Martínez Pérez, Elizabeth

Patrones mutacionales en cáncer

Elizabeth Martínez Pérez, Repositorio Institucional Digital de Acceso Abierto,
Marzo de 2021, pp. 148,
http://ridaa.unq.edu.ar,
Universidad Nacional de Quilmes, Secretaría de Posgrado,
Doctorado en Ciencia y Tecnología

Patrones Mutacionales en Cáncer

TESIS DOCTORAL

Elizabeth Martínez Pérez

emartinezperez1990@gmail.com

Resumen

El cáncer es uno de los problemas de salud más importantes a nivel mundial y, a pesar de avances recientes en diagnóstico y terapia, la mortalidad sigue siendo inaceptablemente alta. Es la segunda causa de muerte con una estimación global de 14 millones de nuevos casos y 8.2 millones de muertes cada año. Es una enfermedad de evolución clonal, donde una célula adquiere mutaciones y/o alteraciones epigenéticas que la vuelven más apta. El resultado de un tratamiento está fuertemente determinado por las alteraciones en su genoma; de ahí la idea de las terapias personalizadas dirigidas a una o múltiples dianas moleculares. La terapia inmunológica ha demostrado buenos resultados en el tratamiento de algunos tumores, sin embargo solo un 35% de los pacientes se benefician de esta por lo que es necesario encontrar mejores marcadores para predecir su efectividad.

El principal objetivo de esta tesis ha sido encontrar patrones entre las mutaciones en los pacientes con cáncer utilizando técnicas bioinformáticas y el análisis de datos públicamente disponibles. Hemos abordado tres temas, las relaciones entre las mutaciones por su utilidad en las terapias dirigidas, las mutaciones que generan resistencia a medicamentos y la carga mutacional total (TMB) como biomarcador de pacientes con beneficio clínico ante la inmunoterapia.

Como resultados se obtuvieron relaciones entre mutaciones con significación estadística, algunas de ellas son excluyentes, mientras que otras son comutaciones.

Tanto el tipo de relación como las mutaciones implicadas en ella son importantes al momento de seleccionar una terapia multi dirigida. Se encontraron 3 características de las mutaciones drivers (o sea causantes de la enfermedad), lo que nos permite sugerir mutaciones como drivers que aún no son consideradas como tal. Además, se determinó que la mayoría de las mutaciones de resistencia a drogas ocurren cerca del sitio de unión a la misma y en el caso de las quinasas también en el lazo de activación. Por último, se definieron paneles de genes tumor-específicos para predecir con mayor precisión el TMB. Estos paneles son más económicos que los paneles comerciales (Foundation-One y MSK-Impact) ya que requieren secuenciar menos pares de bases. Además contribuimos al estudio de proteínas desordenadas (o con regiones desordenadas) durante estadías en el exterior como colaboraciones.

Palabras Claves: cáncer, mutaciones, patrones mutacionales, terapias dirigidas, multiterapia, inmunoterapia, carga mutacional total, proteínas desordenadas.

Abstract

Cancer is one of the most important health problems in the world, and despite recent advances in diagnosis and therapy, mortality remains unacceptably high. It is the second leading cause of death with a global estimate of 14 million new cases and 8.2 million deaths each year. Cancer is a disease of clonal evolution, where cells acquire mutations and / or epigenetic alterations that give them adaptive advantages with respect to the normal cells. The goal of a treatment is strongly determined by the alterations in his/her genome; hence the idea of personalized therapies directed at one or multiple molecular targets. Immunotherapy has shown good results in the treatment of some tumors, however only 35% of patients benefit from it, so it is necessary to find better markers to predict its effectiveness. The main goal of this thesis has been to find patterns between mutations in cancer patients using bioinformatic techniques and the analysis of public available data. We have approached three problems: the relationships between mutations due to their usefulness in designing targeted therapies; the study of mutations that generate drug resistance; and the prediction of the total mutational burden (TMB) as a biomarker of patients immunotherapy benefit.

Our results show a relationship between mutations with statistical significance some of them are mutually exclusive, while others co-occur. Knowing both, the type of relationship and the mutations are important to select a multi-targeted therapy. We found 3 characteristics of driver mutations that allow us to suggest new mutations as drivers not considered as such yet. By other hand, we determined that most of the drug resistance mutations occur near their binding site and, in the case of kinases also in the activation loop. Finally, tumor-specific gene panels were defined to accurately predict TMB. These panels are cheaper than currently used commercial panels (Foundation-One and MSK-Impact) as they require fewer base pairs to be sequenced. Also, we contribute to the study of disordered proteins (or with disordered regions) during my stays abroad in collaborative works.

Keywords: cancer, mutations, mutational patterns, target therapy, multi-targeted therapy, immunotherapy, total mutational burden, disordered proteins.

Universidad Nacional de Quilmes

Tesis en opción del título

Doctora en Ciencia y Tecnología

Patrones Mutacionales en Cáncer

Aspirante: Lic. Elizabeth Martínez Pérez[1,2]

Directora: Dra. Cristina E. Marino Buslje[1]

Co-director: Dr. Gustavo Daniel Parisi[2]

Consejera de estudios: Dra. Silvina Fornasari[2]

Lugar de Trabajo:

[1] Laboratorio de biología estructural. Fundación Instituto Leloir (FIL)

[2] Departamento de Ciencia y Tecnología. Universidad Nacional de Quilmes (UNQ)

Buenos Aires, 1 de marzo del 2021

Publicaciones relacionadas al tema de la tesis

- Panels and models for accurate prediction of tumor mutation burden in tumor samples. **Elizabeth Martínez-Pérez**, Miguel Angel Molina-Vila & Cristina Marino-Buslje. *En proceso de publicación*.
- Comutation and exclusion analysis in human tumors: A tool for cancer biology studies and for rational selection of multitargeted therapeutic approaches. Human Mutation. Volume 40, Issue 4. Pages 413-425. 01/2019. Soledad Ochoa*, **Elizabeth Martínez-Pérez***, Diego Javier Zea, Miguel Angel Molina-Vila & Cristina Marino-Buslje. (*) Co-primer autor. https://doi.org/10.1002/humu.23705.

Publicaciones de colaboraciones

- Disprot expanded: homology transfer, from Disprot proteins to ortholog proteins. **Elizabeth Martínez-Pérez**, Mátyás Pajkos, Silvio Tosatto, Toby J. Gibson, Zsuzsanna Dosztanyi & Cristina Marino-Buslje. *En proceso de publicación*.
- PED 4.0: a comprehensive extension and update of the database of disordered protein ensembles. *Aceptada en Nucleic Acids Research*. 01/2021. Tamas Lazar, **Elizabeth Martínez-Pérez**, Federica Quaglia, András Hato, et. al.
- Short linear motif candidates in the cell entry system used by SARS-CoV-2 and their potential therapeutic implications. *Aceptada en Science Signaling*. 01/2021. Bálint Mészáros, Hugo Sámano-Sánchez, Jesús Alvarado-Valverde, Jelena Čalyševa, **Elizabeth Martínez-Pérez**, Renato Alves, Manjeet Kumar, Friedrich Rippmann, Lucía B. Chemes and Toby J. Gibson.
- DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Research. Volume 48, Issue D1. Pages D269–D276. 01/2020. András Hatos, et.al. **Mi nombre es el 34 de 64 autores.** https://doi.org/10.1093/nar/gkz975.

Eventos

- 1st Congress of Women in Bioinformatics and Data Science LA. Argentina. 09/2020. Oral y Poster: Models to predict Total Mutational Burden in Cancer.
- 10th Congreso Argentino de Bioinformática y Biología Computacional (CAB2C). Argentina. 11/2019. Oral y Poster: Models to predict Total Mutational Burden in Cancer.
- 9th CAB2C. Argentina. 11/2018. Poster: Models to predict Mutational Burden in Cancer.

- 8th CAB2C. Argentina. 11/2017. Oral y Poster: Landscape of Drug Resistance Mutations in Cancer.
- Oncology Conference. Argentina. 10/2017.
- 7th CAB2C. Argentina. 11/2016. Poster: Towards a better cancer classification: mutational patterns of loci and cancer types.

Becas de Investigación en el exterior

- Finalización de los proyectos: Transferencia por homología y Servidor PED. IDPfun. EMBL, Heidelberg, Alemania. 06/2020 al 10/2020.
- Servidor estructuras de proteínas desordenadas (PED). IDPfun. EMBL, Heidelberg, Alemania. 12/2019 al 05/2020.
- Transferencia por homología en proteínas desordenadas. IDPfun. EMBL, Heidelberg, Alemania. 09/2019 al 11/2019.
- Ontología de proteínas desordenadas. IDPfun. EMBL, Heidelberg, Alemania. 09/2018 al 12/2018.

Becas de Estudio en Argentina

- Beca de Finalización de Doctorado. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). 04/2019 al 03/2021
- Beca Inicial de Doctorado. Agencia Nacional de promoción Científica y Tecnológica (AGENCIA). 07/2016 al 03/2019
- Beca Total para pago de aranceles de la carrera de doctorado. Universidad nacional de Quilmes (UNQ). 06/2016 al 03/2021

Cursos extracurriculares

- Proteínas desordenadas. IDPfun. Brixen. Italia. 01/2020
- Escuela de modelado de biomoléculas. CELFI. UBA. Argentina. 08/2018

Actividades extracurriculares

- Escuelita de verano, bioinformática. Docente y apoyo. FIL. Argentina. 2019.
- Escuelita de verano, bioinformática. Docente y apoyo. FIL. Argentina. 2018.
- Noche de los Museos. FIL. Argentina. 2017
- Escuelita de verano, bioinformática. Docente y apoyo. FIL. Argentina. 2017.
- Noche de los Museos. FIL. Argentina. 2016.

A **Cristina Marino Buslje**, directora y amiga. Siempre ha estado conmigo durante toda la tesis brindando apoyo, ayuda y conocimiento tanto en lo científico como en lo personal. Le doy mil gracias por haberme dado la oportunidad de hacer el doctorado con ella en Buenos Aires, en un instituto tan prestigioso como la Fundación Instituto Leloir (FIL). Me ha guiado durante el doctorado, pero nunca ha impuesto un objetivo, me ha dejado las riendas de mi investigación para poder elegir y decidir con el objetivo de que yo crezca tanto en el ámbito profesional como personal. No tengo forma de agradecerle todo lo que ha hecho para que yo tuviera las mejores condiciones, incluyendo trámites de migraciones y vivienda. Ella y su familia siempre nos han tenido presente en cada celebración de navidad y fin de año.

A **Gustavo Parisi**, co-director. Fue la chispa que nos ayudó a inscribir el doctorado en la UNQ, tras recibir muchos traspiés de la UBA, todo sin conocerme y antes de yo ingresar a la Argentina. Su ayuda fue vital para cumplir uno de los requisitos que el cónsul argentino en Cuba solicitaba para otorgar la visa de estudiante.

A **Silvina Fornasari**, consejera de estudios. Siempre presente y ayudando con cada presentación de avance de tesis.

A mi comité de seguimiento de mi tesis en la FIL, personas que durante los 5 años de la tesis han seguido el avance de mi tesis de forma incondicional. **Ignacio Ponzoni**, que cada año viaja desde Bahía Blanca a Buenos Aires solo a escucharme sobre mi avance de tesis y aportar nuevas ideas. Y **Mónica Castro** que aunque teniendo un bebe muy chico y trabajar en el Instituto de Oncología Ángel H. Roffo ha seguido mi tesis sin descanso año tras año. **Andrea Gamarnik** que aunque no se desarrolla directamente en el tema de mi tesis siempre ha estado presente. A los tres les agradezco la infinidad de ayudas y debates para mejorar la calidad de mi investigación.

A las **secretarías de posgrado, egreso y títulos de la UNQ**, por cada año realizar los trámites necesarios para renovar mi permiso de residencia como estudiante en la Argentina. Además de estar presentes en cada duda o solicitud relacionada con el doctorado en la UNQ.

A los **compañeros de grupo**, Elin Teppa, Diego Zea, Franco Simonetti, Javier Iserte, Soledad Ochoa, Fernando Orti y Álvaro Navarro. Los cuales siempre me han apoyado y ayudado a resolver mis problemas e inquietudes tanto científicas como personales. También con los que compartimos laboratorio pero son de otro grupo, Ariel Chernomoretz, Andrés Rabinovich, Maximiliano Beckel y Bruno Kaufman.

Agradecimientos

A **todas las personas con las que he colaborado**, con los cuales he obtenido lindos trabajos de investigación: Miguel Angel Molina Vila, Silvio C.E. Tosatto, Peter Tompa, Zsuzsanna Dosztanyi, Támás Lazar, Mátyás Pajkos, András Hatos, Federica Quaglia, Lucia Beatriz Chemes, Toby J. Gibson, entre otros.

Al consorcio internacional **IDPfun**, que bajo el proyecto de intercambio de personal de investigación e innovación Marie Skłodowska-Curie financiado por H2020 (Marie-Skłodowska-Curie RISE Nº 778247) ha brindado a tres laboratorios de argentina, incluyendo el mío a tener intercambios científicos con otros 5 países europeos. Por este proyecto he tenido la oportunidad de participar en colaboraciones con otros laboratorios.

A **Toby J. Gibson y su grupo** en el Laboratorio Europeo de Biología Molecular en la ciudad de Heidelberg Alemania que me ha acogido en todas las colaboraciones del proyecto IDPfun. Y luego durante la pandemia de COVID-19 que no pude volar por el cierre de las fronteras me acogió como una integrante más de su laboratorio. Y le agradezco muchísimo por toda su ayuda tanto profesional como personal durante todas mis estadías.

A mis **amigos de Cuba** que me han ayudado en muchas situaciones, Maricel Meneses Gómez, Yudith Cañizares Carmenate y Juan Alberto Castillo Garit.

A mi familia de los cuales estuve lejos, pero recibí mucho apoyo para superar todas las adversidades. Especial agradecimiento a mis padres **Orlando Amado Martínez Baños** y **Mirta Odalys Pérez Anoceto**, mi hermana **Yenisbel Martínez Pérez** y tía **Ana María Martínez Baños** que siempre me han dado los mejores consejos para seguir adelante y también han sido los más sacrificados debido a la distancia. Mis padres me enseñaron a ser independiente y confiar en mis ideas y conocimientos, siempre haciendo caso a una frase de mi papá "*Busca el por qué de las cosas"*.

