reclassification versus cancer stages. The I, IA and IB stages do not have patients classified as M1, unlike IIA, IIB, III, IIIA and IIIB stages which have a low proportion of M1 patients. Most patients in stage IV were labelled as M1, which may imply a specific correlation between these variables. Despite this classification that considers the spread of cancer to lymph nodes, there is still a clear imbalance of classes.

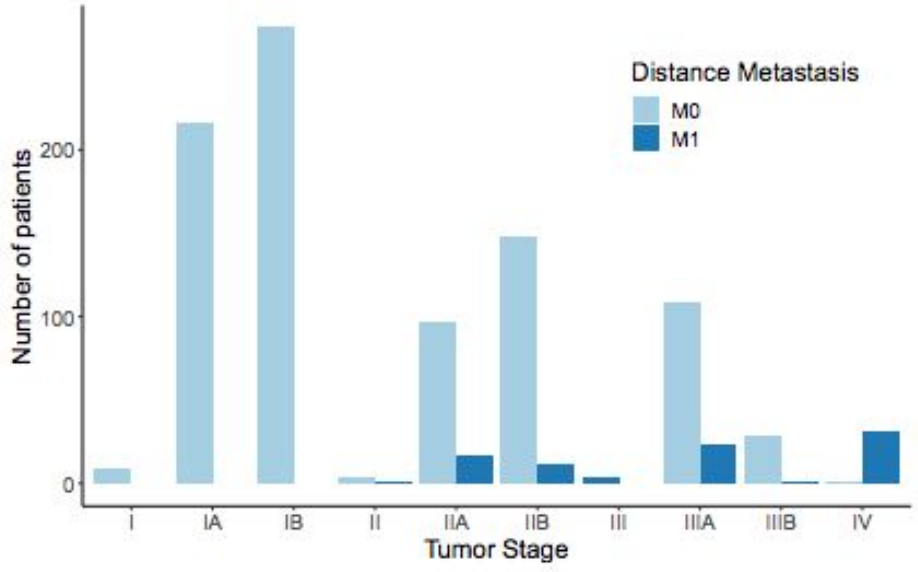


Figure 3. Distribution of new classification of distant metastasis by tumour stage.

1.4.3 Classification of patients using RF models

We built three RF classification models to classify metastatic status in Pan-Lung cancer patients. We tested different preprocessing, classification and validation strategies, such as dimensionality reduction, data separation methods, and sampling types. We compared the performance measures for class M1 (metastasis) since it is the least frequent class, and also its results vary according to the methods and models applied. Meanwhile, class M0 obtains performance measures above 0.9 for all models and methods. In Figure 5, we show the Precision and recall measures for all models considering the preprocessing and validation methods applied. To select the best model, we used the precision and recall metrics, selecting those with the highest values located in the upper right corner of Figure 5.
The models with the best metrics were model 1 that combines the clinical variables and TML without resampling. To select between the two models, we used the confusion matrix of each of them. The model 2 has higher recall than model 1, because model 2 classified 24 patients who had metastases as patients without metastasis (false negatives), while model 1 obtained 25 false negatives. However, model 1 only got 4 false positives, as opposed to model 2 that obtained just 12. Therefore, model 1 has better results in terms of false positives. When evaluating the use of PCA or not in these two models, we used the same approach based on the lowest number of false negatives and positives, obtaining that model 1 without PCA is the best model in terms of recall and precision.

According to the results, we found that the best model corresponds to the clinical variables together with the TML, where PCA was not applied, the 5 cross-validation method was used without resampling, obtaining an F1 value of 0.62. Although we obtained a slightly higher F1 value using Hold-out, we discarded it because the results with cross-validation are more robust and limit overfitting. However, the clinical features by themselves also express a good classifying model of the metastatic status, where the tumour stage is one major component to obtain these results. Also, we see that there is no relevant difference between applying PCA or not in terms of precision and recall. We see a kind of grouping of models with PCA and by type of re-sampling method.

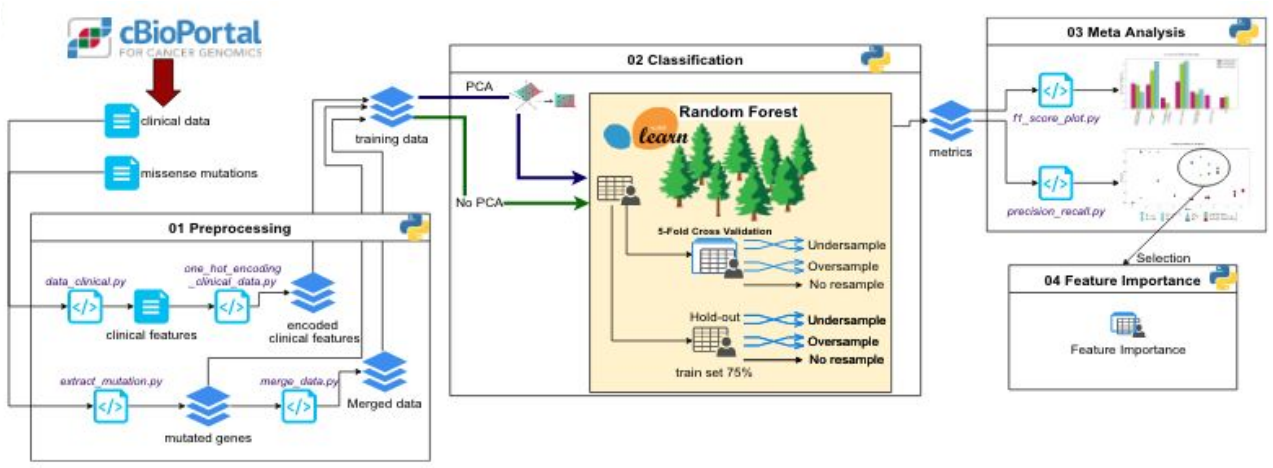
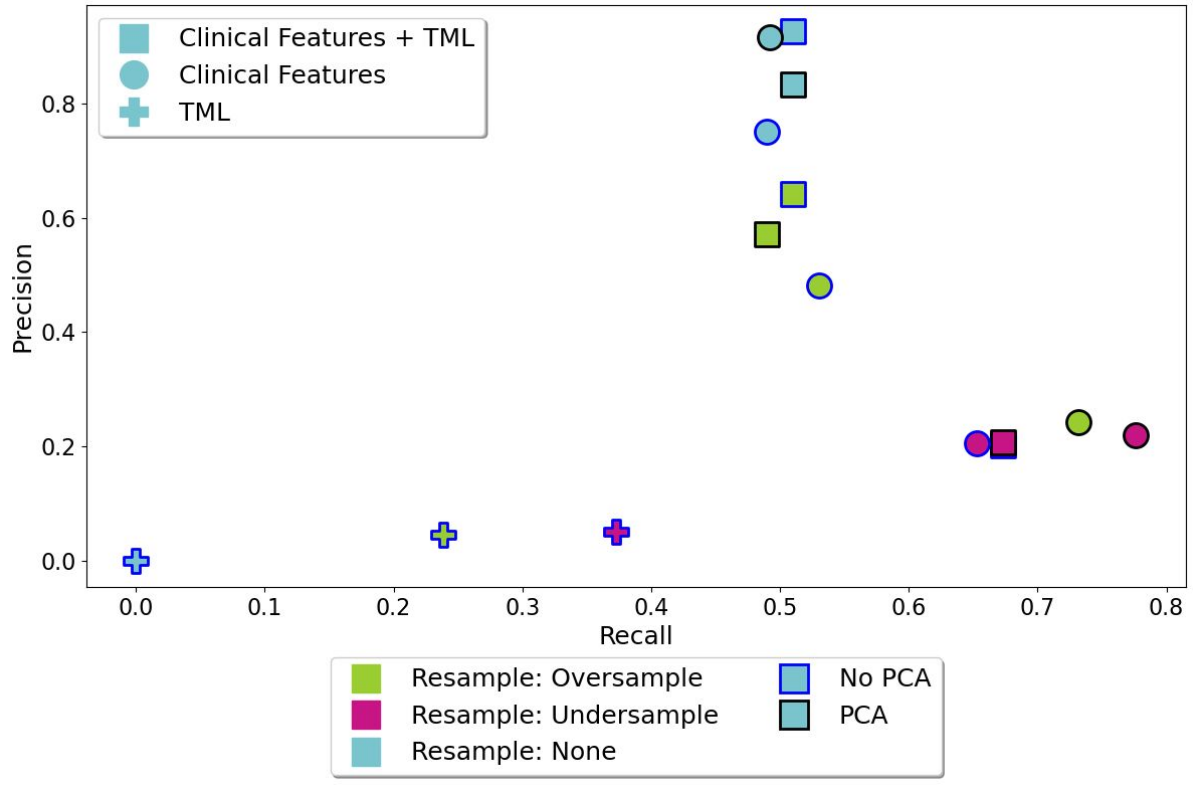


Figure 4. This benchmarking has 4 main stages, where we preprocessed the dataset, we applied different classification and validation strategies, such as dimensionality reduction, data separation methods, and sampling methods. In the last stage, we evaluated all RF models proposed from different strategies. Finally, we selected the best model for the dataset studied and also we got the variable that more contributed to the classification.

A



B

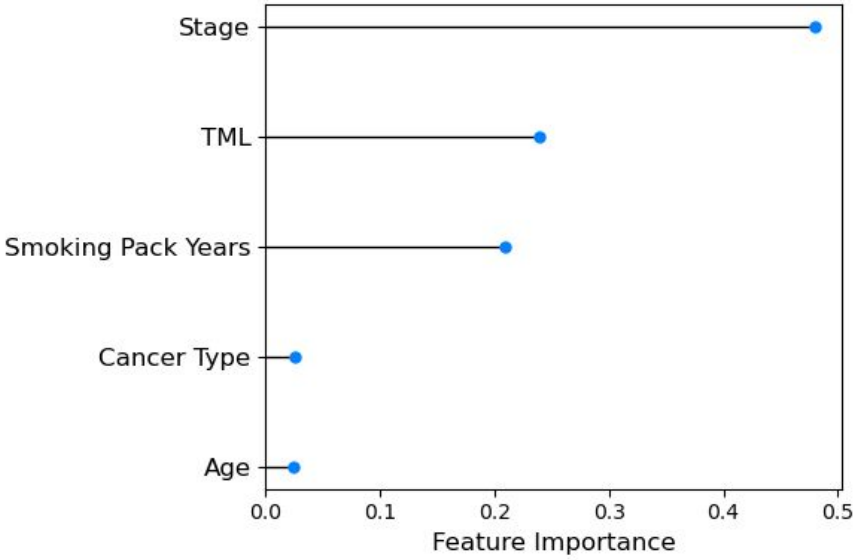


Figure 5 ***A)*** *Precision and Recall for the 3 RF models. The shape indicates the models evaluated, the black and grey border means whether or not PCA was applied. While the colour shows the resampling method used.* ***B)*** *The most important variable for classification model 1.*

1.5 Discussion

In this study, first, we analyzed the association between Total Mutational Load and clinical features in 972 lung adenocarcinoma and Lung Squamous Cell Carcinoma samples. Then we used these results to understand the metastasis status classification models using a benchmarking strategy based on random forest models. We corroborated that smoking is strongly associated with the total mutational load in LUAD. LUAD patients who have never smoked have a lower total mutational load than those who have actively smoked during their lifetime. Unlike LSCCs that have never smoked, which have a large number of mutations. In the study of (Steeg 2016), they found that patients with this pathology accumulate many passenger mutations, suggesting that LSCC is no longer just a smoker's disease since 14.7% (95% CI, 12.1% –17.4%) of their patients were never smokers. Therefore, smoking seems to be a relevant factor to explain the increase in the number of mutations in patients with lung adenocarcinoma.