A mi esposo, **Maykel Torres Lorenzo** que ha estado a mi lado incondicionalmente por estos 5 años de estudio en el extranjero, sacrificando su beneficio personal por el bienestar familiar. Es más que esposo, es un gran amigo.

A **toda mi familia y la de mi esposo** que siempre me han apoyado y han estado presentes durante este período.

La tesis se enmarca en el período de julio 2016 a marzo 2021. Primero tuve una beca inicial de doctorado de AGENCIA por 3 años y luego una de finalización de doctorado del CONICET por 2 años. Además conté con una beca total de la UNQ para el pago de los aranceles relacionados al doctorado.

El tema de la tesis es “Patrones mutacionales en cáncer”, lo que conlleva un fuerte estudio de bioinformática y de conocimientos sobre proteínas y la enfermedad. La tesis se encuentra estructurada en 10 capítulos que describiremos a continuación.

La **Introducción** aborda el cáncer como enfermedad de alto impacto globalmente, sus tratamientos y la resistencia a medicamentos que se presentan después de los tratamientos. También se habla de 3 bases de datos de muestras de cáncer. Por último presenta el objetivo general y los objetivos específicos de la tesis.

Los **Materiales y métodos**, son generales a todos los trabajos realizados durante la tesis, pues cada trabajo en particular cuenta con materiales y métodos específicos Aborda temas como lenguajes de programación para bioinformática, pruebas estadísticas y análisis de datos más utilizados durante la tesis.

La **Dependencias entre mutaciones**, presenta el primer estudio realizado durante la tesis que fue en conjunto con Soledad Ochoa para su tesis de maestría. Este estudio caracteriza las mutaciones en cáncer atendiendo a sus dependencias como exclusiones y co-mutaciones. Este trabajo se encuentra publicado en la revista *Human Mutations* (https://doi.org/10.1002/humu.23705).

La **Resistencia adquirida a medicamentos**, presenta el segundo estudio realizado durante la tesis. Se analizan las relación entre mutaciones relacionadas a resistencia a medicamentos. Si bien es un estudio corto y que no presentó grandes resultados, puede servir de apoyo a este tipo de análisis.

La **Carga mutacional total**, presenta el tercer estudio y último relacionado al tema de la tesis. Se enfoca en obtener un conjunto de genes específicos por tipo de cáncer que permitan, sumado a un modelo matemático, predecir la carga mutacional total de una muestra de cáncer. Se proponen modelos que permiten este cálculo en 14 tipos de cáncer. Este estudio se encuentra en proceso de publicación.

En lo que se refiere a **Colaboraciones**, se presentan 4 trabajos en los cuales colaboré durante mis estadías de investigación en el exterior. El primero de ellos relacionado a la **ontología de proteínas desordenadas**. Dicha ontología se encuentra integrada a la base de datos de anotaciones proteínas desordenadas, llamada DisProt. En la publicación de la versión 8.0 de la base de datos se encuentra la ontología como uno de los aspectos mejorados en la actualización.

Este trabajo se encuentra publicado en la revista *Nucleic Acids Research* (https://doi.org/10.1093/nar/gkz975). El segundo, relacionado a la **transferencia por homología de anotaciones de desorden y términos de ontología** a otras proteínas no presentes en DisProt. Este estudio se enmarca en encontrar esas otras proteínas y seleccionar los parámetros para transferir las anotaciones. Este trabajo se encuentra en proceso de escritura para su posterior publicación. El tercer trabajo se relaciona con la **actualización de la base de datos PED**, base de datos de conformaciones estructurales de proteínas desordenadas. La base de datos contaba con una infraestructura obsoleta y baja cantidad de datos. Por lo que con este trabajo se mejoró la cantidad y calidad de la misma. La actualización de PED a su versión 4.0 se encuentra aceptada para su publicación en enero de 2021 en la revista *Nucleic Acids Research*. El cuarto trabajo se relaciona con **proteínas desordenadas y COVID-19** ya que durante el período de la tesis ocurrió la pandemia. Este trabajo está aceptado para ser publicado en la revista *Science Signaling* en enero de 2021.

Las **Conclusiones**, son generales del tema de la tesis, mostrando los principales resultados de los 3 estudios enmarcados en la tesis.

Las **Referencias bibliográficas**, que contiene todos los artículos, libros, etc, citados y consultados durante la tesis.

Los **Anexos**, muestran datos complementarios para la tesis que son tablas muy largas para escribirla dentro del texto principal.

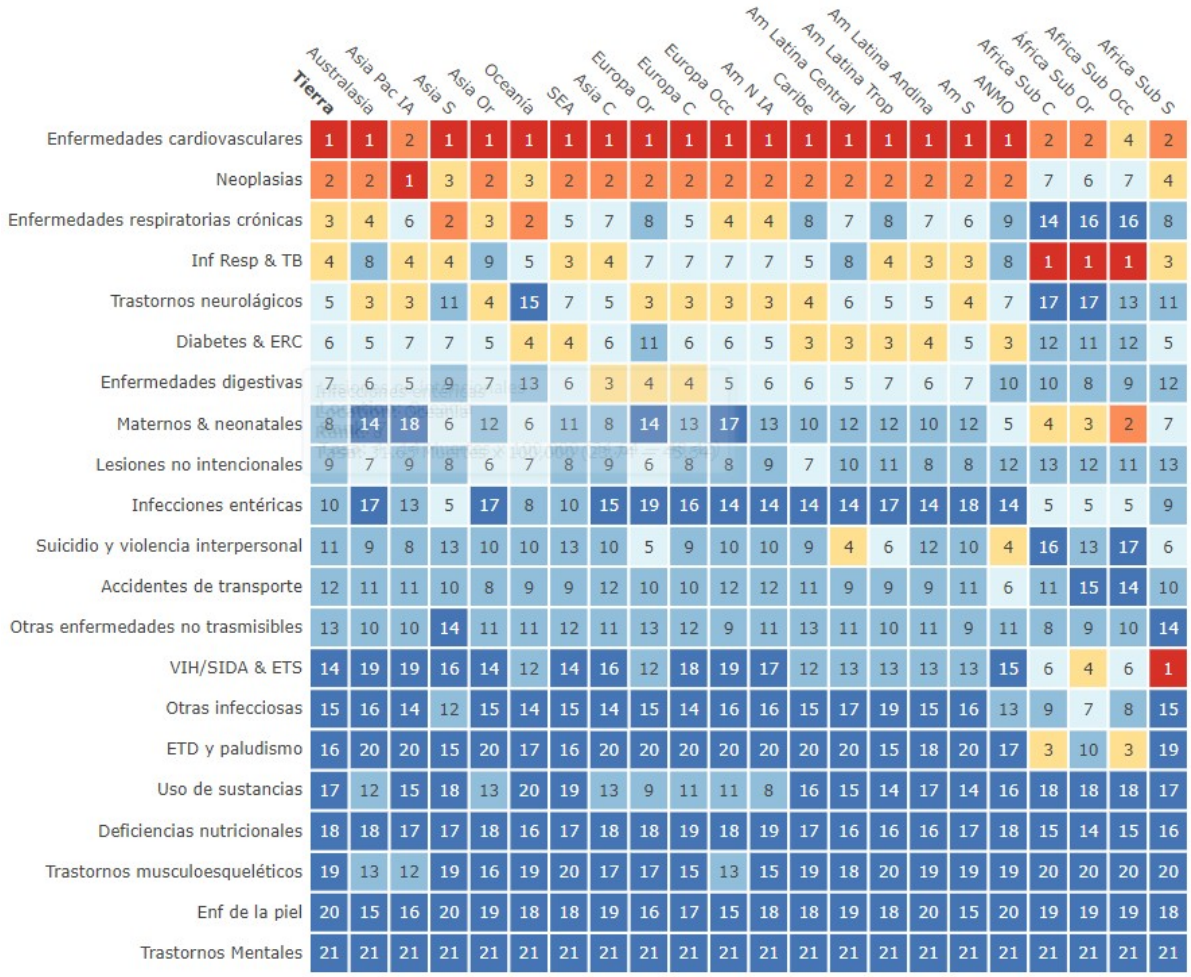
Los **Índices**, como bien indica contiene índices para el acceso fácil, cómo son los de figuras, tablas y ecuaciones, abreviaturas y genes/proteínas mencionada en la tesis.

1. Introducción

El cáncer es una enfermedad de evolución clonal[1], donde una célula adquiere mutaciones y/o alteraciones epigenéticas que la vuelven más apta[2]. Las 6 características distintivas conocidas como "Hallmarks of cancer"[3] le brindan a las células cancerosas la adaptabilidad, estas son: mantenimiento de señalización proliferativa, evasión de supresores de crecimiento, resistencia a la muerte celular, habilitar la inmortalidad replicativa, inducir angiogénesis y la activación de invasión y metástasis. El uso de estas características ayuda a la célula cancerosa a tener identidad propia y comportarse como un organismo invasivo y provocar la muerte del individuo.

1.1. Cáncer una enfermedad de alta incidencia y mortalidad

Entre las principales tareas del Instituto de Métricas y Evaluación de la Salud (IHME) es recaudar la información de las causas de muertes, atendiendo al país, rango de edad, sexo[4]. Presentan una herramienta de visualización de la información de forma fácil e intuitiva[5]. El motivo de la muerte lo desglosan en 4 niveles de agrupación, en el segundo nivel de agrupación aparecen 21 términos de motivos de muerte siendo uno de ellos “neoplasias” que engloba a todos los tipos de cáncer. En los datos del 2017, se reporta que la muerte relacionada al cáncer es la segunda de mayor incidencia a nivel mundial (**Figura 1**), siendo la primera las enfermedades cardiovasculares.

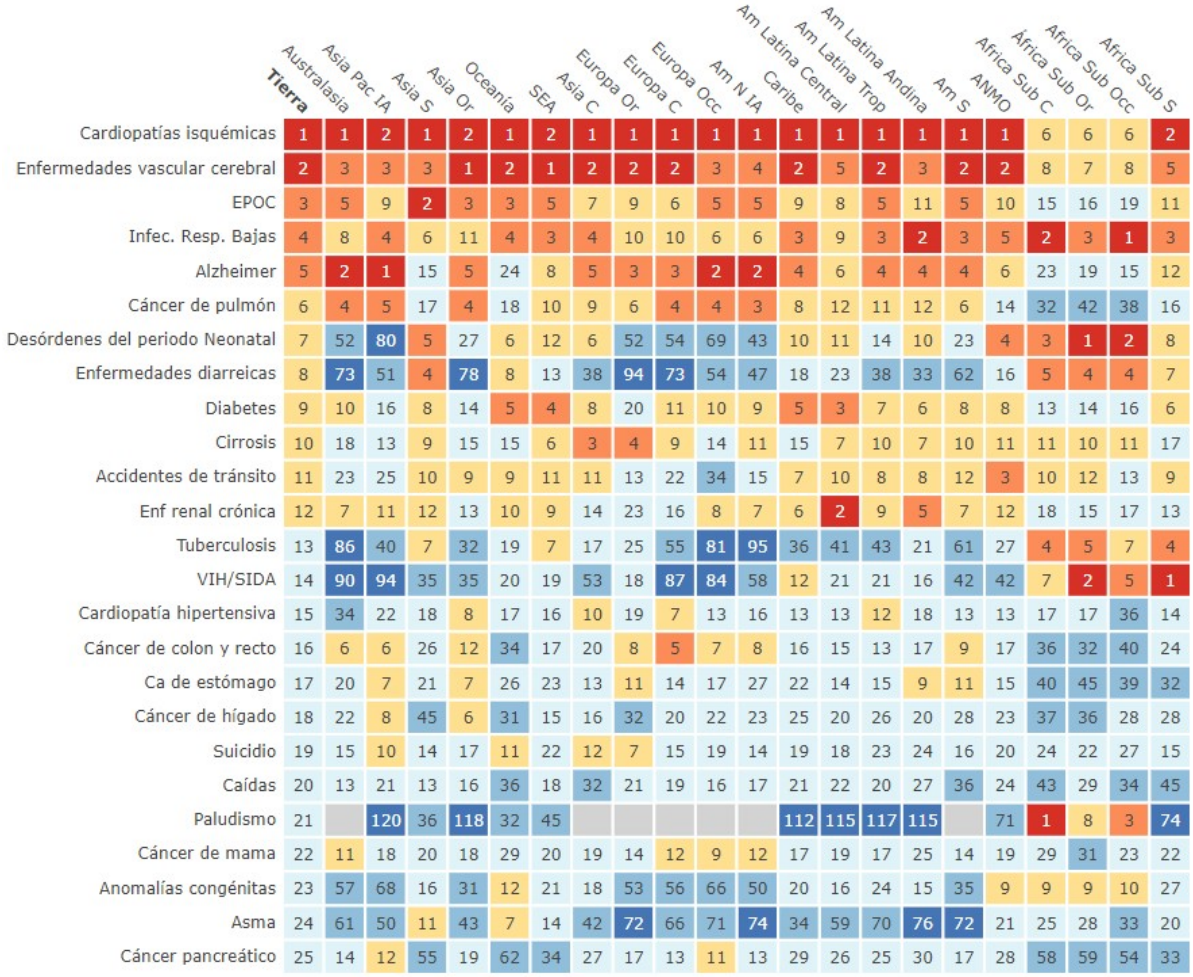


Tierra	Australasia	Asia Pac IA	Asia S	Asia Or	Oceanía	SEA	Asia C	Europa Or	Europa C	Europa Occ	Am N IA	Am Latina Caribe	Am Latina Central	Am Latina Trop	Am Latina Andina	Am S	Africa Sub ANMO	Africa Sub C	Africa Sub Or	Africa Sub Occ	Africa Sub S
Enfermedades cardiovasculares	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	4	2
Neoplasias	2	2	1	3	2	3	2	2	2	2	2	2	2	2	2	2	2	7	6	7	4
Enfermedades respiratorias crónicas	3	4	6	2	3	2	5	7	8	5	4	4	8	7	8	7	6	9	14	16	16	8
Inf Resp & TB	4	8	4	4	9	5	3	4	7	7	7	7	5	8	4	3	3	8	1	1	1	3
Trastornos neurológicos	5	3	3	11	4	15	7	5	3	3	3	3	4	6	5	5	4	7	17	17	13	11
Diabetes & ERC	6	5	7	7	5	4	4	6	11	6	6	5	3	3	3	4	5	3	12	11	12	5
Enfermedades digestivas	7	6	5	9	7	13	6	3	4	4	5	6	6	5	7	6	7	10	10	8	9	12
Maternos & neonatales	8	14	18	6	12	6	11	8	14	13	17	13	10	12	12	10	12	5	4	3	2	7
Lesiones no intencionales	9	7	9	8	6	7	8	9	6	8	8	9	7	10	11	8	8	12	13	12	11	13
Infecciones entéricas	10	17	13	5	17	8	10	15	19	16	14	14	14	14	17	14	18	14	5	5	5	9
Suicidio y violencia interpersonal	11	9	8	13	10	10	13	10	5	9	10	10	9	4	6	12	10	4	16	13	17	6
Accidentes de transporte	12	11	11	10	8	9	9	12	10	10	12	12	11	9	9	9	11	6	11	15	14	10
Otras enfermedades no trasmisibles	13	10	10	14	11	11	12	11	13	12	9	11	13	11	10	11	9	11	8	9	10	14
VIH/SIDA & ETS	14	19	19	16	14	12	14	16	12	18	19	17	12	13	13	13	13	15	6	4	6	1
Otras infecciosas	15	16	14	12	15	14	15	14	15	14	16	16	15	17	19	15	16	13	9	7	8	15
ETD y paludismo	16	20	20	15	20	17	16	20	20	20	20	20	20	20	15	18	20	17	3	10	3	19
Uso de sustancias	17	12	15	18	13	20	19	13	9	11	11	8	16	15	14	17	14	16	18	18	18	17
Deficiencias nutricionales	18	18	17	17	18	16	17	18	18	19	18	19	17	16	16	16	17	18	15	14	15	16
Trastornos musculoesqueléticos	19	13	12	19	16	19	20	17	17	15	13	15	19	18	20	19	19	19	20	20	20	20
Enf de la piel	20	15	16	20	19	18	18	19	16	17	15	18	18	19	18	20	15	20	19	19	19	18
Trastornos Mentales	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21	21

Figura 1. Principales causas de muertes en el mundo en el 2017.

Datos del IHME por el visualizador de GBD[6]. El eje "x" son las regiones, el eje "y" las posibles causas de muerte ordenadas por la cantidad de muertes de mayor a menor. El número en cada casilla del gráfico indica el orden obtenido para la causa de muerte en la región.

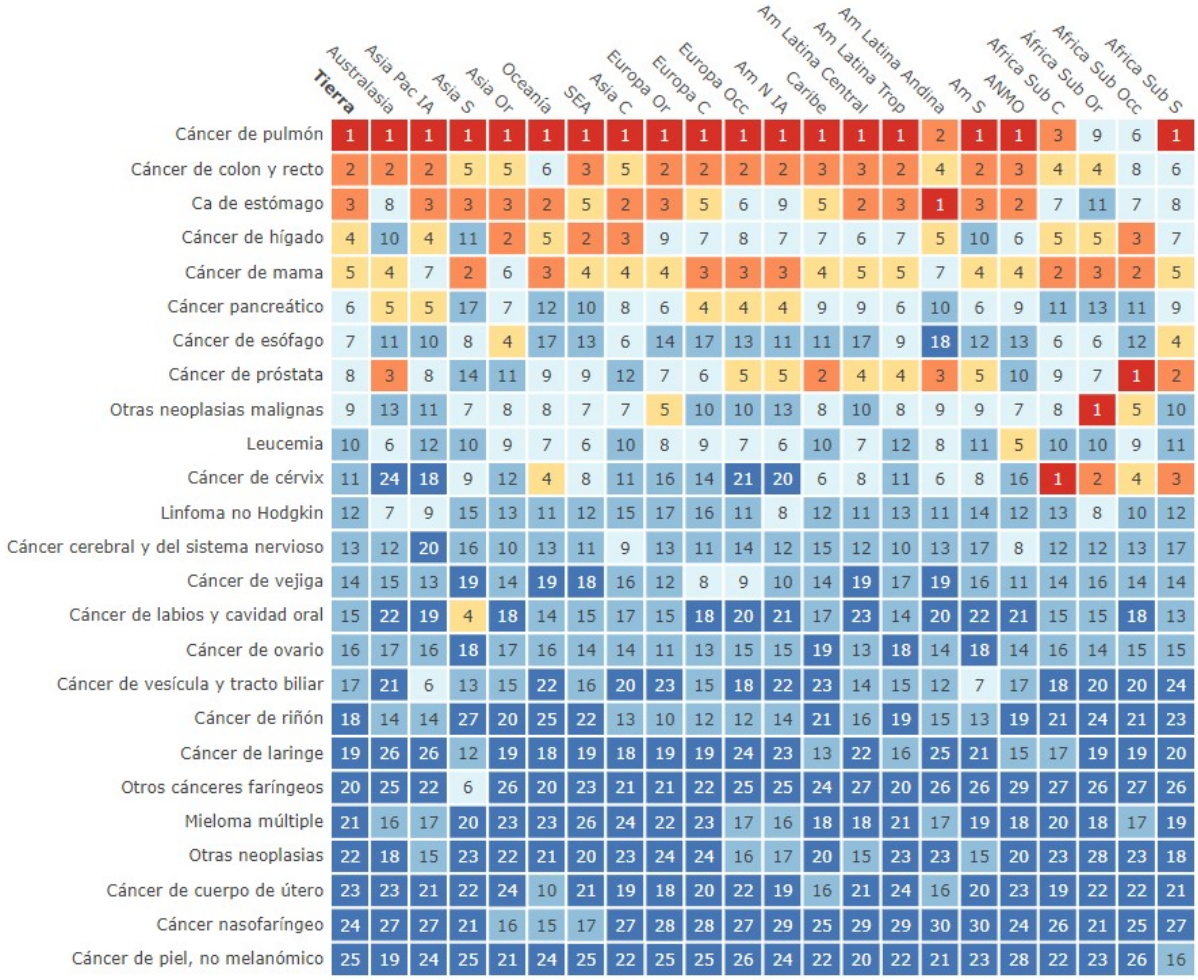
Si bajamos al tercer nivel (**Figura 2**), ya se subdivide neoplasias en los tipos de cáncer y encontramos que el cáncer de pulmón es el sexto de mayor incidencia a nivel mundial, seguido por colorrectal, gástrico, hígado, mama y páncreas en los puestos: 16, 17, 18, 22 y 25 respectivamente.



Tierra	Australasia	Asia Pac IA	Asia S	Asia Or	Oceanía	SEA	Asia C	Europa Or	Europa C	Europa Occ	Am N IA	Am Latina Caribe	Am Latina Central	Am Latina Trop	Am Latina Andina	Am S	ANMO	África Sub C	África Sub Or	África Sub Occ	África Sub S
Cardiopatías isquémicas	1	1	2	1	2	1	2	1	1	1	1	1	1	1	1	1	1	6	6	6	2
Enfermedades vascular cerebral	2	3	3	3	1	2	1	2	2	2	3	4	2	5	2	3	2	2	8	7	8	5
EPOC	3	5	9	2	3	3	5	7	9	6	5	5	9	8	5	11	5	10	15	16	19	11
Infec. Resp. Bajas	4	8	4	6	11	4	3	4	10	10	6	6	3	9	3	2	3	5	2	3	1	3
Alzheimer	5	2	1	15	5	24	8	5	3	3	2	2	4	6	4	4	4	6	23	19	15	12
Cáncer de pulmón	6	4	5	17	4	18	10	9	6	4	4	3	8	12	11	12	6	14	32	42	38	16
Desórdenes del periodo Neonatal	7	52	80	5	27	6	12	6	52	54	69	43	10	11	14	10	23	4	3	1	2	8
Enfermedades diarreicas	8	73	51	4	78	8	13	38	94	73	54	47	18	23	38	33	62	16	5	4	4	7
Diabetes	9	10	16	8	14	5	4	8	20	11	10	9	5	3	7	6	8	8	13	14	16	6
Cirrosis	10	18	13	9	15	15	6	3	4	9	14	11	15	7	10	7	10	11	11	10	11	17
Accidentes de tránsito	11	23	25	10	9	9	11	11	13	22	34	15	7	10	8	8	12	3	10	12	13	9
Enf renal crónica	12	7	11	12	13	10	9	14	23	16	8	7	6	2	9	5	7	12	18	15	17	13
Tuberculosis	13	86	40	7	32	19	7	17	25	55	81	95	36	41	43	21	61	27	4	5	7	4
VIH/SIDA	14	90	94	35	35	20	19	53	18	87	84	58	12	21	21	16	42	42	7	2	5	1
Cardiopatía hipertensiva	15	34	22	18	8	17	16	10	19	7	13	16	13	13	12	18	13	13	17	17	36	14
Cáncer de colon y recto	16	6	6	26	12	34	17	20	8	5	7	8	16	15	13	17	9	17	36	32	40	24
Ca de estómago	17	20	7	21	7	26	23	13	11	14	17	27	22	14	15	9	11	15	40	45	39	32
Cáncer de hígado	18	22	8	45	6	31	15	16	32	20	22	23	25	20	26	20	28	23	37	36	28	28
Suicidio	19	15	10	14	17	11	22	12	7	15	19	14	19	18	23	24	16	20	24	22	27	15
Caídas	20	13	21	13	16	36	18	32	21	19	16	17	21	22	20	27	36	24	43	29	34	45
Paludismo	21		120	36	118	32	45						112	115	117	115		71	1	8	3	74
Cáncer de mama	22	11	18	20	18	29	20	19	14	12	9	12	17	19	17	25	14	19	29	31	23	22
Anomalías congénitas	23	57	68	16	31	12	21	18	53	56	66	50	20	16	24	15	35	9	9	9	10	27
Asma	24	61	50	11	43	7	14	42	72	66	71	74	34	59	70	76	72	21	25	28	33	20
Cáncer pancreático	25	14	12	55	19	62	34	27	17	13	11	13	29	26	25	30	17	28	58	59	54	33

Figura 2. Principales causas de muertes en el mundo en el 2017 más detallado. EPOC, enfermedad crónica obstructiva pulmonar. Datos del IHME por el visualizador GBD[6]. El eje "x" son las regiones, el eje "y" las posibles causas de muerte ordenadas por la cantidad de muertes de mayor a menor. El número en cada casilla del gráfico indica el orden obtenido para la causa de muerte en la región.

Si nos enfocamos solo en las muertes por cáncer obtendremos el ranking de la **Figura 3,** donde los primeros 10 tipos de cáncer con mayor número de muertes por cada 100000 habitantes en el año 2017 se listan en orden: pulmón, colorrectal, gástrico, hígado, mama, páncreas, esófago, próstata, otras neoplasias malignas y leucemia.



Tierra	Australasia	Asia Pac IA	Asia S	Asia Or	Oceanía	SEA	Asia C	Europa Or	Europa C	Europa Occ	Am N IA	Am Latina Caribe	Am Latina Central	Am Latina Trop	Am Latina Andina	Am S	ANMO	África Sub C	África Sub Or	África Sub Occ	África Sub S
Cáncer de pulmón	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	3	9	6	1
Cáncer de colon y recto	2	2	2	5	5	6	3	5	2	2	2	2	3	3	2	4	2	3	4	4	8	6
Ca de estómago	3	8	3	3	3	2	5	2	3	5	6	9	5	2	3	1	3	2	7	11	7	8
Cáncer de hígado	4	10	4	11	2	5	2	3	9	7	8	7	7	6	7	5	10	6	5	5	3	7
Cáncer de mama	5	4	7	2	6	3	4	4	4	3	3	3	4	5	5	7	4	4	2	3	2	5
Cáncer pancreático	6	5	5	17	7	12	10	8	6	4	4	4	9	9	6	10	6	9	11	13	11	9
Cáncer de esófago	7	11	10	8	4	17	13	6	14	17	13	11	11	17	9	18	12	13	6	6	12	4
Cáncer de próstata	8	3	8	14	11	9	9	12	7	6	5	5	2	4	4	3	5	10	9	7	1	2
Otras neoplasias malignas	9	13	11	7	8	8	7	7	5	10	10	13	8	10	8	9	9	7	8	1	5	10
Leucemia	10	6	12	10	9	7	6	10	8	9	7	6	10	7	12	8	11	5	10	10	9	11
Cáncer de cérvix	11	24	18	9	12	4	8	11	16	14	21	20	6	8	11	6	8	16	1	2	4	3
Linfoma no Hodgkin	12	7	9	15	13	11	12	15	17	16	11	8	12	11	13	11	14	12	13	8	10	12
Cáncer cerebral y del sistema nervioso	13	12	20	16	10	13	11	9	13	11	14	12	15	12	10	13	17	8	12	12	13	17
Cáncer de vejiga	14	15	13	19	14	19	18	16	12	8	9	10	14	19	17	19	16	11	14	16	14	14
Cáncer de labios y cavidad oral	15	22	19	4	18	14	15	17	15	18	20	21	17	23	14	20	22	21	15	15	18	13
Cáncer de ovario	16	17	16	18	17	16	14	14	11	13	15	15	19	13	18	14	18	14	16	14	15	15
Cáncer de vesícula y tracto biliar	17	21	6	13	15	22	16	20	23	15	18	22	23	14	15	12	7	17	18	20	20	24
Cáncer de riñón	18	14	14	27	20	25	22	13	10	12	12	14	21	16	19	15	13	19	21	24	21	23
Cáncer de laringe	19	26	26	12	19	18	19	18	19	19	24	23	13	22	16	25	21	15	17	19	19	20
Otros cánceres faríngeos	20	25	22	6	26	20	23	21	21	22	25	25	24	27	20	26	26	29	27	26	27	26
Mieloma múltiple	21	16	17	20	23	23	26	24	22	23	17	16	18	18	21	17	19	18	20	18	17	19
Otras neoplasias	22	18	15	23	22	21	20	23	24	24	16	17	20	15	23	23	15	20	23	28	23	18
Cáncer de cuerpo de útero	23	23	21	22	24	10	21	19	18	20	22	19	16	21	24	16	20	23	19	22	22	21
Cáncer nasofaríngeo	24	27	27	21	16	15	17	27	28	28	27	29	25	29	29	30	30	24	26	21	25	27
Cáncer de piel, no melanómico	25	19	24	25	21	24	25	22	25	25	26	24	22	20	22	21	23	28	22	23	26	16

Figura 3. Principales tipos de cáncer como causas de muerte en el 2017.
Datos del IHME por el visualizados GBD[6]. El eje "x" son las regiones, el eje "y" las posibles causas de muerte ordenadas por la cantidad de muertes de mayor a menor. El número en cada casilla del gráfico indica el orden obtenido para la causa de muerte en la región.