When analyzing the types of smokers, we see that there is a significant difference in the number of mutations between the three types of smokers evaluated, where the patients who have never smoked have fewer mutations than the smokers or reformed patients. This shows that smoking affects the number of mutations.

Our benchmarking consisted of testing different preprocessing, classification and validation strategies, such as dimensionality reduction, data separation methods, and sampling methods. In this way we find the best way to obtain the best classification results for our data set, which consists of not reducing the dimensionality with PCA, using the Hold-Out validation method and not resampling of the data. After applying our benchmarking, we verified that the model with the clinical variables and TML obtained the best results in terms of performance. However, the contribution of TML is low, while the tumour stage is a very relevant variable in the classification. These results indicate that tumour stage II and III samples could be reclassified as metastatic samples being able to help the pathologist to classify samples considering this information. Although we have a large number of samples, the main difficulty in this work is the imbalance of classes in the metastatic status. Getting this type of information is very complex, and also most of the available data sets present this problem. Given this, in this work, we build a benchmarking to obtain the best strategy to classify this type of information. The findings in this work may contribute to the development of diagnostic tools able to classify metastasis status at an early stage using clinical information, such as the cancer type, the smoking history and the age. Although the total mutational load can improve the classification, clinical variables are currently a more available source of information than the number of missense mutations. Furthermore, the benchmarking proposed in this work can be of great help to researchers when they need to analyze complex data sets with unbalanced classes.

1.6 Conclusion

In this work, we found that Age-range, the number of cigarette packs per year and cancer type were significantly associated with the total mutational load of the Lung cancer patients In particular, we corroborated that there is a strong relationship between LUAD and the number of mutations. Interestingly, we found that for squamous, the fact of smoking does not represent an increase in the number of mutations. Also, the genomic findings for this type of cancer are in the early stages of research unlike adenocarcinoma (Heist, Sequist, and Engelman 2012). In this work, we proposed a benchmarking where we analysed different preprocessing stages, validation schemes and sampling methods. This proposed benchmarking allowed us to evaluate and select the best classification strategy considering all the available information, and the characteristics of the data set to be studied. Also, we found that clinical variables, such as cancer type, smoking status, and a number of cigarette packs, can be used together with TML to positively predict if a patient with LUAD or LSCCs can develop metastasis. This contribution could help to develop more effective tests or diagnostic tools that can allow patients with LUAD to receive treatments that are specific to their clinical characteristics, knowing their predisposition to develop metastasis.

References

Adam, George, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. 2020. “Machine Learning Approaches to Drug Response Prediction: Challenges and Recent Progress.” *NPJ Precision Oncology* 4. https://doi.org/10.1038/s41698-020-0122-1.

Cho, Han-Jun, Soonchul Lee, Young Geon Ji, and Dong Hyeon Lee. 2018. “Association of Specific Gene Mutations Derived from Machine Learning with Survival in Lung Adenocarcinoma.” *PLoS ONE* 13 (11). https://doi.org/10.1371/journal.pone.0207204.

Denisov, Evgeny V., Anastasia A. Schegoleva, Polina A. Gervas, Anastasia A. Ponomaryova, Lubov A. Tashireva, Valentina V. Boyarko, Ekaterina B. Bukreeva, Olga V. Pankova, and Vladimir M. Perelmuter. 2019. “Premalignant Lesions of Squamous Cell Carcinoma of the Lung: The Molecular Make-up and Factors Affecting Their Progression.” *Lung Cancer (Amsterdam, Netherlands)* 135: 21–28. https://doi.org/10.1016/j.lungcan.2019.07.001.

Inamura, Kentaro, Yusuke Yokouchi, Maki Kobayashi, Rie Sakakibara, Hironori Ninomiya, Sophia Subat, Hiroko Nagano, et al. 2017. “Tumor B7-H3 (CD276) Expression and Smoking History in Relation to Lung Adenocarcinoma Prognosis.” *Lung Cancer* 103 (January): 44–51. https://doi.org/10.1016/j.lungcan.2016.11.013.

Loyola-González, Octavio, Milton García-Borroto, Miguel Angel Medina-Pérez, José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and Guillermo De Ita. 2013. “An Empirical Study of Oversampling and Undersampling Methods for LCMine an Emerging Pattern Based Classifier.” In *Pattern Recognition*, edited by Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad, Joaquín Salas Rodríguez, and Gabriella Sanniti di Baja, 264–273. Berlin, Heidelberg: Springer Berlin Heidelberg.

Lu, Chaojing, Hezhong Chen, Zhengxiang Shan, and Lixin Yang. 2016. “Identification of Differentially Expressed Genes between Lung Adenocarcinoma and Lung Squamous Cell Carcinoma by Gene Expression Profiling.” *Molecular Medicine Reports* 14 (2): 1483–90. https://doi.org/10.3892/mmr.2016.5420.

Mayakonda, Anand, De-Chen Lin, Yassen Assenov, Christoph Plass, and H. Phillip Koeffler. 2018. “Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer.” *Genome Research* 28 (11): 1747–56. https://doi.org/10.1101/gr.239244.118.

Shi, Ke, Na Li, Meilan Yang, and Wei Li. 2019. “Identification of Key Genes and Pathways in Female Lung Cancer Patients Who Never Smoked by a Bioinformatics Analysis.” *Journal of Cancer* 10 (1): 51–60. https://doi.org/10.7150/jca.26908.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. 2018. “Cancer Statistics, 2018.” *CA: A Cancer Journal for Clinicians* 68 (1): 7–30.

https://doi.org/10.3322/caac.21442.

Song, Yueqiang, Donglai Chen, Xi Zhang, Yuping Luo, and Siguang Li. 2019. “Integrating Genetic Mutations and Expression Profiles for Survival Prediction of Lung Adenocarcinoma.” *Thoracic Cancer* 10 (5): 1220–28. https://doi.org/10.1111/1759-7714.13072.

Steeg, Patricia S. 2016. “Targeting Metastasis.” *Nature Reviews Cancer* 16 (4): 201–18. https://doi.org/10.1038/nrc.2016.25.

Sun, Sophie, Joan H. Schiller, and Adi F. Gazdar. 2007. “Lung Cancer in Never Smokers--a Different Disease.” *Nature Reviews. Cancer* 7 (10): 778–90. https://doi.org/10.1038/nrc2190.

Wang, Yanfang, Haowen Deng, Shan Xin, Kai Zhang, Run Shi, and Xuanwen Bao. 2019. “Prognostic and Predictive Value of Three DNA Methylation Signatures in Lung Adenocarcinoma.” *Frontiers in Genetics* 10 (April). https://doi.org/10.3389/fgene.2019.00349.

Wu, Kui, Xin Zhang, Fuqiang Li, Dakai Xiao, Yong Hou, Shida Zhu, Dongbing Liu, et al. 2015. “Frequent Alterations in Cytoskeleton Remodelling Genes in Primary and Metastatic Lung Adenocarcinomas.” *Nature Communications* 6 (December). https://doi.org/10.1038/ncomms10131.

Yu, Jiaxian, Yueming Hu, Yafei Xu, Jue Wang, Jiajie Kuang, Wei Zhang, Jianlin Shao, Dianjing Guo, and Yejun Wang. 2019. “LUADpp: An Effective Prediction Model on Prognosis of Lung Adenocarcinomas Based on Somatic Mutational Features.” *BMC Cancer* 19 (March). https://doi.org/10.1186/s12885-019-5433-7.

Zappa, Cecilia, and Shaker A. Mousa. 2016. “Non-Small Cell Lung Cancer: Current Treatment and Future Advances.” *Translational Lung Cancer Research* 5 (3): 288–300. https://doi.org/10.21037/tlcr.2016.06.07.

Zhao, Feiyang, Lei Zheng, Alexander Goncearenco, Anna R. Panchenko, and Minghui Li. 2018. “Computational Approaches to Prioritize Cancer Driver Missense Mutations.” *International Journal of Molecular Sciences* 19 (7). https://doi.org/10.3390/ijms19072113.

Supplementary Material

Supplementary table

Table 3. Precision, recall, and F1 results for the evaluated classes (M0 and M1) for the three models used, according to the different modeling techniques.

Label	Resampling	Validation Method	PCA	Accuracy	M0 F1-score	M0 Recall	M0 Precision	M1 F1-score	M1 Recall	M1 Precision
CF + TML	None	5-Fold CrossValidation	FALSE	0.96	0.98	1.00	0.97	0.66	0.51	0.93
CF + TML	Oversample	5-Fold CrossValidation	FALSE	0.95	0.97	0.98	0.97	0.57	0.51	0.64
CF + TML	Undersample	5-Fold CrossValidation	FALSE	0.80	0.88	0.81	0.97	0.31	0.67	0.20
CF + TML	None	Hold-out	FALSE	0.96	0.98	0.99	0.96	0.64	0.50	0.88
CF + TML	Oversample	Hold-out	FALSE	0.95	0.97	0.99	0.96	0.61	0.50	0.78
CF + TML	Undersample	Hold-out	FALSE	0.85	0.91	0.86	0.97	0.43	0.71	0.30
CF + TML	None	5-Fold CrossValidation	TRUE	0.96	0.98	0.99	0.97	0.63	0.51	0.83
CF + TML	Oversample	5-Fold CrossValidation	TRUE	0.94	0.97	0.97	0.96	0.53	0.49	0.57
CF + TML	Undersample	5-Fold CrossValidation	TRUE	0.80	0.89	0.81	0.97	0.32	0.67	0.21
CF + TML	None	Hold-out	TRUE	0.96	0.98	0.99	0.96	0.64	0.50	0.88
CF + TML	Oversample	Hold-out	TRUE	0.96	0.98	1.00	0.96	0.67	0.50	1.00
CF + TML	Undersample	Hold-out	TRUE	0.86	0.92	0.86	0.99	0.49	0.86	0.34
CF	None	5-Fold CrossValidation	FALSE	0.95	0.98	0.99	0.96	0.59	0.49	0.75

CF	Oversample	5-Fold CrossValidation	FALSE	0.93	0.96	0.96	0.97	0.50	0.53	0.48
CF	Undersample	5-Fold CrossValidation	FALSE	0.81	0.89	0.82	0.97	0.31	0.65	0.21
CF	None	Hold-out	FALSE	0.96	0.98	1.00	0.96	0.67	0.50	1.00
CF	Oversample	Hold-out	FALSE	0.93	0.96	0.96	0.96	0.52	0.50	0.54
CF	Undersample	Hold-out	FALSE	0.82	0.90	0.83	0.98	0.41	0.79	0.28
CF	None	5-Fold CrossValidation	TRUE	0.96	0.98	1.00	0.96	0.64	0.49	0.92
CF	Oversample	5-Fold CrossValidation	TRUE	0.82	0.89	0.83	0.98	0.36	0.73	0.24
CF	Undersample	5-Fold CrossValidation	TRUE	0.79	0.87	0.79	0.98	0.34	0.78	0.22
CF	None	Hold-out	TRUE	0.97	0.98	1.00	0.97	0.67	0.54	0.88
CF	Oversample	Hold-out	TRUE	0.85	0.92	0.85	0.99	0.39	0.85	0.25
CF	Undersample	Hold-out	TRUE	0.82	0.90	0.83	0.98	0.30	0.69	0.19
TML	None	5-Fold CrossValidation	FALSE	0.92	0.96	0.99	0.93	0.00	0.00	0.00
TML	Oversample	5-Fold CrossValidation	FALSE	0.59	0.74	0.62	0.91	0.08	0.24	0.05
TML	Undersample	5-Fold CrossValidation	FALSE	0.46	0.62	0.47	0.91	0.09	0.37	0.05
TML	None	Hold-out	FALSE	0.91	0.95	0.96	0.94	0.00	0.00	0.00
TML	Oversample	Hold-out	FALSE	0.60	0.75	0.62	0.93	0.06	0.23	0.03
TML	Undersample	Hold-out	FALSE	0.49	0.64	0.49	0.95	0.10	0.54	0.06

CAPÍTULO 2

CLASSIFICATION OF PRIMARY CANCER BASED ON MUTATION PATTERNS USING RANDOM FOREST METHOD

2.1 Abstract

To date, about 2% of patients diagnosed with cancer in the USA are classified as cancer of unknown primary origin [1]. These patients are usually treated with broad spectrum chemotherapies, since the primary organ is not identified, preventing the application of targeted and more effective therapies. On the other hand, we also know that cancer is a highly mutated entity, where the distribution of somatic mutations varies in the different tumor types. However, the biological processes that generate these mutations are not yet fully understood. In this work we propose the use of somatic mutational patterns, such as SNVs, indels and SVs, to classify the primary organ that causes cancer using random Forest models in the 2,653 samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium. As a result, we obtained that the patterns that combine the information contained in the SNVs, indels and SVs are excellent features to classify the primary tumor with F1 values above 0.7 in most of the evaluated tumors.