También es de importancia evaluar la relación entre la incidencia de cáncer y las muertes por la misma. La incidencia de cáncer en el mundo en el 2018 se estimó de 18.1 millones de casos nuevos y 9.6 millones de muertes[7]. En la **Figura 4** se muestra la incidencia y mortalidad del cáncer en el 2018, en los 10 tipos de cáncer[8] con mayor mortalidad. El porcentaje de muertes vs incidencia (número de muertes dividido el número de incidencias) en algunos tumores es relativamente bajo, ejemplo de esto son: próstata y mama donde esta proporción representa alrededor del 28% y 30% respectivamente. Mientras para otros tipos de cáncer esta proporción es relativamente alta, como son los de hígado y pulmón son de 93% y 84% respectivamente. Para los cánceres donde esta proporción es baja puede estar dado a un pronóstico temprano, mejores tratamientos y tumores menos agresivos.

Por el contrario, los de alta proporción, son cánceres donde es muy acelerado el avance de la enfermedad o también altamente relacionado con factores externos (alcohol, tabaco, virus, etc)

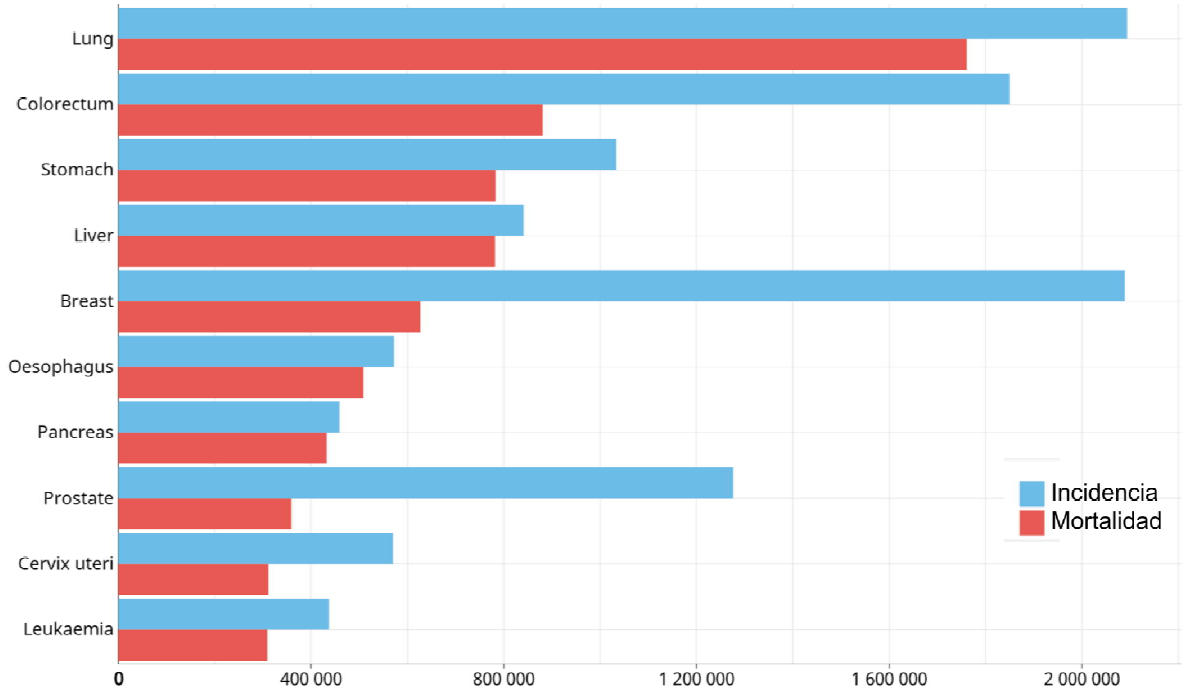


Figura 4. Incidencia y mortalidad en el top 10 de los tipos de cáncer en el 2018. Los tipos de cáncer se ordenaron por mortalidad de mayor a menor. Accedido 2020-04-20 https://gco.iarc.fr/today

La incidencia y mortalidad del cáncer tiene un comportamiento diferente dependiendo de la región (**Figura 5**). Se puede apreciar que los países con mayor incidencia de cáncer son generalmente los más desarrollados, siendo estos también los de mayor cantidad de muertes, exceptuando algunos casos extremos. También se puede apreciar que Australia es el país con mayor incidencia de cáncer, en particular de melanoma (debido a la radiación ultravioleta), pero siendo este cáncer no de tanta mortalidad, este país no es el de mayor cantidad de muertes por cáncer.

Utilizando los datos demográficos de todos los países se estima que para el 2040 la incidencia de cáncer y la mortalidad por esta enfermedad aumentarán en más de 50% el número de casos (**Figura 6**).

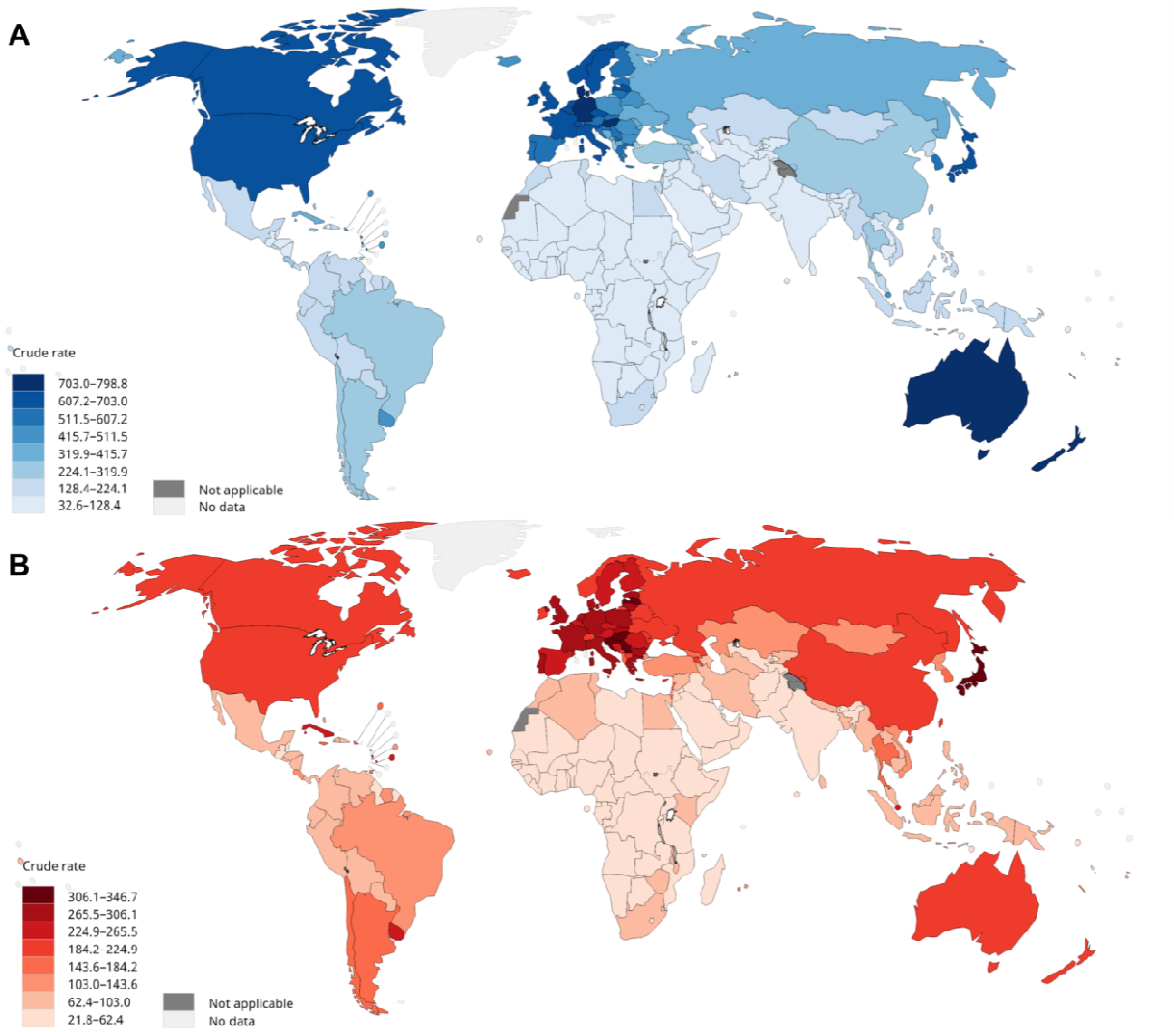


Figura 5. Incidencia y mortalidad del cáncer por región en el 2018.

A: incidencia del cáncer. **B:** muertes por cáncer. Valores cada 100000 habitantes.[9]

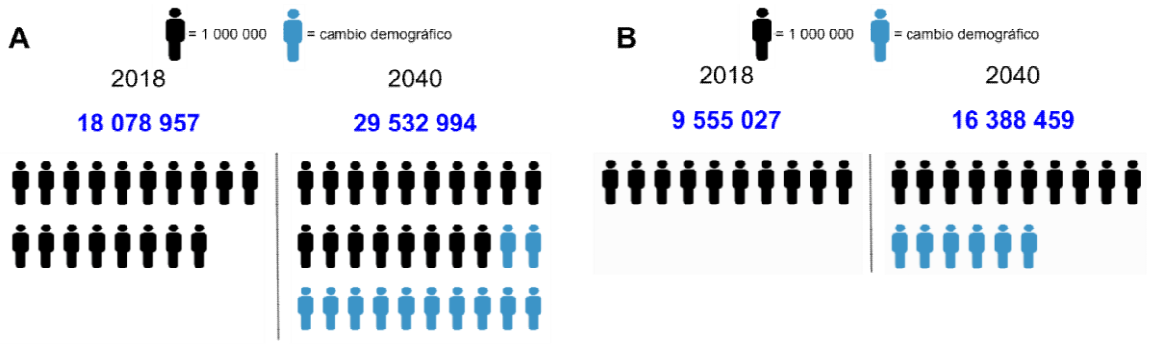


Figura 6. Estimación de incidencia y mortalidad del cáncer para el 2040.

A: incidencia de cáncer en la población. **B:** muertes por cáncer. [10]

1.2. Tratamientos para el cáncer

Existen muchos tipos de métodos de tratamiento del cáncer según el tipo de cáncer y en qué etapa está avanzado. No existe un método o técnica particular para el

tratamiento del cáncer. Las opciones de tratamiento pueden ser quimioterapia, cirugía, radioterapia, terapia hormonal, terapia dirigida, incluida la inmunoterapia, etc. En algunos casos, el plan de tratamiento puede usar una combinación de los métodos de tratamiento (multi-terapia) para tener la máxima efectividad para el tratamiento. El propósito básico de un plan de tratamiento del cáncer es tener una cura para el cáncer y, cuando no es posible una cura completa, el plan de tratamiento debe ser suprimir el cáncer a un estado subclínico y mantener el estado normal para que el sujeto conduzca a un estado normal de calidad de vida.[11]

La identificación de objetivos ideales es esencial para un desarrollo exitoso de terapias moleculares dirigidas en cáncer. Una de las bases de la aparición de cáncer está dictada por la alteración del perfil genético que conduce a la mutación o cambios en las proteínas y los receptores que promueven la supervivencia y la proliferación celular. Estas alteraciones genéticas específicas que pueden distinguir las células cancerosas de las células normales, se pueden utilizar como objetivos moleculares en el desarrollo de fármacos moleculares específicos[12].

1.2.1. Cirugía

El tratamiento del cáncer mediante un procedimiento quirúrgico se practica comúnmente para los cánceres no hematológicos. El procedimiento quirúrgico puede proporcionar una cura completa o parcial para el cáncer. El tratamiento del cáncer mediante un procedimiento quirúrgico puede realizarse para cánceres localizados y el tumor es de tamaño pequeño. Algunos de los tratamientos para el cáncer mediante procedimientos quirúrgicos son la mastectomía del cáncer de mama, el tumor cerebral mediante neurocirugía, la prostatectomía para el cáncer de próstata, el cáncer de riñón, el cáncer de pulmón, el cáncer de hígado, etc. El procedimiento quirúrgico no puede eliminar completamente las células cancerosas, por lo que puede volver a crecer en un nuevo tumor y extenderse a otras partes del cuerpo[11].

1.2.2. Quimoterapia

El método de quimioterapia se lleva a cabo mediante el uso de medicamentos, para interfierir con el crecimiento de tumores o incluso destruir las células cancerosas. Este método puede causar efectos secundarios graves, ya que puede perjudicar células o tejidos sanos. Algunas veces la quimioterapia es administrada como dos o más medicamentos al mismo tiempo; llamándose quimioterapia combinada.[11]

1.2.3. Radioterapia

El método de radioterapia utiliza altas dosis de radiación, generalmente radiación ionizante para matar las células cancerosas y destruir los tejidos tumorales. Este tratamiento se usa comúnmente junto con la cirugía para extirpar o reducir el tamaño de los tumores. La radioterapia puede dañar las células normales e inducir los efectos secundarios de las células normales debido a estas radiaciones ionizantes. Se usa comúnmente para tratar la mayoría de los tipos de tumores como los de cerebro, mama, cuello uterino, laringe, hígado, pulmón, páncreas, próstata, piel, estómago, útero, etc[13].

1.2.4. Terapia hormonal

La terapia hormonal combate el tipo de cáncer que dependen de estos químicos para crecer y propagarseal, al cambiar la cantidad de hormonas en el cuerpo. Este método de tratamiento se utiliza en cánceres de mama, sistema reproductivo y próstata.

1.2.5. Terapia dirigida y multidirigida

La terapia dirigida molecular se refiere al uso de medicamentos u otras sustancias que se dirigen a moléculas específicas (objetivos moleculares) para bloquear el crecimiento y la propagación de las células cancerosas[14].