2.2 Introduction

Cancer is a complex disease that arises from the combined effects of multiple genetic and epigenetic variation, being the second cause of death worldwide with more than 8 million deaths per year (Campbell et al. 2020). Recent studies have determined that cancers are highly mutated entities, where these have different distributions of somatic mutations (Akdemir et al. 2020). In recent years, a new approach has been incorporated, where the complete mutational scenario of various tumours is studied considering both passenger and driver mutations to characterize cancer based on mutational patterns, capturing finally the mutational signatures of each tumour type.
With the advent of NGS and the development of genomics have been possible to identify a large number of somatic mutations classified into every somatic point mutation, copy-number change and structural variant (SV) in a given cancer.

One of the most representative initiatives that seek to capture genetic variation in cancer through complete genome sequencing is the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). This consortium contains 2,658 whole-cancer genomes and their matching normal tissues across 38 tumour types (Campbell et al. 2020)..

Currently, the way to identify a tumour type is by the organ that originates this tumour, also considering its histopathology. Recent studies have shown that when comparing the results of broad-spectrum chemotherapy with a therapy directed at the cells of the tumour of origin, the latter proves to be more effective (Jiao et al. 2020).
Consequently, classification based on patterns of somatic mutations could be useful to identify patients where the site of origin of the tumor is unknown, in order to find the most adequate and effective therapeutic strategy for these patients.

Here we used a Random Forest (RF) model on whole-genome sequencing (WGS) data from 2653 samples from the PCAWG project to predict Tumor type, based on the profile of all types of mutations, including single base substitutions, short insertions and deletions (indels) and structural variants.

2.3 Methods

2.3.1 Data set and preprocessing

We based this work on data generated by Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). This consortium collected genome 2.658 donors across 38 tumours, which whole genome sequencing of 2605 primary tumours and 173 in the metastatic status was arranged.
The pipelines to identify the somatic mutations, for example, Somatic single-nucleotide variations (SNVs), small insertions and deletions (indels), copy-number alterations (CNAs) and SVs are described in (Campbell et al. 2020).

In this work, WGS data for 2653 samples from PCAWG project for 37 tumour types were used, including Liver-HCC (315 samples), Panc-AdenoCA (238 samples), Prost-AdenoCA (210 samples) and Breast-AdenoCA (198 samples) among the most representative tumour types (the full set of tumour types is available in Supplementary Table S4). Tumour types with less than three samples were not considered (total of 4 tumour types). Data were obtained in plain text tabular format, including individual-level mutations and associated annotations, as described in Supplementary Text. We analyzed 1,234,564 Structural Variants (SV), 3,921,229 short insertions and deletions (indels) and 43,778,859 Single nucleotide Variants (SNVs) (See Figure 10).

2.3.2 Feature engineering

To evaluate the predictive value of genomic features, we created mutation profiles for SNV, indels, CNA, and structural variations SV. First, we classified each somatic SNVs in 96 classes considering the mutation and the 3-bp sequence context as described by (Alexandrov et al. 2013). These 96 classes are known as SNV mutation classes. We built a SNV matrix, where the columns correspond to 96 SNV mutation classes, while each row corresponds to a sample. Therefore, each position (i,j) of the matrix indicates the frequency of the j^{th} class in the i^{th} sample normalized. The SomaticSignatures package R from Bioconductor was used for this processing. Then, we classified the indels according to their length in base pair (bp). We selected 5 categories for each insertion and deletion: 1 bp, 2 bp, 3 bp, 4 bp and >5 bp. Then, we generated a matrix of indels frequencies by sample for each category. The Structural variants (SV) are longer insertions and deletions, duplications, inversions, translocations of at least 50 bp (Kosugi et al. 2019). We analyzed only inversions, indels, duplications and translocations. For each sample, we obtained the number of SVs mutations building a frequency matrix as we previously described. We

calculated its GC% for each 2-kb sequence context for all SNVs, and then we stratified each mutation by the GC% quantile distribution. We obtained a matrix with the frequency of the 4 quantile categories for all samples. To evaluate the predictive power of the chromosomes regions, we divided the genome into bins of size 1 Mbp and 10 Mbp. Then, we created a matrix for each bin size with a total number of SNV by bin for each sample. We excluded the sex chromosomes in this analysis. In the case of SNV per 1 Mbp bin, we generated 2897 features, hence we must reduce the dimensionality across Principal Component Analysis (PCA). Thus, we used as predictors the 100 first principal components.

To explore the relationship between mutation classes by tumor types and organ system we used a technique for dimension reduction called Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes, Healy, and Melville 2018).

2.3.4 Tumour-type class prediction by Random Forests

We implemented a multiclass classification approach through the Random Forest (RF) algorithm to identify tumour type based on 96 classes of single base substitutions, indels and structural variants, %GC quantile and SNV bins for 2653 samples. The input data consisted of the previously described mutation matrices and a vector of known tumour types as the response variable. We built 7 models using different combinations of somatic mutation classes as predictors:

- **Model 1:** tumor type ~ SNVs classes
- **Model 2:** tumor type ~ INDELS
- **Model 3:** tumor type ~ SVs
- **Model 4:** tumor type ~ SNVs + INDELS + SVs
- **Model 5:** tumor type ~ %GC quantile
- **Model 6:** tumor type ~ SNV per 1 Mbp bin
- **Model 7:** tumor type ~ SNV per 10 Mbp bin

For each model, we partitioned the input matrices into training and testing data using a proportion of 75% and 25%, respectively. All tumour types were modelled together. However, RF internally trained a model for each tumour type separately, where the response was coded as a binary variable indicating class membership. For the partition and training process, we used a *createDataPartition()* and *train()* functions from the caret R package. R version 3.6.2 (2019-12-12) was used in all analyses.
The RF algorithm performance is affected by a number of parameters, such as the type of resampling, among others. We explored the combination of parameters that maximized performance by setting the "search" argument of the trainControl() function to "random". This tried random combinations for parameters until a given set of tries, which was set to 10 (tuneLength argument). The model was further

optimized by minimizing classification error in a set of 10 cross-validations. Parameter tuning, that is the selection of the best hyperparameters in the model [7], was done in each cross-validation step. The Accuracy metric was used to select the optimal model in the training models process.

2.3.5 Model performance assessment

Once the model was trained, we assessed its performance in the test dataset. We computed the accuracy as a global metric to evaluate different feature combinations. We also evaluated the performance of prediction on each primary tumour site by the precision, recall, and F1 score:

$$\text{Accuracy} = \Sigma_{yi} TP(y_i) / \Sigma_{yi} Pred(y_i)$$
$$\text{Recall}(y_i) = TP(y_i) / true(y_i)$$
$$\text{Precision} = TP(y_i) / Pred(y_i)$$
$$F1 = 2*((Precision(y_i) * Recall(y_i))/(Precision(y_i) + Recall(y_i))$$

Where:
y_i : tumour type class
$TP(y_i)$: number of true positives
$Pred(y_i)$: number of the membership predictions for the given class
$True(y_i)$: number of true members of the class

2.4 Results

We obtained an SNVs matrix with 2635 samples and 96 features that correspond to the mutation classes. To understand the relationship between tumor types and SNVs data, we applied the UMAP algorithm. We obtained clusters of SNVs classes where for some tumor types the groups are easily identifiable, for example, Kidney RCC, Skin melanoma, Liver-HCC, among others (See Figure 6). However, when we include the other types of mutations such as SNVs, SVs and indels, the structure obtained presents a greater separation between the groups, being able to identify more tumor types.

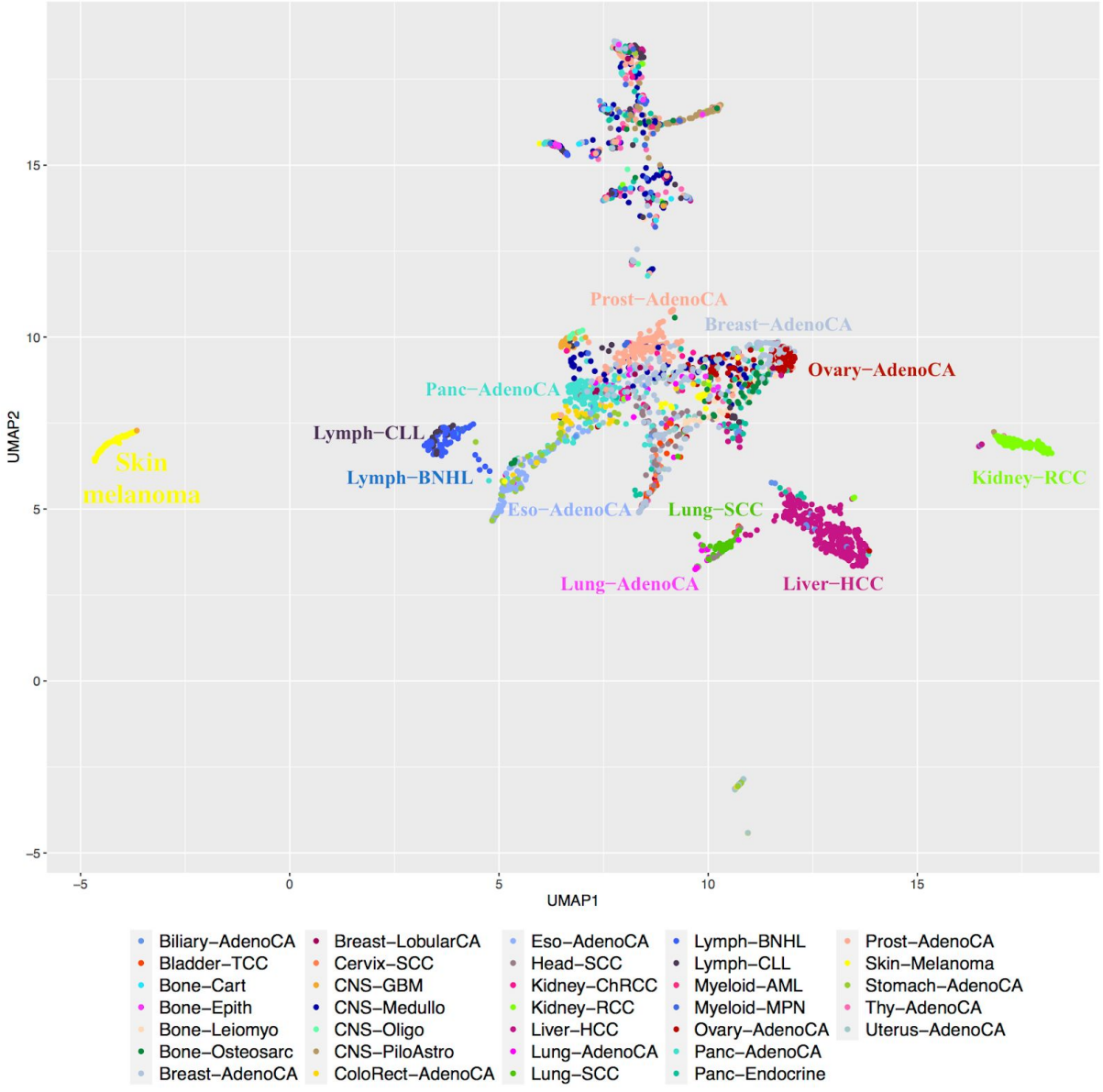


Figure 6. UMAP projection of the PCAWG dataset, color-coded by tumor type labels.