Las terapias dirigidas actúan sobre antígenos de la superficie celular, factores de crecimiento, receptores o las vías de transducción de señales que regulan la progresión del ciclo celular, la muerte celular, la metástasis y la angiogénesis[15]. Los agentes utilizados en la terapia molecular dirigida se clasifican en moléculas pequeñas, anticuerpos monoclonales, vacunas inmunoterapéuticas contra el cáncer y terapia génica[16]. Estos medicamentos pueden bloquear las señales que favorecen la promoción del crecimiento de células cancerosas, interferir con la regulación del ciclo celular y / o inducir la muerte celular de células cancerosas[16]. Con respecto a sus acciones específicas, los medicamentos también pueden dificultar la progresión e invasión del tumor o resensibilizar el tumor resistente a otros tratamientos cuando se usan como complementos de la quimioterapia[17].

La terapia dirigida molecular es una estrategia útil en el tratamiento del cáncer solo o en combinación con agentes de quimioterapia estándar.

La resistencia a los medicamentos asociada a la terapia molecular dirigida en el tratamiento del cáncer puede ser causada por un alto grado de heterogeneidad clonal del cáncer, heterogeneidad genética intratumoral, regulación epigenética y

complejidad de señalización celular que puede hacer que la célula cancerosa se adapte bajo la presión selectiva de los regímenes terapéuticos. Por lo tanto, la predicción del marcador pronóstico para la heterogeneidad intratumoral y la identificación de nuevos objetivos terapéuticos son esenciales para desarrollar fármacos que puedan bloquear de manera óptima la progresión de las células cancerosas que ya son resistentes a los tratamientos anteriores. Además, la identificación de perfiles de expresión de proteínas específicas de tumor y redes de señalización que actúan individual o colectivamente para conferir resistencia a los medicamentos antes de la terapia, puede ser la forma más efectiva de predecir posibles resultados terapéuticos.[14]

1.2.6. Inmunoterapia

La inmunoterapia contra el cáncer ha cambiado el paradigma del tratamiento del cáncer; estas terapias tienen como objetivo mejorar las respuestas inmunitarias antitumorales con menos efectos fuera del objetivo, que las quimioterapias y otros agentes que matan directamente las células cancerosas. Las inmunoterapias se dividen en varias clases: inhibidores de puntos de control, citoquinas activadoras de linfocitos, células CAR T y otras terapias celulares, anticuerpos agonistas contra receptores coestimuladores, vacunas contra el cáncer, virus oncolíticos y anticuerpos biespecíficos[18].

Los puntos de control inmunitarios mantienen las respuestas inmunitarias adecuadas y protegen los tejidos sanos del ataque inmunológico[19]. Los inhibidores más comunes han sido el bloqueo de la proteína de muerte celular programada 1 (PD-1) o de su ligando PD-L1 y la inhibición de antagonistas del antígeno 4 asociado a los linfocitos T citotóxicos (CTLA4)[20,21]. Cuando las células T se activan, por ejemplo, en respuesta a la inflamación, expresan PD-1, lo que les permite reconocer células anormales y cancerosas[22,23]. Para evadir el reconocimiento y la eliminación por las células T, las células tumorales expresan PD-L1, que se une a PD-1 en las células T para volverlas inactivas[22,24]. Por lo tanto, bloquear esta interacción con un anticuerpo monoclonal (mAb) que se dirige a PD-1 o PD-L1 activa a las celulas T, causando la muerte de células tumorales mediada por células T. Otro punto de control inmunológico, CTLA4, es una molécula co-inhibitoria que regula el grado de activación de las células T. Las interacciones entre CTLA4 y sus ligandos, CD80 y CD86, inhiben la actividad de las células T y, por lo tanto, promueven la progresión tumoral[25]. Al bloquear la interacción entre CTLA4 y estos

ligandos, las células T permanecen activas y pueden reconocer y destruir células tumorales. En la **Figura 7** se muestra el mecanismo de bloqueo de puntos de control inmunológico[26]. Se han aprobado cinco anticuerpos contra de PD-1 o PD-L1 y un inhibidor de CTLA4 para tratar varios cánceres en función de las mejoras en la supervivencia general en comparación con las quimioterapias tradicionales[27].

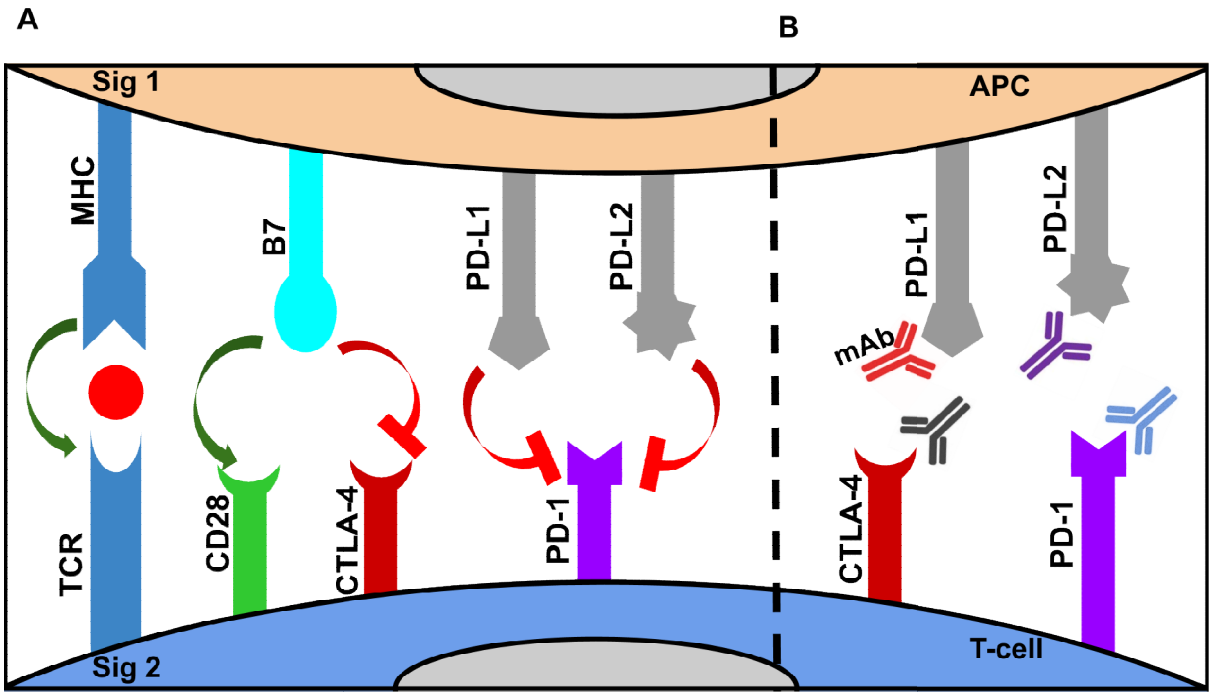


Figura 7. Mecanismo de bloqueo de puntos de control inmunológico.
La activación de las células T está determinada en última instancia por el equilibrio de las señales coestimuladoras versus co-inhibidoras. **A:** Tras la activación de las células T (señal 1 + señal 2), los receptores co-inhibidores como CTLA-4 y PD-1 se regulan positivamente. CTLA-4 se une con mayor avidez a las moléculas B7 CD80 / CD86 y supera la unión de CD28, lo que lleva a la inhibición de la activación de las células T. **B:** Los anticuerpos monoclonales dirigidos contra receptores inhibidores bloquean su capacidad para unirse a su ligando respectivo, lo que favorece la señal coestimuladora y prolonga así la activación de las células T y la respuesta inmune antitumoral. Tomado de Steven et al[26].

Los ensayos clínicos han demostrado que la inmunoterapia mejora la supervivencia total (OS) de pacientes con tumores avanzados de carcinoma de células no pequeñas de pulmón (NSCLC), melanoma, carcinoma de células renales, cáncer urotelial, linfoma de Hodgkin y otros[28–33]. Estos resultados han llevado a la aprobación de inhibidores de puntos de control inmunitarios (ICI) para el tratamiento

de primera o segunda línea de varios tipos de neoplasias malignas. Sin embargo, no todos los pacientes metastásicos responden al tratamiento con inmunoterapia y solo el 15-35% de los casos obtienen un beneficio clínico duradero[34]. Además, los costos del tratamiento son elevados y un alto porcentaje de pacientes experimentan efectos adversos graves, como la neumonitis relacionada con el sistema inmunitario[35]. En consecuencia, se necesitan biomarcadores predictivos para identificar los subconjuntos de pacientes con mayor probabilidad de responder a la inmunoterapia[36].

1.3. Resistencias a los medicamentos para el cáncer

La resistencia a los medicamentos es un fenómeno bien conocido que se produce cuando las enfermedades se vuelven tolerantes a tratamientos farmacéuticos. Este concepto se consideró por primera vez cuando las bacterias se volvieron resistentes a ciertos antibióticos, pero desde entonces se han encontrado mecanismos similares en otras enfermedades, incluyendo el cáncer. Aunque muchos tipos de cáncer son inicialmente susceptibles a la quimioterapia, con el tiempo pueden desarrollarse resistencia a través diversos mecanismos, como mutaciones de DNA y cambios metabólicos que promueven la inhibición y degradación de drogas. Entre los mecanismos de resistencia se incluyen: la inactivación de la droga o el flujo de degradación de la misma o la alteración de la diana del medicamento, la reparación del daño del DNA, la inhibición de la muerte celular y la transición epitelial-mesenquimal[37].

En su definición más simple, la terapia contra el cáncer funciona como sistema de tres componentes: (i) una terapia; que se dirige (ii) a una población de células cancerosas; dentro de (iii) un entorno *host* particular. La resistencia a los medicamentos en cáncer puede darse como una resistencia primaria o adquirida luego de la exposición a la droga (**Figura 8**); sin embargo, en la práctica, muchos tumores son o se vuelven resistentes debido a la superposición de combinaciones de estos factores.[38]

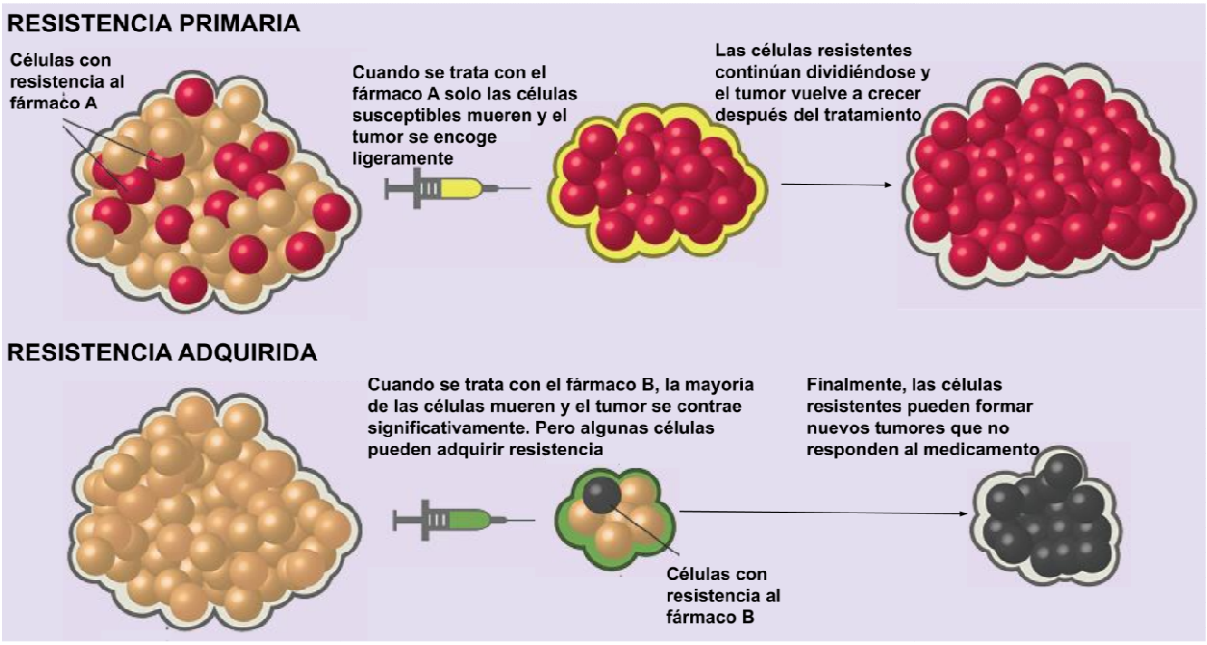


Figura 8. Resistencia primaria y adquirida.

Editado y traducido al español.[39]

1.4. Mutaciones en la secuencia de proteína

Como se ha dicho anteriormente, el cáncer es una enfermedad de evolución clonal[1], donde una célula adquiere mutaciones y/o alteraciones epigenéticas que la vuelven más apta[2]. Esto conlleva al estudio de las mutaciones en pacientes con cáncer como una de las principales líneas de investigación sobre el cáncer.

1.4.1. Tipo de mutaciones según impacto en la secuencia de la proteína

El desarrollo y la función de un organismo están controlados en gran parte por genes. Las mutaciones pueden provocar cambios en la estructura y/o función de una proteína o causar una disminución o pérdida completa de su expresión[40].

Una mutación que implica un cambio en un solo par de bases, a menudo llamada mutación puntual, o una eliminación de algunos pares de bases, generalmente afecta la función de un solo gen. Los tipos de mutación se pueden clasificar en [40]: i) Mutación por cambio (missense), que da como resultado una proteína en la que un aminoácido se sustituye por otro; ii) Mutación de fin de proteína (nonsense), en la que la mutación genera un codón de terminación, lo que lleva a la terminación prematura de la traducción, iii) Mutación silenciosa (*silent*), es la que la mutación no conduce a un cambio de aminoácido, por lo que la proteína no se ve afectada; iv) Mutación por cambio de marco de lectura (*frameshift*), la introducción o deleción de nucleótidos no múltiplo de 3, lleva a que la traducción desde ese punto de una

secuencia proteica completamente cambiada, hasta el encuentro de un codón stop, v) Las deleciones, siempre que sea un número de nucleótidos múltiplo de 3 darán como resultado la eliminación de un pequeño número de aminoácidos contiguos; vi) Las inserciones, de manera similar a las deleciones, si se insertan un número de bases múltiplo de 3, resultará en la adición de unos pocos aminoácidos dentro de la secuencia de la proteína. Otros números darán cambio de marco de lectura; y vii) Luego hay mutaciones que son la combinación de estas últimas 2 llamadas *indel*. En la **Tabla 1** se muestra ejemplos de mutaciones de pequeña escala, y una secuencia ejemplo tanto a nivel de DNA como de proteína.