In Figure 7, we present the overview of the methodology applied in this study. This methodology consists of two main stages, preprocessing and classification, respectively. In the first stage, we build the features based on the somatic mutations such as SNVs, Indels and SVs. For each type of mutation, we build a frequency matrix with their respective features. The %GC quartile, SNV per 1 Mbp and 10 Mbp bins are based on SNV classes. Then, we added the tumor types labels to each mutation matrix described before. We filtered the dataset by the number of samples for each tumour type. Therefore, we removed Myeloid-MDS (2 samples), Breast-DCIS (3 samples), Lymph-NOS (2 samples) and Cervix-AdenoCA (2 samples) from the data. Therefore, 33 tumour types were analyzed in this study.
In the classification stage, each mutation matrix, with its respective labels, is divided into training and testing data sets. For each training dataset, we applied the Random Forest algorithm using a 10 fold-cross validation, selecting the best hyperparameters in the models. This process is known as tuning parameters. Before, we evaluated the testing dataset for each training model, obtaining the performance measures.

In the classification Stage, we evaluated seven mutational models: SNVs, SVs, indels, triplet mutations (SNV classes + indels + SVs), %GC quartile, SNV per 1 Mbp bin and 10 Mbp bins. Each dataset was divided into training and test sets using a ratio of 75% and 25%, respectively. We use the training sets to apply the tuning parameters, where we train ten internal models with different hyperparameters, to improve precision values. Once the training process is finished, we obtained the best internal model for each mutational model, which is used to evaluate the testing dataset, generating the performance scores.

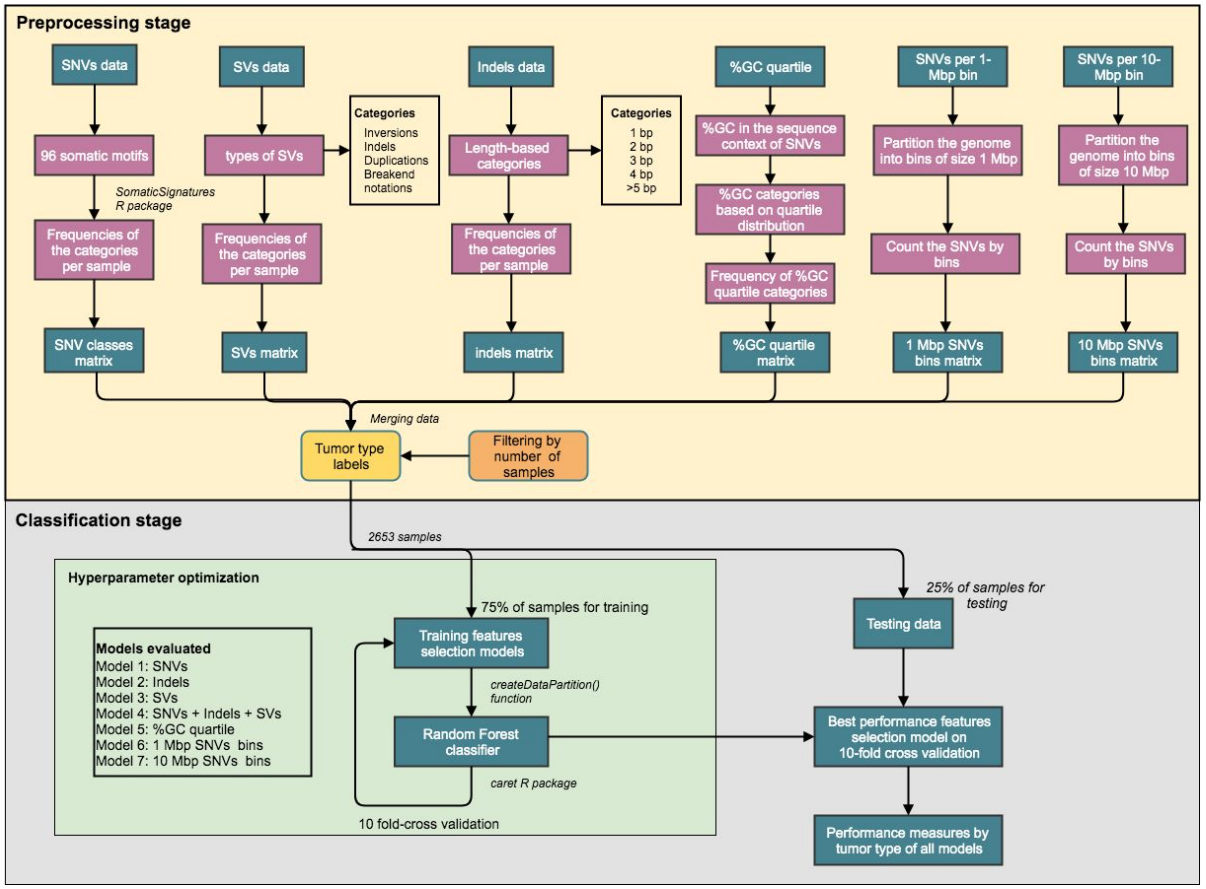


Figure 7. **Overview of the preprocessing and classification stages.** The first stage consists in the preprocessing of SNVs, SVs and indels to build features which will later be used to classify the tumor type. Specifically, we created seven matrices with the frequencies of the types of mutations already mentioned, in addition to including the genomic context, we incorporated the number of SNVs per 1 Mbp and 10 Mbp bin sizes. In the next stage, the matrices are partitioned into a training set and a test set. Seven mutational models are trained using 10 cross-validation and evaluating different parameters of the random forest algorithm to improve precision results. In this process several models are generated, where the best in terms of precision is chosen. Therefore, seven final models with the best precision values are obtained to evaluate the testing set. Finally, we obtained the performance scores.

2.4.1 Classification performance

The Accuracy, precision, recall and F1 score For each model evaluated across tumour types are found in the supplementary material. In Figure 8, we present the ordered performance results of the RF models for the 33 tumour types evaluated. We used the F1 score to compare the models since it is an overall measure of a model's accuracy that combines precision and recall. Therefore, a high F1 means low false positives and low false negatives. We show the best F1 score of the 7 RF models reached for each tumour type, where the colour circle indicates the model. The kidney-RCC, Lymph-BNHL, Skin-Melanoma, Lymph-CLL, CNS-GBM, Liver-HCC had a highest F1 score (over 0.85), being reached mostly by the SNVs and triplete mutation models. Otherwise, the Myeloid−AML, Cervix−SCC, Breast−LobularCA, Bone−Epith, Bone−Cart and Biliary−AdenoCA had really bad precision and recall values, throwing NA values for F1. Interestingly, within tumour types with an F1 greater than 0.5, we found that for CNS−PiloAstro the best F1 comes from the indels model, being the only tumour type, where indels are more predictive than the other proposed models. The Structural variants and %GC quartile turned out not to be predictive in these tumour types.

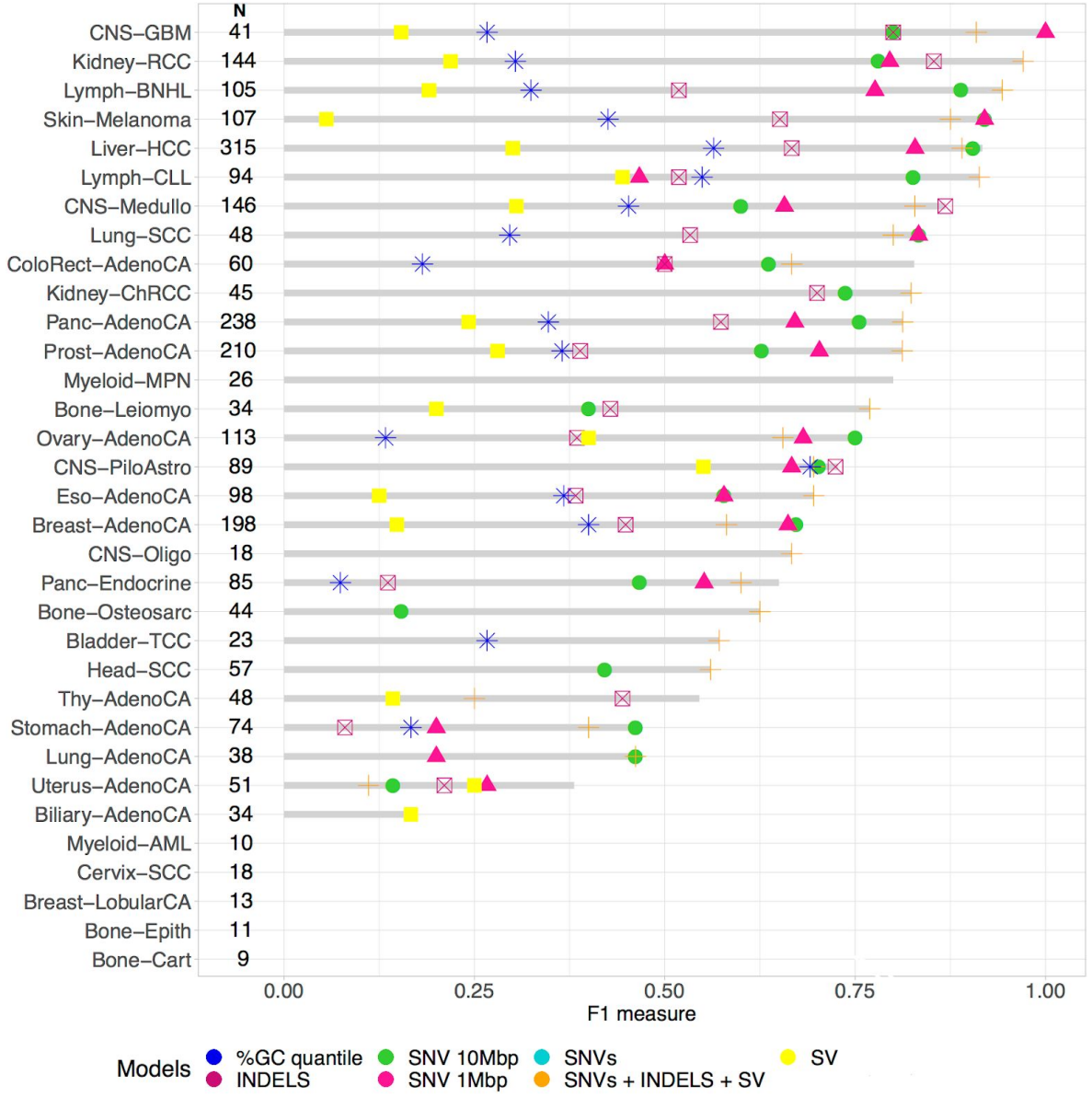


Figure 8. **F1 scores of all models evaluated for 33 tumor types.** The colors and shape of each point indicate the mutational model used.