Tipo de mutación	**Donde**	**Secuencia**
Normal	Dna	AUG GCC TGC AAA CGC TGG
	Proteína	MET ALA CYS LYS ARG TGP
Silente (Silent)	Dna	AUG GC**T** TGC AAA CGC TGG
	Proteína	MET ALA CYS LYS ARG TGP
Codón de fin (Nonsense, stop codon)	Dna	AUG GCC TG**A** AAA CGC TGG
	Proteína	MET ALA **STOP**
Cambio de aminoácido (Missense)	Dna	AUG GCC **G**GC AAA CGC TGG
	Proteína	MET ALA **ALA** LYS ARG TGP
Cambio de marco por inserción de un nucleótido (frameshift)	Dna	AUG GCC**C**TGC AAA CGC TGG
	Proteína	MET ALA **LEU GLN THR LEU**
Cambio de marco por deleción de un nucleótido (frameshift)	Dna	AUG GC**-** TGC AAA CGC TGG
	Proteína	MET **ALA GLU ASN ALA**
Inserción de un codón (insertion)	Dna	AUG GCC **CTG** TGC AAA CGC TGG
	Proteína	MET ALA **LEU** CYS LYS ARG TGP
Deleción de un codón (deletion)	Dna	AUG GCC **---** AAA CGC TGG
	Proteína	MET ALA **---** LYS ARG TGP
Cambio complejo, inserción, cambio y deleción (complex change, indel)	Dna	AUG GCC**C**TGC **G**AA **-**GC TGG
	Proteína	MET ALA **LEU ARG THR** TGP

Tabla 1. Ejemplo de los distintos tipos de mutaciones.

En negrita y rojo se señalan los cambios que ocurren en el DNA y en negrita y azul el cambio resultante a nivel de proteína.

El segundo tipo importante de mutación implica cambios a gran escala en la estructura cromosómica y puede afectar el funcionamiento de numerosos genes, lo que tiene como resultado importante consecuencias fenotípicas. Tales mutaciones

(o anomalías) cromosómicas pueden implicar la deleción o inserción de varios genes contiguos, la inversión de genes en un cromosoma o el intercambio de grandes segmentos de DNA entre cromosomas no homólogos[40].

1.4.2. Mutaciones adaptativas (*drivers*) y neutrales (*passengers*)

Como cualquier proceso evolutivo, la acumulación de mutaciones en el cáncer cuenta con mutaciones neutrales y mutaciones adaptativas. Un tumor sólido típico tiene docenas de sustituciones, pero sólo unas cuantas participan en la enfermedad y se conocen como *drivers*. El resto son llamadas *passengers* pues su ocurrencia no otorga ningún beneficio a la célula y no se relacionan con la enfermedad.

Hay tres tipos de genes que pueden recibir mutaciones *driver*: oncogenes, supresores de tumores y genes de estabilidad[41]. Los oncogenes mutados se activan promoviendo la división celular constitutiva o en condiciones en las que normalmente no lo harían. La activación puede deberse a translocaciones cromosómicas, amplificaciones o mutaciones en residuos regulatorios[41]. Tales mutaciones son contadas, poco desestabilizantes y concentradas en regiones específicas de la proteína[42]. MYC, las GTPasas regulatorias de la familia RAS y los receptores quinasa son algunos ejemplos[43].

Los genes supresores de tumores, al contrario que los oncogenes, reducen su actividad por sustituciones no sinónimas de residuos esenciales o codones de *stop*, inserciones/deleciones, silenciamiento epigenético o RNAs no codificantes, inhibiendo de esta manera su capacidad de inducir la apoptosis o el arresto celular[41]. En general, los supresores de tumores exhiben más mutaciones, sobre todo disruptivas, a todo lo largo de su secuencia[42]. Ejemplos puntuales son TP53, RB1 y PTEN[43].

Los genes de estabilidad normalmente reparan los daños que el DNA sufre durante la replicación o la exposición a mutágenos; cuando las mutaciones los inactivan se produce inestabilidad genómica; al igual que con los supresores de tumores, ambos alelos de estos genes deben desactivarse para tener efecto, por lo que el fenómeno de pérdida de la heterocigosidad es común. Sorprendentemente, las predisposiciones hereditarias suelen deberse a mutaciones germinales de genes de estabilidad como BRCA1, BRCA2, MLH1 y MSH2[41].

1.5. Datos sobre muestras de cáncer

Actualmente los hospitales no están obligados a registrar en bases de datos nacionales o internacionales los resultados de secuenciación parcial o total de las

muestras de cáncer ni las características cuantitativas o cualitativas de una muestra ni del paciente. Existen varias bases de datos en línea donde se registran algunos datos de las muestras de cáncer y de los pacientes enmarcados en determinados proyectos donde los datos se hacen públicos o con acceso limitado para los científicos. Las tres principales son: el catálogo de mutaciones somáticas en cáncer (COSMIC)[44], datos del consorcio internacional del genoma del cáncer (ICGC-DCC)[45,46] y el atlas del genoma del cáncer (TCGA)[47]. Cada una de estas bases de datos dienen sus características particulares, y también entre ellas existen datos solapados.

COSMIC en su web https://cancer.sanger.ac.uk/cosmic, se define como: es el Catálogo de mutaciones somáticas en el cáncer, es el recurso más grande y completo del mundo para explorar el impacto de las mutaciones somáticas en el cáncer humano. International Cancer Genome Consortium. ICGC-DCC en sus webs, una del consorcio https://icgc.org/ otra del portal de datos https://dcc.icgc.org/, plantea que el ICGC se estableció para lanzar y coordinar una gran cantidad de proyectos de investigación que comparten un objetivo común de desentrañar los cambios genómicos presentes en muchas formas de cáncer que contribuyen a la carga de la enfermedad en las personas en todo el mundo; y se almacena sus datos en el portal DCC. Mientras que en la web de TCGA https://www.cancer.gov/tcga, dice que un programa histórico de genómica del cáncer, que es un esfuerzo conjunto entre el Instituto Nacional del Cáncer y el Instituto Nacional de Investigación del Genoma Humano y que comenzó en 2006 reuniendo a investigadores de diversas disciplinas e instituciones múltiples.

Estas tres bases de datos se diferencian entre sí, por el soporte que las actualiza, los datos que almacenan y la accesibilidad de los datos almacenados. Algunos de los datos que se guardan en estas bases de datos son las mutaciones, fusión de genes, datos de metilación y expresión de genes, etc., así como datos clínicos de la muestra. Los datos de mutaciones dependen de su forma de secuenciación ya que puede ser secuenciado todo el genoma (WGS) o todo el exoma (WES) o parte de ellos con panel específicos. COSMIC tienen un libre acceso a todos sus datos almacenados, necesitando solo registrarte para descargar la información. Mientras que el portal de ICGC-DCC tiene restricciones sobre los datos almacenados siendo algunos de uso público y otros de uso restringido al que se le puede solicitar acceso. Por último solo se puede tener acceso a los datos almacenados en TCGA haciendo una solicitud formal pidiendo determinados estudios y tipo de datos.

También vale la pena recalcar que COSMIC en cada actualización recopila datos tanto de ICGC-DCC como de TCGA y además incorpora datos de literatura. No resulta útil trabajar con las 3 bases de datos juntas porque tienen mucha información solapada, y además COSMIC verifica ese solapamiento comprobando algunos de los identificadores. Aunque en cada nueva versión de COSMIC es posible que no estén los últimos datos actualizados de ICGC-DCC y TCGA, lo más posible es que estén en la versión siguiente, por lo que consideramos COSMIC como la más completa de las 3 bases de datos para el estudio de mutaciones.

1.6. Patrones mutacionales

Como bien lo indica su nombre los patrones mutacionales, no es más que encontrar características distintivas entre las mutaciones provenientes de un grupo de muestras. Particularmente en el cáncer podemos buscar y encontrar estos patrones tanto a nivel de nucleótidos (DNA) como de aminoácidos (proteína). El análisis de patrones de mutaciones somáticas es una herramienta poderosa para comprender la etiología de los cánceres humanos[48]. En los primeros estudios se identificaron sustituciones de un solo punto debido a fumar, exposición a la luz ultravioleta, consumo de aflatoxinas, ingesta de productos que contienen ácido aristolóquico, entre otros[49–52]. El advenimiento de las tecnologías de secuenciación paralela masiva[53] permitió una evaluación barata y eficiente de las mutaciones somáticas en un genoma de cáncer. Esto proporcionó una oportunidad sin precedentes para examinar los patrones mutacionales somáticos mediante la secuenciación de múltiples genes, de todas las regiones codificantes del genoma humano (secuenciación del exoma completo), o de la secuencia completa del genoma (secuenciación del genoma completo)[54].

1.7. El objetivo principal

Es la búsqueda de “patrones” entre las mutaciones en pacientes con cáncer. Esto con la finalidad de contribuir a la selección de terapias tanto mono como multi dirigidas. Además también obtener patrones que permitan cuantificar la carga mutacional de las muestras de cáncer y así identificar si un paciente se beneficiaría de la terapia inmunológica.

Ambos trabajos tienen enfoques completamente distintos pero los une la forma (encontrar patrones en las mutaciones) y el objetivo (mejorar la selección del tipo de terapia a aplicar a los paciente de cáncer). También se pretende encontrar patrones

en resistencia a fármacos pero debido a la falta de datos, no ha sido posible profundizar en este puto en este momento.

Para ello se utilizarán los datos públicos de COSMIC, lenguajes de programación como R y Python y pruebas estadísticas incorporadas en los paquetes de estos lenguajes.

Los **objetivos específicos** son:

- Encontrar relaciones de exclusión y co-mutación entre las mutaciones presentes en las muestras de cáncer para sugerir terapias. Ya sea el uso de múltiples fármacos cuando dos o más mutaciones co-ocurren o desalentar la combinación de ciertos fármacos ya que dado el patrón mutacional del paciente, no resultarían efectivas.
- Encontrar características con las cuales se puedan diferenciar mutaciones *driver* de *no-diver* y haciendo uso de ellas poner sugerir posibles *drivers*.
- Estudiar las mutaciones que se consideran de resistencia a medicamentos en el entorno 3D de la proteína.
- Encontrar relaciones entre las mutaciones de resistencia a medicamentos.
- Encontrar conjuntos de genes (paneles) y modelos matemáticos asociados capaces de predecir la carga mutacional total para ser utilizados como biomarcadores y así decidir el uso de la inmunoterapia en pacientes.
- Determinar si este conjunto de genes sería tumor específico o no.

2. Materiales y métodos generales

En este capítulo se describen los materiales y métodos generales. Dentro de cada uno de los capítulos de Resultados se explican en detalles los métodos utilizados en dicho capítulo.

2.1. Lenguajes de programación en bioinformática: R y Python

El procesamiento de datos científicos se puede realizar con muchas herramientas, algunas de ellas privadas y otras de acceso libre. En esta tesis he utilizado R[55] como lenguaje de programación. Haciendo uso de este lenguaje y sus librerías nos permite procesar y graficar datos, además cuenta con una amplia comunidad y soporte de versiones y tiene un enfoque al análisis estadístico. Es uno de los lenguajes de programación más utilizados en investigación científica, siendo además muy popular en los campos de aprendizaje automático (machine learning), minería de datos, investigación biomédica, bioinformática y matemáticas financieras. Además R brinda un gran conjunto de librerías imprescindibles para el trabajo con datos biotecnológicos y su visualización, como son:

- GenomicFeatures and GenomicRanges[56], biomaRt[57,58], Biostring[59], rtracklayer[60], etc.: para el trabajo con coordenadas genómicas y secuencias.
- Seqinr[61] y rphast[62]: para realizar alineamientos y obtener regiones del alineamiento.
- Caret[63], stats[55], survival[64,65]: para creación de modelos y análisis estadísticos.
- Ggplot2[66], ggridges[67], VennDiagram[68], survminer[69], plotROC[70]: para realizar distintos tipos de gráficos.
- Rjson[71]: manejo de datos tipo json.
- Httr[72]: para realizar consultas a las web programáticamente.

Otro de los lenguajes utilizados ha sido Python[73]. Python es un lenguaje de programación fácil de aprender que está ganando terreno entre científicos. Esto es probable porque es fácil de usar, pero lo suficientemente potente como para lograr la mayoría de los objetivos de programación. Con Python se puede comenzar a hacer programación real muy rápidamente. Los científicos están utilizando Python para la visualización molecular, la anotación genómica, la manipulación de datos y muchas otras aplicaciones.[74]

Python también brinda un gran conjunto de librerías imprescindibles para el trabajo con datos biotecnológicos y su visualización, como son:

- Biopython[75]: manejo de estructuras PDB, secuencias, alineamientos, etc.
- Matplotlib[76,77]: gráficos de todo tipo.

Por último he realizado numerosos programas *Ad Hoc* en múltiples lenguajes para la solución de problemas o cálculos específicos de la tesis.

2.2. Pruebas estadísticas utilizadas

El análisis estadísticos de datos es uno de los aspectos fundamentales de cualquier estudio, ya que la estadística apoyándose en pruebas es capaz de analizar una hipótesis y decirnos si los resultados son o no producto del azar. En este epígrafe hablaremos de las pruebas estadísticas que hemos utilizado en los estudios. (tomado de https://www.statisticshowto.datasciencecentral.com/)

La prueba de Chi-cuadrado para la independencia compara dos variables en una tabla de contingencia para ver si están relacionadas. En un sentido más general, es una prueba para ver si las distribuciones de variables categóricas difieren entre sí.

La prueba exacta de Fisher es una prueba estadística que se usa cuando se tienen dos variables nominales y se desea averiguar si las proporciones de una variable nominal son diferentes a los valores de la otra variable nominal, se utiliza en el análisis de tablas de contingencia. Para los experimentos con un pequeño número de muestras (menos de 1000), el de test de Fisher es más preciso que la prueba de chi-cuadrado.

La prueba de Mann-Whitney es una prueba no paramétrica para comparar a dos muestras independientes. Nos dice cuán significativas son las diferencias entre los grupos; en otras palabras, permite saber si esas diferencias (medidas en medias / promedios) podrían haber ocurrido por casualidad.