In the following figure, we plot the precision and recall values for the triplet mutational model since it is the model that obtains the best results in most of the tumor types. Interestingly, we found that the tumor types with the best precision and recall ratios are the same that presented a good grouping in the UMAP clustering.

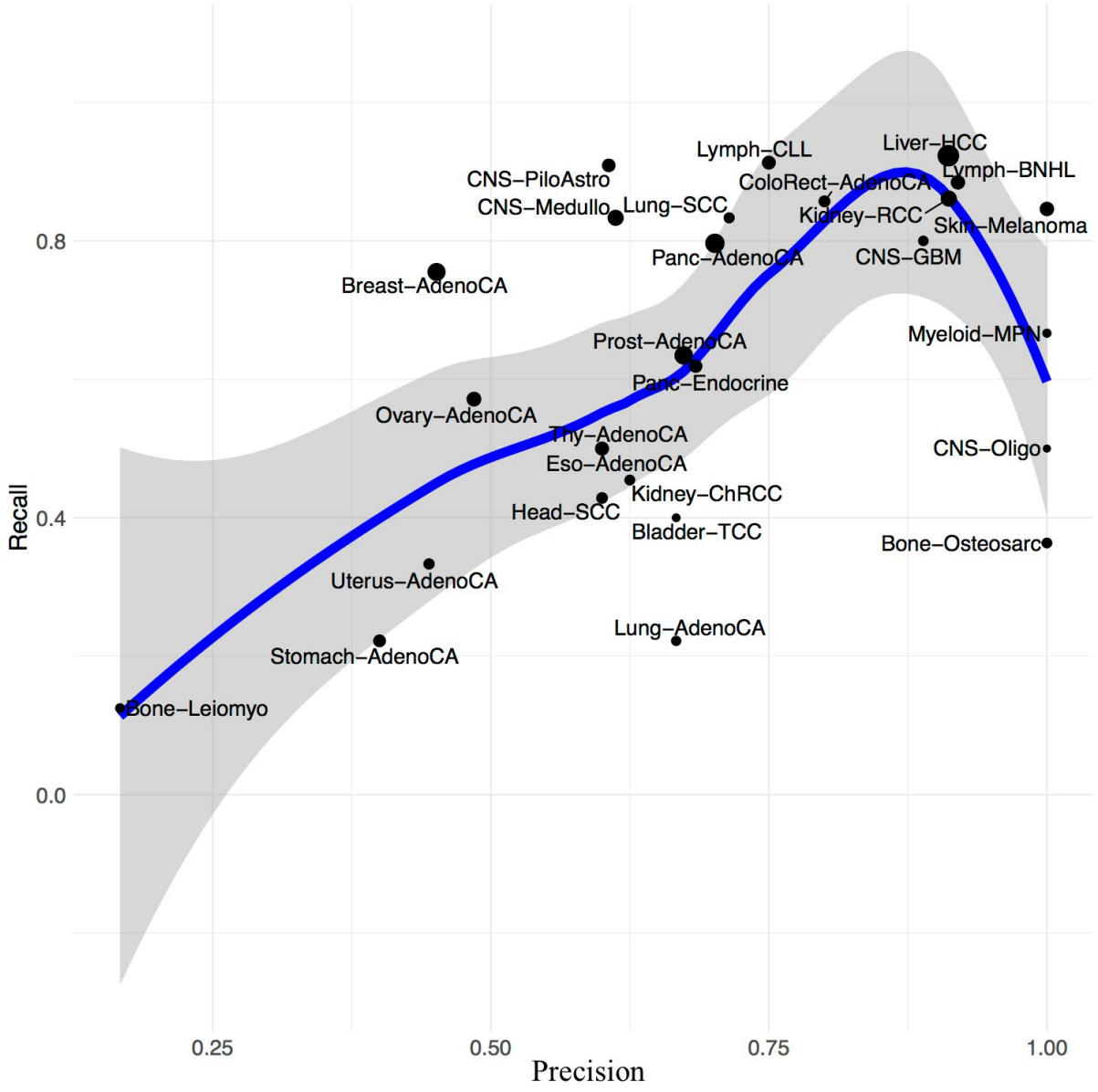


Figure 9. The relationship between the Recall and Precision for triplet model is shown for each tumour type. The dot size represents the number of samples for each tumor type. The blue line represents a regression line fit using LOESS regression, while the grey area represents a 95% confidence interval.

In Table S5 we expose the F1 scores of the 33 tumor types for all mutational models. The SNVs model obtained the best F1 scores in 13 tumor types, while the triplete mutational model obtained 15 tumor types with the best results. Furthermore, the SVs and the %GC quantile were not predictive of any of the tumor types analyzed in this work. In Table 4 we present a summary of the performance of each model for the classification of tumour types based on the proportion of high F1 scores. Triplets are the ones that best classify the greatest number of tumours. However, the SNV classes are the ones that contribute in an important way to this classification, since indels and SVs by themselves do not generate good classifications. In Table 5 we show the details of the results for all tumor types.

Table 4**.** Proportion of tumor types classified with an F1 greater than 0.7 for each model evaluated and the proportion of tumor types where the model evaluated turned out to be the best.

	SNV	**indels**	**SV**	**Triplete**	**F1_GC**	**SNV 1Mbp**	**SNV 10Mbp**
Number of F1 scores > 0.7	0.39	0.12	0	0.45	0	0.39	0.21
Number of highest F1 scores	0.12	0.03	0	0.18	0	0.21	0.03

2.5 Discussion

In this work, we analyze the classifying power of somatic mutations obtained by whole genome sequencing in the largest collection of tumor types. Some tumour types share similar mutational patterns considering the SNVs, SVs, and indels. For example, When we analyze the clustering of SNV classes generated by UMAP, we obtain that for some tumour types, such as Kidney-RCC, Liver-HCC, Skin-Melanoma, they present a similar mutational profile given the level of grouping of the samples in well-defined clusters.

Interestingly, Lung−AdenoCA and Lung−SCC that corresponds to Non-Small Cell Carcinoma (NSCLC), have a similar mutational pattern in most samples, but in clinical terms, they come from different epithelial cell type (See Figure 11). Furthermore, we noted that adenocarcinomas such as Breast-AdenoCA, Prost-AdenoCA and Panc-AdenoCA, tend to group, but do not reach the Kidney-RCC or Liver-HCC cohesion level. In the case of Eso−AdenoCA and Stomach−AdenoC, these present the same structure when we observe UMAP analysis, this could be due to both being adenocarcinomas, having their origin in epithelial cells. Furthermore, since both originate in organs of the digestive system, they could accumulate mutations due to their exposure to environmental factors, such as drugs, smoke, among others (Parsa 2012).

The SNV and triplet models were the ones that obtained the best F1 score results in most of the tumour types evaluated. Also, the genomic distribution of SNVs obtained good performance results as well as SNVs and triplets models. The accumulation of mutations in the chromosomal context can be explained for the chromatin accessibility to DNA repair complexes, which would be affecting the epigenetic process in the cancer cell (Jiao et al. 2020).

Although there are other works that focus on the classification of tumour types (Jiao et al. 2020; Sun et al. 2019; Lyu and Haque 2018) these use techniques based on deep neural networks, of which it is very complex to decompose the main factors that influence the classifications, unlike the models that use the Random Forest as a classifier.

One of the most important limitations in this work was the low number of samples for some tumour types, such as bone-Cart, myeloid-AML, among others, which could interfere with the classification results.

2.6 Conclusion

Mutational patterns, specifically SNV classes and the triplet, are useful to classify the tumor type, which obtained very good results in terms of performance. These show that there are specific mutational processes that generate mutations that cause cancer to appear in particular organs. Therefore, the use of these mutational features would allow us to define more adequately the therapeutic strategy in patients where it is not possible to identify the type of tumour they have through histology.

References

Akdemir, Kadir C., Victoria T. Le, Justin M. Kim, Sarah Killcoyne, Devin A. King, Ya-Ping Lin, Yanyan Tian, et al. 2020. “Somatic Mutation Distributions in Cancer Genomes Vary with Three-Dimensional Chromatin Structure.” *Nature Genetics*, October, 1–11. https://doi.org/10.1038/s41588-020-0708-0.

Campbell, Peter J., Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, et al. 2020. “Pan-Cancer Analysis of Whole Genomes.” *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Jiao, Wei, Gurnit Atwal, Paz Polak, Rosa Karlic, Edwin Cuppen, Alexandra Danyi, Jeroen de Ridder, et al. 2020. “A Deep Learning System Accurately Classifies Primary and Metastatic Cancers Using Passenger Mutation Patterns.” *Nature Communications* 11 (February). https://doi.org/10.1038/s41467-019-13825-8.

Kosugi, Shunichi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. 2019. “Comprehensive Evaluation of Structural Variation Detection Algorithms for Whole Genome Sequencing.” *Genome Biology* 20 (1): 117. https://doi.org/10.1186/s13059-019-1720-5.

Lyu, Boyu, and Anamul Haque. 2018. “Deep Learning Based Tumor Type Classification Using Gene Expression Data.” *BioRxiv*, July, 364323. https://doi.org/10.1101/364323.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *ArXiv:1802.03426 [Cs, Stat]*, December. http://arxiv.org/abs/1802.03426.

Parsa, N. 2012. “Environmental Factors Inducing Human Cancers.” *Iranian Journal of Public Health* 41 (11): 1–9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3521879/.

Probst, Philipp, Marvin N. Wright, and Anne-Laure Boulesteix. 2019. “Hyperparameters and Tuning Strategies for Random Forest.” *WIREs Data Mining and Knowledge Discovery* 9 (3): e1301. https://doi.org/10.1002/widm.1301.

Sun, Yingshuai, Sitao Zhu, Kailong Ma, Weiqing Liu, Yao Yue, Gang Hu, Huifang Lu, and Wenbin Chen. 2019. “Identification of 12 Cancer Types through Genome Deep Learning.” *Scientific Reports* 9 (1): 17256. https://doi.org/10.1038/s41598-019-53989-3.

“Unknown Primary - Statistics.” 2012. Cancer.Net. June 25, 2012. https://www.cancer.net/cancer-types/unknown-primary/statistics.

Supplementary Material

Supplementary text

Input data format

Input data is provided in a text file such that each row corresponds to one somatic mutation. Columns in the tab-separated input data file are the following:

1. chrom, chromosome where the mutation occurs.
2. pos, chromosome position where mutation occurs.
3. ref, reference allele in the source VCF file.
4. alt, alternative allele in the source VCF file.
5. sample, unique PCAWG sample identified (UUID)
6. seq, 2048 bp of flanking DNA sequence + somatic mutations occurring in this region and sample centered at the mutation position (chrom:pos)
7. genic, 1 if mutation occurs in a gene, else 0
8. exonic, 1 if mutation occurs in an exon, else 0
9. strand, + or - if mutation occurs in a gene (+ if pyrimidine reference base occurs on positive strand, else -); ? if mutation overlaps with two genes of opposite orientation; = if mutation does not occur in a gene
10. rt, mean replication timing across five cell lines (source: PCAWG).

Sequence is encoded with an extended DNA alphabet consisting of {A, C, G, T} and letters for different mutation types (see /csc/epitkane/data/PCAWG2019/mutation_codes_sv_mei.tsv for encoding). For example, C:G>A:T substitution is encoded as '$'.

Example: somatic C:G>A:T substitution. Note the '$' in sequence (highlighted in orange). This $ occurs in the middle of the sequence which is position 14:35860669.

chrom pos ref alt sample seq genic exonic strand rt
14 35860669 G T bcf858fd-cc3b-4fde-ab10-eb96216f4366
GCCACAGTTTTGGAGGCTGCAAAGTCCAAGATCAAGATGCTAGCAGATTCTGTGTCTG
GTGAGGGCCAGCTTTTCTATTTGAAGATGAGGCCTTCTTGCTGTGTCTGCACGTGGTGG
AAAGGGCCAGCTAGCTTGCTGGGGCTTCTTCTATAAGGGAACTAATCCCATTCAGGAG
GGCTCCATCCTTATGACCTAATCACTTCCCAAGGCCCCACCTGACACCATCACCTTGGG
GATTCAGATTTCAACATATGGATTTTGAAGGGACACTAACATTCAGTTCATTGTAATAGC
CTTAGGTGCCATCTTGACACCCAAGACTTGGCTGCTGGCAAAGAGGCTCCAGGTAGCT
CTGGTCCCAGCTCAGGATACTATCCCAGGAGGGTTTTTTTTTCTTTTTTTCATTTTGGAG
ATAGGGTCTCACTCTGTCACCTAGGCTAGAGTGCAGTGGCACAATCATGGCTCACTGC
AGCCTCAGCCTCCTGGGCTTAAATGATCCTCCCACCTCAGCCTTCCGAGTGTGTGGGA
TCCCAGGTGCATGTCACCAAGCCCAGCTAATTTTTTTTTTTTTTGAGAAAGAGTCTTACT
CTGTTGCCCAGGCTGGAGTGCAGTGGCGCAATCTTGGCTCACTGCAACCTCCACCTCC
CAGGTTCAAGTGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGCATGTG
CCACTATGTGCACCTAATTTTTGTATTTTTAGTAGAGATGGGGTTTCACCATGTTGGCCA
GGTTGGTCTCGAACTCCTGACCTCAAGTGATCCTCCTGCCTTGGCCTCCCAATGTGCTG

GGATTACAAGTGTGAGCCACCACACCCAGCCCTGGGATGGTTTTTGACAGCAGAAACC
TACCAGCAAAGGAGAAAAAACTCTGCCTCCTCAGAGGCTGCAGGGAAGCCACGGCTGA
GGTAATTGTGTGATTAAGTTTTGAGCCCATTTCCAGCTGACCCACTGGGGTGGAGATGG
GTATAGAATGAAGACGCCTGCGAGGC**$**TGTCTGAGGGGAGACCTGGGAAGGCCAGGT
CCTGGGAAGGCCAACAACCAGGAACACCCTGGGTGGCTCTCATGACAGAGACA&CCTT
CAGAGCAATCGATCATTCTCCATTGGAGGTTCAGACCAACCTGGCTCCCACTGGAGGC
CCTACTCTCTACACCATGTAACCCTGGGCTAGTGGCTTAACCTCTCTGAGACTCAATTC
CTCATCTCTAAAGCAGGGTGACCGCAGCTGCCTTGCAGAGTGGCTGTGAGGCTTAGCT
ATGACTCATGATGTCTGTAAAGCTCTGGCAAAGTGCCAGACACATAGTAGGCACTCAGT
AGGTGACAGCTTCAATTAGGATGATTAATTATCACTAATGTGAAGAGCATCTGACAGCC
CTCCCTGTAGGCACGTTGTTCCCTGCAAGCCAAAACTTCCTGTTCCCCTACTCATAGGG
TGGCCAGGACCTGAGGAACCAGATCCTCTAAGGACAAACTGGAACATCTAGTCTTCCT
CTTGACCCTCCCTGACTTTGTCCCTTGCCAGTGGGGAGCCCCCTTCACCAGGCCCTTC
CCAGCCTCCAACACCTGCTCAGCGCAGCCCTGACCCACGGGAGCTGCCTTGAGGCTG
GGGGGCATGACCCACCAGGTGGAGGCATGAAACGGAGGGCTGTCATGGGCCCACCC
CAGGTGCAAGCCTGGAAGGCCCTGGACCTCTGGGAGTCCTCTCAAGTCATGGAAGGT
GAGAGAAGCCTCTGCTCCTGTGGTCAGCAGTGGGCCAGGTGGCAGATCGCTGACGGT
AATGGGGATGCAACTCAGTAGGGAAGCCCGACCTGGTTAGAGAGAGATGAACACCCCC
TCCTGGGGCCTTGCCTGGAGCCCTCTGTGAGGGGGTCAGCAAATCAGGCAGAACCTT
GGGCCCTCAGCCCTCCCCATCCTGCCTCTGTAATGCAGAAAGTGGGCTTTATTATAAAT
GCTTTCCTTTTCTTCTACCACTGGTGGGAAAGGCGGGGAAGGAGAGAGCCTTGCTCTA
GTGGCAGC 0.000000 0.000000 = 77.417648

Supplementary tables

Table 5. Description of samples in the input dataset by tumor type.

Tumour type	Average Age	Number of male samples	Number of female samples	Total sample size
Biliary-AdenoCA	64.50	19	15	34
Bladder-TCC	65.35	15	8	23
Bone-Cart	-	7	2	9
Bone-Epith	-	6	5	11
Bone-Leiomyo	-	19	15	34
Bone-Osteosarc	-	21	23	44
Breast-AdenoCA	55.83	1	197	198
Breast-DCIS	52	0	3	3
Breast-LobularCA	54.92	0	13	13
Cervix-AdenoCA	39	0	2	2
Cervix-SCC	39.33	0	18	18
CNS-GBM	57.73	28	13	41
CNS-Medullo	12.34	79	67	146
CNS-Oligo	40.72	9	9	18
CNS-PiloAstro	9.57	42	47	89
ColoRect-AdenoCA	64.77	30	30	60
Eso-AdenoCA	68.32	84	14	98
Head-SCC	52.63	47	10	57
Kidney-ChRCC	49.44	26	19	45
Kidney-RCC	60.39	90	54	144
Liver-HCC	65.51	226	89	315
Lung-AdenoCA	65.29	18	20	38
Lung-SCC	66.4	38	10	48
Lymph-BNHL	50.49	55	50	105
Lymph-CLL	62.59	64	30	94
Lymph-NOS	32	1	1	2
Myeloid-AML	47.9	7	3	10

Myeloid-MDS	75.5	1	1	2
Myeloid-MPN	56.58	12	14	26
Ovary-AdenoCA	60.65	0	113	113
Panc-AdenoCA	65.64	119	119	238
Panc-Endocrine	56.99	55	30	85
Prost-AdenoCA	58.93	210	0	210
Skin-Melanoma	56.96	69	38	107
Stomach-AdenoCA	64.93	56	18	74
Thy-AdenoCA	52.1	11	37	48
Uterus-AdenoCA	67.63	0	51	51

Table 6. F1 score by tumor types for all models evaluated. The largest F1 values for each type of tumour are highlighted in bold.

Tumor types	Organ system	SNV	indels	SV	Triplete	F1_GC	SNV 1Mbp	SNV 10Mbp
Lymph-CLL	blood, bone marrow, & hematopoietic sys	0.824	0.519	0.444	**0.930**	0.549	0.600	0.467
Myeloid-AML	blood, bone marrow, & hematopoietic sys	-	-	-	-	-	-	-
Myeloid-MPN	blood, bone marrow, & hematopoietic sys	**0.800**	-	-	-	-	-	-
Bone-Cart	bones & joints	-	-	-	0.667	-	-	-
Bone-Epith	bones & joints	-	-	-	-	-	-	-
Bone-Leiomyo	bones & joints	0.143	0.429	0.200	0.500	-	0.667	-
Bone-Osteosarc	bones & joints	0.533	-	-	0.462	-	0.588	-
CNS-GBM	brain, & cranial nerves, & spinal cord, (excl. ventricle, cerebellum)	0.842	0.800	0.154	0.900	0.267	0.800	**1.000**
CNS-Medullo	brain, & cranial nerves, & spinal cord, (excl. ventricle, cerebellum)	0.706	**0.868**	0.305	0.778	0.452	0.611	0.657
CNS-Oligo	brain, & cranial nerves, & spinal cord, (excl. ventricle, cerebellum)	0.667	-	-	0.400	-	-	-
CNS-PiloAstro	brain, & cranial nerves, & spinal cord, (excl. ventricle, cerebellum)	0.727	0.724	0.551	**0.727**	0.691	0.588	0.667

Breast-AdenoCA	breast	0.565	0.449	0.148	0.531	0.400	**0.723**	0.662
Breast-LobularCA	breast	-	-	-	-	-	-	-
Cervix-SCC	cervix uteri	-	-	-	0.667	-	-	-
Eso-AdenoCA	esophagus	0.545	0.383	0.125	0.585	0.367	**0.714**	0.578
Biliary-AdenoCA	gallbladder & extrahepatic bile ducts	-	-	0.167	-	-	-	-
Head-SCC	gum, floor of mouth, & other mouth	0.500	-	-	0.500	-	0.583	-
Kidney-ChRCC	kidney	0.526	0.700	-	**0.824**	-	0.308	-
Kidney-RCC	kidney	0.886	0.853	0.219	**0.985**	0.304	0.706	0.795
ColoRect-AdenoCA	large intestine, (excl. appendix)	**0.828**	0.500	-	0.769	0.182	0.759	0.500
Biliary-AdenoCA	liver	-	-	0.167	-	-	-	-
Liver-HCC	liver	0.917	0.667	0.300	0.889	0.564	**0.944**	0.829
Lung-AdenoCA	lung & bronchus	0.333	-	-	0.333	-	0.533	0.200
Lung-SCC	lung & bronchus	0.769	0.533	-	**0.833**	0.296	0.800	0.833
Lymph-BNHL	lymph nodes	**0.902**	0.519	0.190	0.889	0.324	0.767	0.776

Ovary-AdenoCA	ovary	0.525	0.385	0.400	0.656	0.133	**0.741**	0.682
Panc-AdenoCA	pancreas	0.746	0.574	0.242	0.769	0.347	**0.814**	0.671
Panc-Endocrine	pancreas	0.650	0.136	-	0.690	0.074	0.400	0.552
Prost-AdenoCA	prostate gland	0.653	0.389	0.280	0.750	0.365	**0.893**	0.703
ColoRect-AdenoCA	rectum	**0.828**	0.500	-	0.769	0.182	0.759	0.500
Skin-Melanoma	skin	0.917	0.651	0.056	0.894	0.426	**0.960**	0.920
Stomach-AdenoCA	stomach	0.286	0.080	-	0.222	0.167	0.308	0.200
Thy-AdenoCA	thyroid gland	0.545	0.444	0.143	**0.800**	-	-	-
Bladder-TCC	urinary bladder	0.500	-	-	0.286	0.267	-	-