La prueba de Kruskal-Wallis es un método no paramétrico para comparar si las medianas de dos o más grupos son diferentes. Es una extensión de la prueba de Mann-Whitney para 3 o más grupos.

La prueba post-hoc de Dunn es una prueba de comparación múltiple no paramétrica post hoc, es decir, se aplica luego de detectar diferencias significativas con una prueba de Kruskal-wallis y ayuda a analizar las diferencias significativas entre los grupos.

La prueba de correlación de Pearson es una medida lineal entre dos variables aleatorias cuantitativas. La correlación de Pearson es independiente de la escala de medida de las variables.

La prueba de Log-rank es una prueba no paramétrica utilizada para analizar la supervivencia de pacientes en ensayos clínicos contrastando las funciones de supervivencia de dos poblaciones. Puede usarse en presencia de datos censurados (pacientes vivos pero que se perdieron del análisis).

2.3. Redes de datos

Si bien el estudio de las redes tiene una larga historia, con raíces en la teoría de grafos y la sociología, el capítulo moderno de la ciencia de las redes surgió solo durante la primera década del siglo XXI. Estamos rodeados de sistemas que son desesperadamente complicados, estos que captan el hecho de que es difícil derivar su comportamiento colectivo del conocimiento de los componentes del sistema. Dado el importante papel que juegan los sistemas complejos en nuestra vida diaria, en la ciencia y en la economía, su comprensión, descripción matemática, predicción y eventualmente control es uno de los mayores desafíos intelectuales y científicos del siglo XXI[78].

Si queremos entender un sistema complejo, primero necesitamos saber cómo interactúan sus componentes entre sí. En otras palabras, necesitamos un mapa de su diagrama de cableado. Una red es un catálogo de los componentes de un sistema a menudo llamados nodos o vértices y las interacciones directas entre ellos, llamados enlaces o aristas. Esta representación de red ofrece un lenguaje común para estudiar sistemas que pueden diferir mucho en naturaleza, apariencia o alcance[78].

En la ciencia de redes, a menudo distinguimos las redes por alguna propiedad elemental del gráfico subyacente. Los tipos de redes más comunes se presentan en la **Figura 9**. Hay que tener en cuenta que muchas redes reales combinan varias de estas características de red elementales. Por ejemplo, la WWW es un grafo dirigido con múltiples conexiones y con auto-interacciones; la red de llamadas móviles está dirigida y ponderada, sin bucles automáticos. Algunas propiedades importantes de los grafos son el número de nodos (N) y enlaces (L) y una propiedad clave de cada nodo es su grado, que representa el número de enlaces que tiene con otros nodos[78].

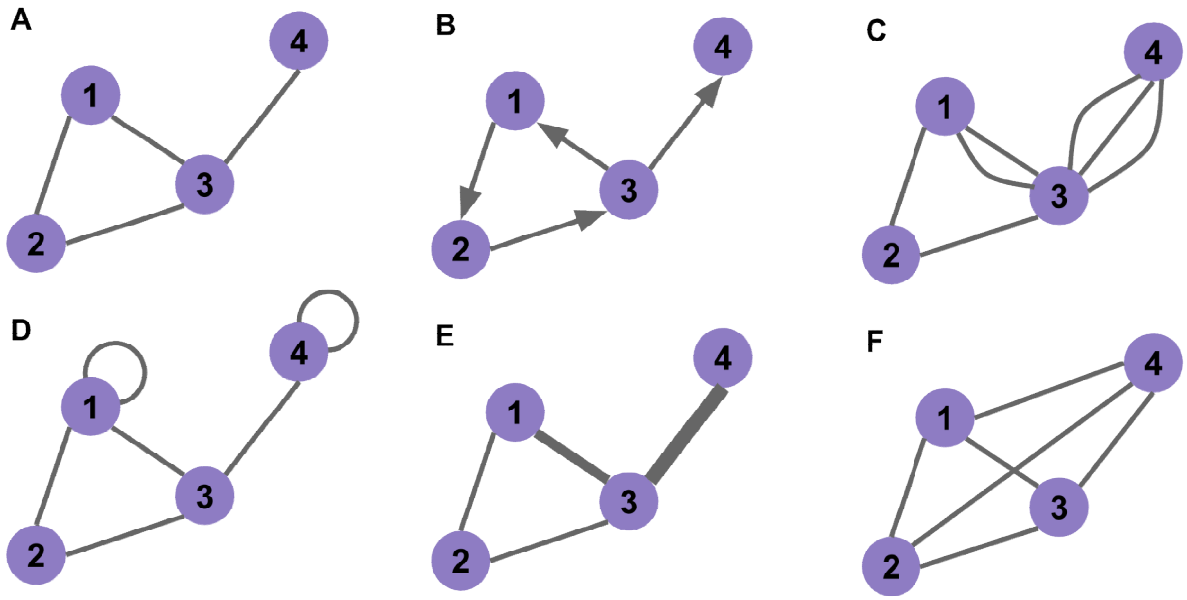


Figura 9. Ejemplo de tipos de grafos / redes más comunes.
Los nodos se presentan en círculos color violeta, un número por cada nodo. Los enlaces se definen como las conexiones entre los nodos. **A**: No dirigido, cuando ninguno de los enlaces tiene dirección. **B**: Dirigido, cuando todos los enlaces tienen dirección. **C**: Multi-grafo, cuando dos nodos se conectan por más de un enlace. **D**: Con bucles, cuando un nodo posee un enlace a sí mismo. **E**: Peso en los enlaces, cuando los enlaces tienen un valor de importancia, por ejemplo la longitud de una carretera que conecta a dos ciudades. **F**: Completo, cuando todos los nodos se conectan con todos.

En esta tesis se han utilizado las redes de datos no dirigidas, con múltiples conexiones y sin bucles entre nodos para visualizar las relaciones de dependencia entre mutaciones en diferentes tipos de cáncer. A estas redes se les analizó la conectividad y el grado de los nodos.

2.4. Modelos de regresión lineal

A menudo se requiere resolver problemas que implican conjuntos de variables de las cuales se sabe que tienen alguna relación inherente entre sí[79]. Por ejemplo, en un medicina quizá se sepa que determinados factores como edad, estatura, calorías de ingesta y pérdida de calorías por ejercicio está relacionado con el peso de la persona. Podría ser de interés desarrollar un método de pronóstico, es decir, un procedimiento que permita estimar el peso de una persona conociendo estos datos. A la edad, estatura, calorías de ingesta diaria y pérdida diaria de calorías por ejercicio se les denomina variables independientes o regresores. Y el peso de la

persona sería la variable dependiente o respuesta. Una forma razonable de relación entre la respuesta Y y los regresores x_i es la relación lineal. Como en este ejemplo, en muchas aplicaciones habrá más de un regresor, es decir, más de una variable independiente que ayude a explicar a Y, a esto se le llama regresión múltiple y se podría escribir como:

$$\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}$$

Ecuación 1. Regresión lineal múltiple

Donde Y es la variable respuesta, los x_i (i = 1, 2, …, n) serían las n variables independientes y los parámetros lineales β_i sus coeficientes de regresión. Cada coeficiente de regresión β_i se estima por medio de de los datos muestrales, usando el método de los mínimos cuadrados.

La prueba t que se utiliza con más frecuencia en la regresión múltiple, es aquella que prueba la importancia de los coeficientes individuales. Con frecuencia estas pruebas contribuyen a lo que se denomina selección de variables, con la cual el analista intenta llegar al modelo más útil, es decir, a la elección de cuál regresor utilizar. Aquí debemos destacar que, si se encuentra que un coeficiente es insignificante, la conclusión sería que la variable explica una cantidad insignificante de la variación de y en la presencia de los demás regresores del modelo[79].

Cuando se está interesado en eliminar variables de una regresión ya que además de llegar a una ecuación de pronóstico funcional, debe encontrar la “mejor regresión” que implique sólo variables que sean predictores útiles se suelen utilizar criterios como son el coeficiente de determinación múltiple (R^2) y la raíz del error cuadrático medio (RSE). La cantidad R^2 tan sólo indica qué proporción de la variación total de la respuesta Y es explicada por el modelo ajustado. Con frecuencia se reportan $R^2 \times 100\%$ e interpretan el resultado como el porcentaje de variación explicado con el modelo propuesto.

En esta tesis se utilizaron modelos de regresión lineal para encontrar los genes que deben de estar presentes en paneles para predecir con calidad la carga mutacional total de pacientes en diferentes tumores.

3. Dependencias entre mutaciones

En este capítulo abordaremos el estudio relacionado a la dependencia entre mutaciones provenientes de pacientes con cáncer. Estas dependencias son fundamentales para comprender si estas relaciones son o no específicas por tejido. Utilizando este conocimiento, si una terapia es útil en determinado tumor, la misma terapia puede ser utilizada en otro tumor que presente similares dependencias entre mutaciones. Además se pueden sugerir combinaciones de terapia atendiendo a las dependencias encontradas entre las mutaciones. Este trabajo fue realizado en colaboración con Soledad Ochoa para su proyecto de maestría, además se encuentra publicado en la revista Human Mutation[80], a continuación se muestran los detalles del estudio.

3.1. Introducción

Hay evidencias de que las alteraciones genéticas en los genes relacionados con el cáncer se agrupan dentro de un conjunto limitado de vías biológicas esenciales[81–83]. Sin embargo, la vía puede ser alterada por mutaciones somáticas u otros cambios en diferentes genes. Por ejemplo, en el glioblastoma multiforme (GBM), la vía p53 se regula negativamente en el 87% de los tumores; pero la base genética de esta regulación negativa varía de paciente a paciente. Las alteraciones reportadas incluyen mutaciones somáticas o deleción homocigota de TP53, eliminación del gen inhibidor de CDKN2A y amplificación de dos genes que codifican MDM2 y MDM4[81].

Los proyectos de perfiles genéticos de tumores a gran escala han revelado alteraciones mutuamente excluyentes en muchos pacientes, incluidas mutaciones *driver* en genes específicos[84]. Por ejemplo, las mutaciones de TP53 y la amplificación de MDM2 rara vez se producen juntas en GBM y pocos pacientes albergan ambas lesiones. Ejemplos adicionales incluyen la exclusión mutua entre mutaciones de APC y CTNNB1 en el cáncer colorrectal , ambos involucrados en la vía de señalización de beta-catenina; o mutaciones en BRAF y KRAS, codificando proteínas de la vía de señalización RAS/RAF[81]. En cáncer seroso de ovario se ha observado una exclusión mutua entre mutaciones de BRCA1 y BRCA2 y el silenciamiento epigenético de BRCA1, mientras que las mutaciones en EGFR y KRAS nunca ocurren simultáneamente en cáncer de células no pequeñas de pulmón[81].

También los perfiles del cáncer han descubierto varios casos de alteraciones concurrentes, sugiriendo que algunos cambios en vías asociadas pueden provocar efectos complementarios en vez de redundantes[85]. Los ejemplos incluyen la eliminación de Fosfatidilinositol 3,4,5-trifosfato 3-fosfatasa y PTEN concomitante con la amplificación de la ERBB2 en cáncer de mama[86], activando mutaciones de MET cuando el gen supresor de tumor de la enfermedad de Hippel-Lindau (VHL) se elimina en el carcinoma renal[87] y la pérdida de CDKN2A junto con mutaciones activadoras en el protooncogen BRAF en melanoma[81,88].

A pesar de toda esta evidencia, en la actualidad no existe un análisis sistemático de mutaciones codependientes en todos los tipos de cáncer. Los únicos estudios disponibles, tratan genes completos en lugar de mutaciones específicas[81,89–93], ignorando el hecho de que diferentes mutaciones en el mismo gen pueden tener efectos muy diferentes, por ejemplo, mutaciones en EGFR G735S, G796S y E804G

inducen su activación oncogénica en cáncer de próstata, mientras que R841K no tiene relevancia funcional[94].

Una mejor y más amplia comprensión de las codependencias entre mutaciones es relevante en muchos aspectos, como la clasificación del tumor, el diagnóstico o el tratamiento de elección. En este sentido, las relaciones de codependencia entre las alteraciones genéticas puede evidenciar epistasis mutacional[95] y resalta la necesidad de una terapia multidirigida. Varios fármacos antitumorales de nueva generación se dirigen a proteínas albergando mutaciones *driver* específicas. Sorafenib es activo contra las células de carcinomas renales y hepáticos que tienen la mutación BRAF.V600E[88]; imatinib contra tumores del estroma gastrointestinal con mutaciones V560G, K642E, N822H o N822K en el gen de serina/treonina quinasa (KIT) o la mutación V561D en PDGFRA[96]; gefitinib, erlotinib y afatinib contra cánceres de pulmón de células no pequeñas con deleciones del exón 19 o la mutación puntual L858R en EGFR[97]; y dabrafenib o vemurafenib contra BRAF.V600E en melanoma[98]. Sin embargo, estos tratamientos centrados en una sola alteración son seguidos casi invariablemente por una recaída debido a la selección de células resistentes[99]. Los enfoques multi-dirigidos contra las alteraciones perjudiciales concurrentes tienen el potencial para retrasar el inicio de la resistencia, y una mejor comprensión de las coocurrencias podría ser de ayuda en este contexto. Debe recordarse que algunos de estos enfoques ya están en uso clínico, como la combinación de BRAF e Inhibidores de MEK (MAP2K1 y MAP2K2) en melanoma metastásico[100], mientras que varios otros actualmente se están probando en ensayos clínicos[101].

3.2. Materiales y Métodos

3.2.1. Mutaciones *driver*

Nuestro conjunto de mutaciones *driver* está compuesto por mutaciones descritas por Molina-Vila et al.[102] y almacenadas en la base de datos Kin-Driver[43]. Además tiene otras mutaciones *driver* encontradas en literatura que son mutaciones en la proteínas RAS[103], PIK3CA[104] y TP53[105]. Obteniendo un conjunto de 289 mutaciones *driver* (**Anexo 1**).

3.2.2. Procesamiento de los datos

Se utilizaron los datos de la versión 75 de COSMIC, obteniendo de esta forma la información de 1178444 muestras de 47 tejidos de origen con 193 histologías y 716

sub-histologías. Las muestras aportan un total de 2812088 mutaciones de 2128846 posiciones secuenciadas. Se definió como tipo de cáncer a la combinación única de tejido de origen, histología y sub-histologías, por ejemplo: lung/carcinoma/adenocarcinoma. Para contar mutaciones únicas, se mapean los identificadores ENSEMBL del transcrito a identificadores UNIPROT de proteínas únicas y se concatenaron con la mutación, por ejemplo P15056.V600E.