Supplementary Figures

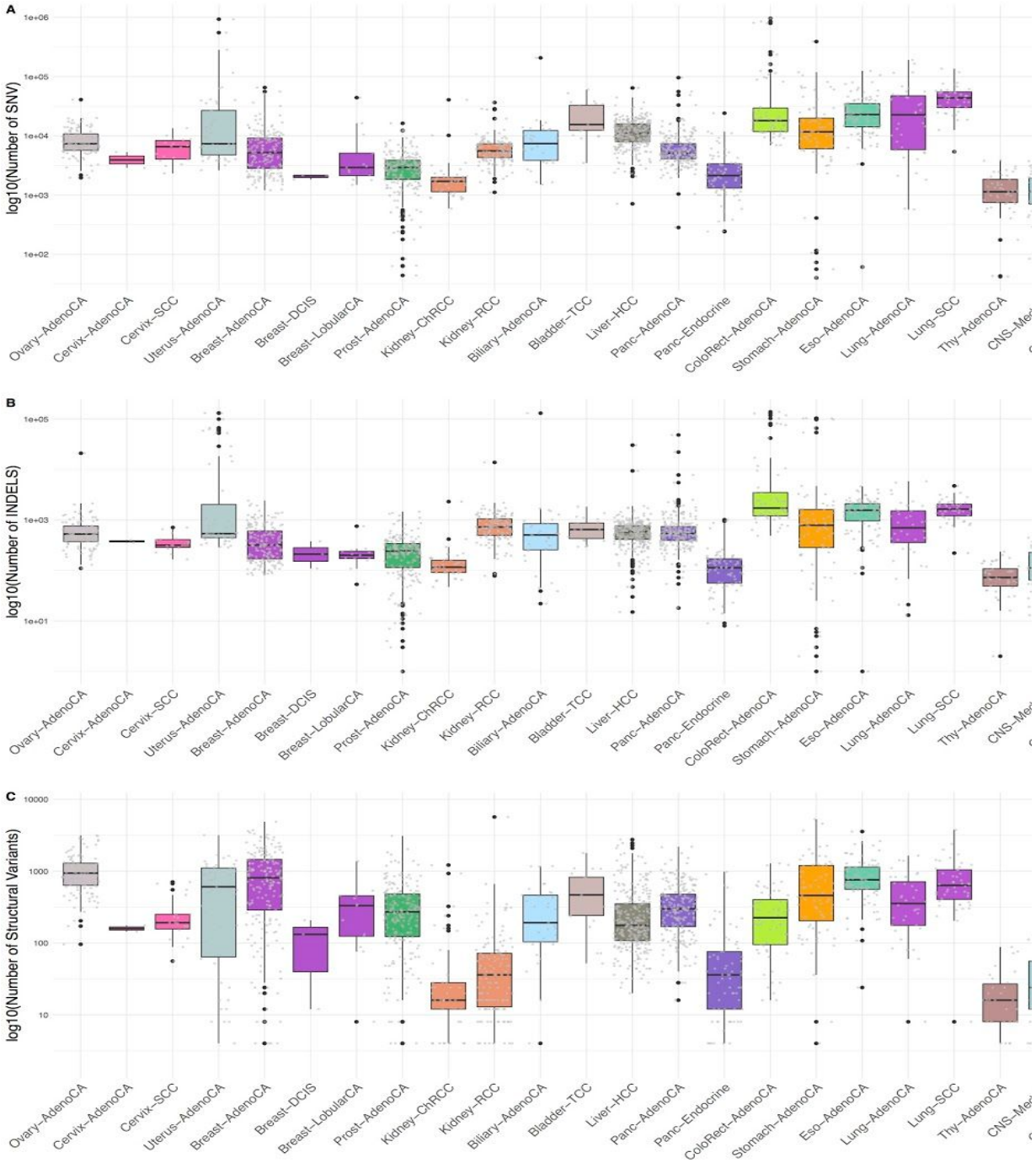


Figure 10**. Distribution of number of mutations by tumor type.** Number of SNVs (upper), INDELs (middle), and SV (bottom) are grouped and colored by organ.

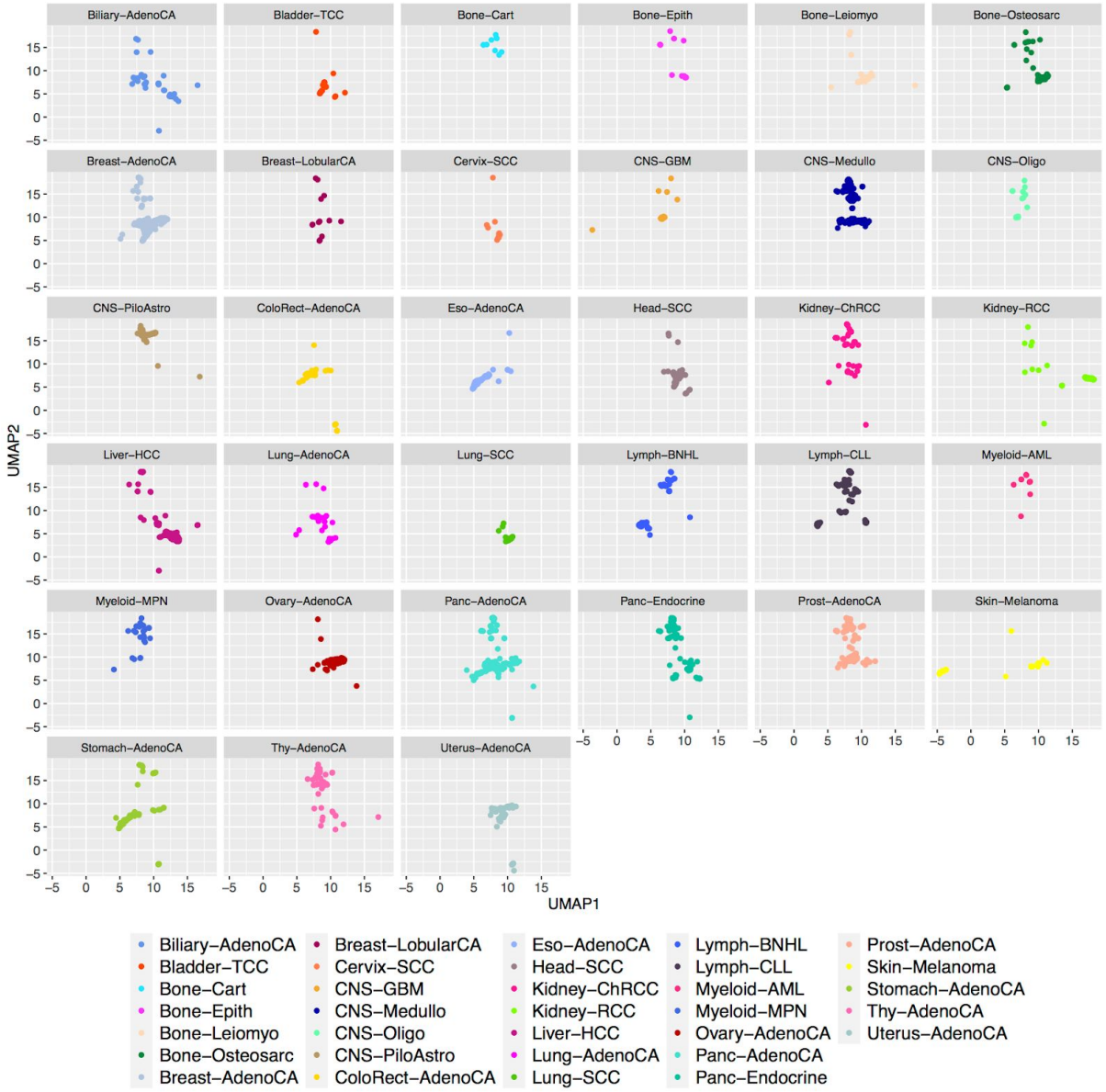


Figure 11. **UMAP of SNVs by tumor type.**

CONCLUSIONES GENERALES

El principal objetivo de esta tesis fue implementar nuevas estrategias computacionales para la identificación de patrones o relaciones entre los datos genómicos y/o clínicos en el estudio del cáncer. Para cumplir con este objetivo se definieron 3 objetivos específicos:

1. Determinar la asociación entre Carga mutacional tumoral y las características clínicas en pacientes con cáncer de pulmón.
2. Clasificar el estado metastásico de pacientes con cáncer de pulmón basándose en la relación de Carga Mutacional Total y las características clínicas.
3. Identificar patrones somáticos que permitan clasificar el tipo tumoral de origen en muestras provenientes del Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium.

En el primer capítulo de esta tesis, abordamos los primeros dos objetivos específicos. En este capítulo se encontró que la carga mutacional total está asociada significativamente al estado de fumador, subtipo tumoral y al rango de edad. Además, el modelo clasificatorio de Random forest, que incluye la Carga Mutacional Total, el estado de fumador, el rango de edad y el grado tumoral, obtuvo el mejor resultado de rendimiento con un puntaje de F1 de 0.62.

En el Segundo capítulo de esta tesis se abordó el tercer objetivo propuesto. En este capítulo se propuso que la acumulación de los distintos tipos de mutaciones somáticas permiten clasificar con un excelente rendimiento el tipo tumoral en 33 tipos de cánceres superando en la mayoría de los cánceres el puntaje F1 de 0.7. Específicamente, el mejor modelo generado es el que combina tanto las Variaciones de nucleótido único (*Single Nucleotide Variants*, SNVs), pequeñas inserciones y deleciones (Indels) y las Variaciones Estructurales (*Structural Variants*, SV). Este resultado indica que los procesos biológicos que determinan los patrones mutacionales somáticos están asociados con la diferenciación celular, y por lo tanto pueden ser usados para identificar tipos tumorales.

BIBLIOGRAFÍA

Akdemir, Kadir C., Victoria T. Le, Justin M. Kim, Sarah Killcoyne, Devin A. King, Ya-Ping Lin, Yanyan Tian, et al. 2020. “Somatic Mutation Distributions in Cancer Genomes Vary with Three-Dimensional Chromatin Structure.” *Nature Genetics*, October, 1–11. https://doi.org/10.1038/s41588-020-0708-0.

Alzahani, Salha M., Afnan Althopity, Ashwag Alghamdi, Boushra Alshehri, and Suheer Aljuaid. 2015. “An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction.” *Lecture Notes on Information Theory* 2 (4). https://doi.org/10.12720/lnit.2.4.310-315.

Birkbak, Nicolai Juul, Bose Kochupurakkal, Jose M. G. Izarzugaza, Aron C. Eklund, Yang Li, Joyce Liu, Zoltan Szallasi, et al. 2013. “Tumor Mutation Burden Forecasts Outcome in Ovarian Cancer with BRCA1 or BRCA2 Mutations.” *PLoS ONE* 8 (11). https://doi.org/10.1371/journal.pone.0080023.

Bramer, Max. 2007. *Principles of Data Mining*. Undergraduate Topics in Computer Science. London: Springer-Verlag. https://doi.org/10.1007/978-1-84628-766-4.

“Cancer.” n.d. Accessed October 28, 2020. https://www.who.int/westernpacific/health-topics/cancer.

Cheng, Feixiong, Junfei Zhao, and Zhongming Zhao. 2016. “Advances in Computational Approaches for Prioritizing Driver Mutations and Significantly Mutated Genes in Cancer Genomes.” *Briefings in Bioinformatics* 17 (4): 642–56. https://doi.org/10.1093/bib/bbv068.

Choi, M., H. Kadara, J. Zhang, E. R. Parra, J. Rodriguez-Canales, S. G. Gaffney, Z. Zhao, et al. 2017. “Mutation Profiles in Early-Stage Lung Squamous Cell Carcinoma with Clinical Follow-up and Correlation with Markers of Immune Function.” *Annals of Oncology* 28 (1): 83–89. https://doi.org/10.1093/annonc/mdw437.

Damodaran, Senthilkumar, MD, PhD, Michael F. Berger, PhD, and Sameek Roychowdhury, MD, and PhD. n.d. “Clinical Tumor Sequencing: Opportunities and Challenges for Precision Cancer Medicine.” *Journal of Clinical Oncology*. Accessed April 10, 2017. http://meetinglibrary.asco.org/content/11500175-156.

Dimitrakopoulos, Christos M., and Niko Beerenwinkel. 2017. “Computational Approaches for the Identification of Cancer Genes and Pathways.” *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 9 (1). https://doi.org/10.1002/wsbm.1364.

Fischer, Andrej, Christopher JR Illingworth, Peter J Campbell, and Ville Mustonen. 2013. “EMu: Probabilistic Inference of Mutational Processes and Their Localization in the Cancer Genome.” *Genome Biology* 14 (4): R39. https://doi.org/10.1186/gb-2013-14-4-r39.

Gao, Jianjiong, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, et al. 2013. “Integrative Analysis of

Complex Cancer Genomics and Clinical Profiles Using the CBioPortal." *Science Signaling* 6 (269): pl1. https://doi.org/10.1126/scisignal.2004088.

Garraway, Levi A., and Eric S. Lander. 2013. "Lessons from the Cancer Genome." *Cell* 153 (1): 17–37. https://doi.org/10.1016/j.cell.2013.03.002.

Greenman, Christopher, Philip Stephens, Raffaella Smith, Gillian L. Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, et al. 2007. "Patterns of Somatic Mutation in Human Cancer Genomes." *Nature* 446 (7132): 153–58. https://doi.org/10.1038/nature05610.

Heo, Seong Gu, Youngil Koh, Jong Kwang Kim, Jongsun Jung, Hyung-Lae Kim, Sung-Soo Yoon, and Ji Wan Park. 2017. "Identification of Somatic Mutations Using Whole-Exome Sequencing in Korean Patients with Acute Myeloid Leukemia." *BMC Medical Genetics* 18 (March). https://doi.org/10.1186/s12881-017-0382-y.

Hofree, Matan, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. 2013. "Network-Based Stratification of Tumor Mutations." *Nature Methods* 10 (11): 1108–15. https://doi.org/10.1038/nmeth.2651.

Horak, Peter, Stefan Fröhling, and Hanno Glimm. 2016. "Integrating Next-Generation Sequencing into Clinical Oncology: Strategies, Promises and Pitfalls." *ESMO Open* 1 (5). https://doi.org/10.1136/esmoopen-2016-000094.

Huret, Jean-Loup, Philippe Dessen, and Alain Bernheim. 2001. "Atlas of Genetics and Cytogenetics in Oncology and Haematology, Updated." *Nucleic Acids Research* 29 (1): 303–304. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC29834/.

Jacob, Louis, Moritz Freyn, Matthias Kalder, Konstantinos Dinas, and Karel Kostev. 2018. "Impact of Tobacco Smoking on the Risk of Developing 25 Different Cancers in the UK: A Retrospective Study of 422,010 Patients Followed for up to 30 Years." *Oncotarget* 9 (25): 17420–29. https://doi.org/10.18632/oncotarget.24724.

Kim, Yoo-Ah, Dong-Yeon Cho, Phuong Dao, and Teresa M. Przytycka. 2015. "MEMCover: Integrated Analysis of Mutual Exclusivity and Functional Network Reveals Dysregulated Pathways across Multiple Cancer Types." *Bioinformatics* 31 (12): i284–i292. https://doi.org/10.1093/bioinformatics/btv247.

Kou, Tadayuki, Masashi Kanai, Shigemi Matsumoto, Yasushi Okuno, and Manabu Muto. 2016. "The Possibility of Clinical Sequencing in the Management of Cancer." *Japanese Journal of Clinical Oncology* 46 (5): 399–406. https://doi.org/10.1093/jjco/hyw018.

Lee, Choong Ho, and Hyung-Jin Yoon. 2017. "Medical Big Data: Promise and Challenges." *Kidney Research and Clinical Practice* 36 (1): 3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3.

Lee, Ju-Seog. 2016. "Exploring Cancer Genomic Data from the Cancer Genome Atlas Project." *BMB Reports* 49 (11): 607–11. https://doi.org/10.5483/BMBRep.2016.49.11.145.

Lefebvre, Celine, Thomas Bachelot, Thomas Filleron, Marion Pedrero, Mario

Campone, Jean-Charles Soria, Christophe Massard, et al. 2016. “Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis.” *PLoS Medicine* 13 (12). https://doi.org/10.1371/journal.pmed.1002201.

Li, Shiyong, Yoon-La Choi, Zhuolin Gong, Xiao Liu, Maruja Lira, Zhengyan Kan, Ensel Oh, et al. 2016. “Comprehensive Characterization of Oncogenic Drivers in Asian Lung Adenocarcinoma.” *Journal of Thoracic Oncology* 11 (12): 2129–40. https://doi.org/10.1016/j.jtho.2016.08.142.

Libbrecht, Maxwell W., and William Stafford Noble. 2015. “Machine Learning Applications in Genetics and Genomics.” *Nature Reviews Genetics* 16 (6): 321–332. https://doi.org/10.1038/nrg3920.

Lindquist, Karla J., Pamela L. Paris, Thomas J. Hoffmann, Niall J. Cardin, Rémi Kazma, Joel A. Mefford, Jeffrey P. Simko, et al. 2016. “Mutational Landscape of Aggressive Prostate Tumors in African American Men.” *Cancer Research* 76 (7): 1860–68. https://doi.org/10.1158/0008-5472.CAN-15-1787.

Loyola-González, Octavio, Milton García-Borroto, Miguel Angel Medina-Pérez, José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and Guillermo De Ita. 2013. “An Empirical Study of Oversampling and Undersampling Methods for LCMine an Emerging Pattern Based Classifier.” In *Pattern Recognition*, edited by Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad, Joaquín Salas Rodríguez, and Gabriella Sanniti di Baja, 264–273. Berlin, Heidelberg: Springer Berlin Heidelberg.

Lynch, Chip M., Behnaz Abdollahi, Joshua D. Fuqua, Alexandra R. de Carlo, James A. Bartholomai, Rayeanne N. Balgemann, Victor H. van Berkel, and Hermann B. Frieboes. 2017. “Prediction of Lung Cancer Patient Survival via Supervised Machine Learning Classification Techniques.” *International Journal of Medical Informatics* 108 (December): 1–8. https://doi.org/10.1016/j.ijmedinf.2017.09.013.

Malouf, Gabriel G., Siraj M. Ali, Kai Wang, Sohail Balasubramanian, Jeffrey S. Ross, Vincent A. Miller, Philip J. Stephens, et al. 2016. “Genomic Characterization of Renal Cell Carcinoma with Sarcomatoid Dedifferentiation Pinpoints Recurrent Genomic Alterations.” *European Urology* 70 (2): 348–357. https://doi.org/10.1016/j.eururo.2016.01.051.

Nik-Zainal, Serena. 2014. “Insights into Cancer Biology through Next-Generation Sequencing.” *Clinical Medicine* 14 (Suppl 6): s71–77. https://doi.org/10.7861/clinmedicine.14-6-s71.

Nik-Zainal, Serena, Ludmil B. Alexandrov, David C. Wedge, Peter Van Loo, Christopher D. Greenman, Keiran Raine, David Jones, et al. 2012. “Mutational Processes Molding the Genomes of 21 Breast Cancers.” *Cell* 149 (5–10): 979–93. https://doi.org/10.1016/j.cell.2012.04.024.

Pereira, Bernard, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, et al. 2016. “The Somatic Mutation Profiles of 2,433 Breast Cancers Refine Their Genomic and Transcriptomic Landscapes.” *Nature Communications* 7 (May): 11479. https://doi.org/10.1038/ncomms11479.

Riazalhosseini, Yasser, and Mark Lathrop. 2016. “Precision Medicine from the Renal Cancer Genome.” *Nature Reviews Nephrology* 12. https://doi.org/10.1038/nrneph.2016.133.

Vural, Suleyman, Xiaosheng Wang, and Chittibabu Guda. 2016. “Classification of Breast Cancer Patients Using Somatic Mutation Profiles and Machine Learning Approaches.” *BMC Systems Biology* 10 (Suppl 3). https://doi.org/10.1186/s12918-016-0306-z.

Welcome \textbar ICGC Data Portal. n.d. Accessed May 10, 2017. https://dcc.icgc.org/.

Wendl, Michael C., John W. Wallis, Ling Lin, Cyriac Kandoth, Elaine R. Mardis, Richard K. Wilson, and Li Ding. 2011. “PathScan: A Tool for Discerning Mutational Significance in Groups of Putative Cancer Genes.” *Bioinformatics* 27 (12): 1595–1602. https://doi.org/10.1093/bioinformatics/btr193.

Yang, William, Kenji Yoshigoe, Xiang Qin, Jun S Liu, Jack Y Yang, Andrzej Niemierko, Youping Deng, et al. 2014. “Identification of Genes and Pathways Involved in Kidney Renal Clear Cell Carcinoma.” *BMC Bioinformatics* 15 (Suppl 17): S2. https://doi.org/10.1186/1471-2105-15-S17-S2.

Yang, Yadong, Xunong Dong, Bingbing Xie, Nan Ding, Juan Chen, Yongjun Li, Qian Zhang, Hongzhu Qu, and Xiangdong Fang. 2015. “Databases and Web Tools for Cancer Genomics Study.” *Genomics, Proteomics & Bioinformatics* 13 (1): 46–50. https://doi.org/10.1016/j.gpb.2015.01.005.

Zhang, Fan, Chunyan Ren, Kwun Kit Lau, Zihan Zheng, Geming Lu, Zhengzi Yi, Yongzhong Zhao, et al. 2016. “A Network Medicine Approach to Build a Comprehensive Atlas for the Prognosis of Human Cancer.” *Briefings in Bioinformatics* 17 (6): 1044–59. https://doi.org/10.1093/bib/bbw076.

Zhang, Wensheng, Andrea Edwards, Erik K Flemington, and Kun Zhang. 2017. “Significant Prognostic Features and Patterns of Somatic TP53 Mutations in Human Cancers.” *Cancer Informatics* 16 (February). https://doi.org/10.1177/1176935117691267.

Zutter, Mary M., Kenneth J. Bloom, Liang Cheng, Ian S. Hagemann, Jill H. Kaufman, Alyssa M. Krasinskas, Alexander J. Lazar, et al. 2014. “The Cancer Genomics Resource List 2014.” *Archives of Pathology & Laboratory Medicine* 139 (8): 989–1008. https://doi.org/10.5858/arpa.2014-0330-CP.

GLOSARIO

NGS Secuenciación de Nueva Generación, *Next Generation Sequencing*
SNV Variaciones de nucleótido único, *Single Nucleotide Variants*
SV Variaciones Estructurales, *Structural Variants*
TML Carga Mutacional Total, Total Mutational Load
WGS Genoma completo, *Whole-Genome Sequencing*
WES Exoma completo, *Whole-Exome Sequencing*
TCGA Atlas del genoma del cáncer, *The Cancer Genome Atlas*
ICGC Consorcio Internacional de la genómica del cáncer, *International Cancer Genomics Consortium*
CCLE Enciclopedia de línea celular en cáncer, *Cancer Cell Line Encyclopedia*
TARGET *Therapeutically Applicable Research to Generate Effective Treatments*
LUAD Adenocarcinoma de pulmón, Lung Adenocarcinoma
LMA Leucemia Mieloide Aguda, Acute Myeloid Leukemia
LUSC Carcinoma de Células Escamosas de Pulmón, Lung Squamous Cell Carcinoma
HNSC Carcinoma de Células Escamosas de Cabeza y Cuello, Head and Neck Squamous Cell Carcinoma
OS Sobrevida general, *Overall Survival*,
DFS Tiempo libre de la enfermedad, *Disease-free survival*
TMB Carga Mutacional Total, Total Mutational Burden
NSCLC Carcinoma de Células no pequeñas, non-small-cell lung carcinoma