Las mutaciones fueron agrupadas por posición según el tipo de alteración, a continuación las reglas utilizadas para la agrupación:

- Las sustituciones se denominan con aminoácido inicial + posición. Ej: BRAF.V600E y BRAF.V600M son BRAF.V600
- Las deleciones como aminoácido + posición + "del". Ej: EGFR.E746_A750del y EGFR.E746_T751delELREAT son EGFR.E746del
- Las inserciones como aminoácido + posición + "ins". Ej: FLT3.Y599_D600insEYEYEY y FLT3.Y599_D600insR son FLT3.Y599ins
- Las sustituciones complejas, como aminoácido + posición + ">". Ej: EGFR.E746_A750>IP y EGFR.E746_A752>V son EGFR.E746>
- Los cambios de marco de lectura como aminoácido + posición + "fs". Ej: CACNG3.S172fs*21 y CACNG3.S172fs*31 son CACNG3.S172fs

Se filtraron las mutaciones *Nonstop extension*, *Substitution coding silent* y *Whole gene deletion*. El número de muestras se redujo así a 1107460 de 1298 tipos de cáncer con 1615508 posiciones de 19297 proteínas mutadas. Se espera que gran proporción de las mutaciones sean *passenger*, por lo que se filtran todos las mutaciones secuenciadas menos de 1000 veces; el límite se justifica, como se muestra en la **Figura 10**, dado que una gran cantidad de ellas que son *driver* lo pasan, 283 de 285. Además, los 1298 tipos de cáncer incluyen cánceres sin mutaciones y pobremente secuenciados de los que no pueden hacerse aseveraciones, así que se filtran aquellos con menos de 10 muestras secuenciadas y ninguna mutada. El set de datos final consiste en 365096 posiciones de 1329 proteínas y 1098411 muestras de 687 tipos de cáncer.

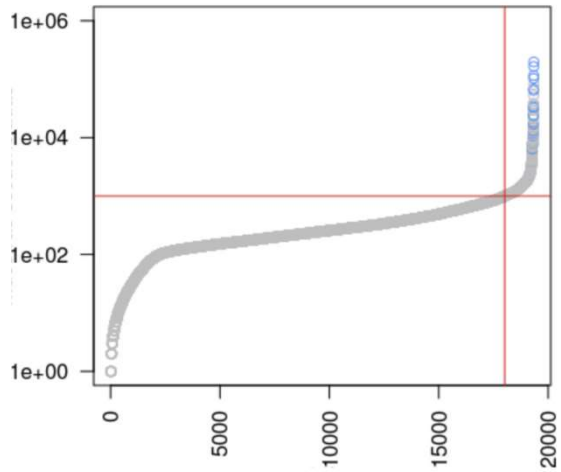


Figura 10. Número de veces secuenciada cada proteína.

El eje X es el número de la proteína y el eje Y es el número de posiciones secuenciadas. Las proteínas se encuentran ordenadas por la cantidad de veces secuenciadas. En color azul se muestran las proteínas con mutaciones *driver*.

3.2.3. Pares de mutaciones relacionadas

Para identificar mutaciones significativamente asociadas, se calcularon las tablas de contingencia de todos los pares de mutaciones co-secuenciados de cada tipo de cáncer. Ejemplo de tabla de contingencia en **Figura 11**, donde m1 y m2 son mutaciones a analizar; a sería cantidad de muestras donde están ambas están mutadas; b la cantidad de muestras donde no está mutado m1 pero si lo está m2; c la cantidad de muestras donde no está mutado m2 pero si lo está m1; y d la cantidad de muestras donde ninguno de los dos está mutado.

Manteniendo sólo los pares con más de 10 muestras mutadas para cada miembro del par, se genera un total de 262375 tablas en 135 tipos de cáncer. Luego de someter cada tabla de contingencia a una prueba de Fisher y corregir los p valores con FDR (*False Discovery Rate*), quedan 43136 pares de mutaciones de 82 tipos de cáncer.

		Mutación m1	
		mutado	no mutado
Mutación m2	mutado	a	b
	no mutado	c	d

Figura 11. Ejemplo de una tabla de contingencia.

Donde m1 y m2 son mutaciones a analizar.

3.2.4. Filtrado de posibles falsos positivos

Se compararon las probabilidades observadas y esperadas de cada par de mutaciones, definidas en **Ecuación 1A-B**. Dicha comparación permite determinar si las bajas probabilidades mutacionales permiten que pares no relacionados pasen la prueba exacta de Fisher corregida. Específicamente, las posiciones poco mutadas podrían parecer excluirse entre sí cuando son poco frecuentes, por lo que su probabilidad conjunta esperada sería similar a la observada. Para filtrar este tipo de falsos positivos, para cada par de mutaciones estimamos el intervalo de confianza del 99% de la probabilidad de mutaciones conjuntas observadas. Si la probabilidad de mutación conjunta esperada cae dentro del intervalo de confianza del observado, el par se descartó, lo que es equivalente a filtrar los puntos cerca de la malla en la **Figura 12**. Los intervalos de confianza se calcularon usando la distribución binomial. La distribución B (n, p) se define para cada par con tamaño n, como el número total de muestras con ambas mutaciones secuenciados y la probabilidad de éxito p, como la probabilidad observada de co-mutación.

A $P(m1 \cap m2) = \frac{a}{t}$ **B** $P(m1) * P(m2) = \frac{a+b}{t} * \frac{a+c}{t} = \frac{(a+b)*(a+c)}{t}$

Ecuación 2. Definición de probabilidad observada (A) y esperada (B) entre mutaciones.

La probabilidad esperada, supone independencia entre las mutaciones. Los valores a, b, c y d son los utilizados en la tabla de contingencia y t es la suma de estos valores (total de muestras)

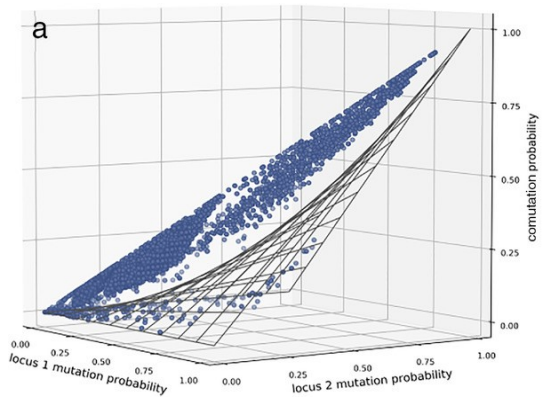


Figura 12. Representación gráfica de la probabilidad observada y esperada.

Los valores de probabilidad observada son representados con los puntos azules, mientras que la malla representa los valores de probabilidad esperada.

Después de este paso se obtuvieron 29489 pares de mutaciones dependientes donde cuya mutación no depende de las bajas frecuencias mutacionales. Estos pares involucran a 495 posiciones en 67 tipos de cáncer de 21 tejidos de origen.
La **Figura 13** muestra la distribución de las mutaciones en los tipos de tejido.

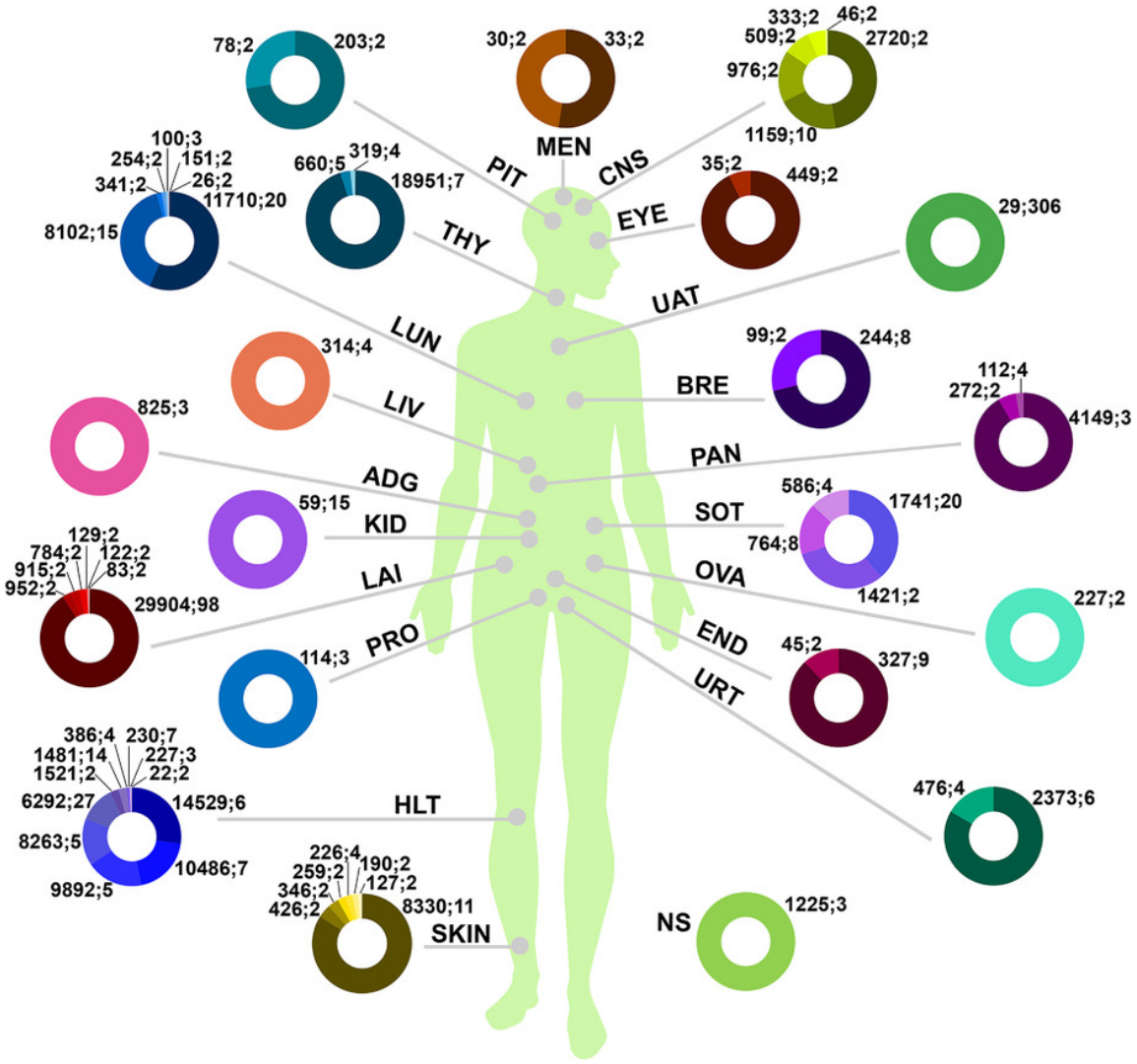


Figura 13. Distribución de mutaciones por los tipos de cáncer.
Cada gráfico de anillo representa los distintos tipos de cáncer en un tejido localizado en la figura humana. Los números x;y representan el número de muestras y el número de mutaciones involucrado en el estudio después de la selección. MEN: meninges; PIT: glándula pituitaria; CNS: sistema nervioso central; THY: tiroide; EYE: ojo; LUN: pulmón; UAT: sistema aero-digestivo superior; BRE: seno; PAN: páncreas; ADG: glándula suprarrenal; LIV: hígado; KID: riñón; SOT: tejido blando; LAI: colorectal; OVA: ovario; PRO: próstata; END: endometrio; HLT: tejido de linfoma y hematopoyético; URT: tracto urinario; SKIN: piel; NS: origen no especificado.

3.2.5. Clasificar las relaciones entre par de mutaciones

Se clasificaron las relaciones entre las mutaciones de cada par atendiendo a sus probabilidades individuales y condicionales de mutación. La clasificación de co-mutación se le asignó a los pares donde alguna de las probabilidades condicionales supere el intervalo de confianza (CI) de la probabilidad individual; y la de exclusión a los pares con ambas probabilidades condicionales inferiores a las individuales. La **Figura 14** muestra la representación gráfica de las reglas utilizadas (**Ecuación 3A-D**). Los 29489 pares se clasificaron en 189 de exclusión y 29300 de co-mutación.

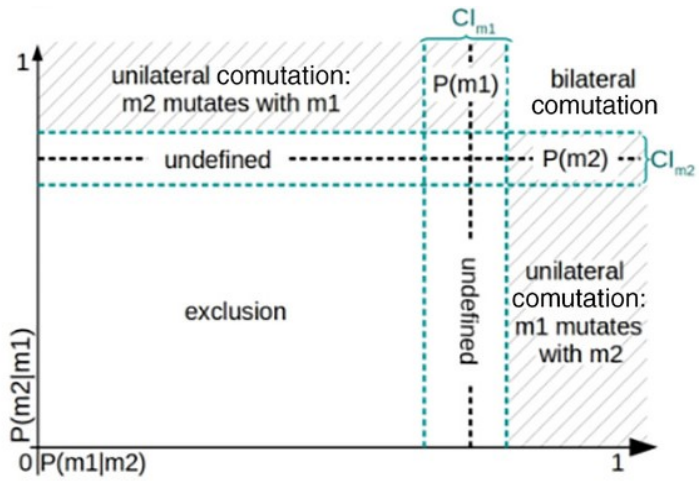


Figura 14. Representación gráfica de las reglas de clasificación para los tipos de relaciones entre mutaciones.

A P(m1 | m2) < CI(m1) y P(m2 | m1) < CI(m2)

B P(m1 | m2) > CI(m1) y P(m2 | m1) > CI(m2)

C P(m1 | m2) > CI(m1) o P(m2 | m1) > CI(m2)

D P(m1 | m2) ∈ CI(m1) o P(m2|m1) ∈ CI(m2)

Ecuación 3. Reglas para clasificar el tipo de relación.

Donde **A, B, C y D:** Reglas para identificar la exclusión mutua, co-mutación bidireccional, co-mutación unidireccional e indefinición respectivamente. Las probabilidades condicionales se definen con los valores de a, b y c de la tabla de contingencia como sigue, P(m1|m2) = a/(a+b) y P(m2|m1) = a/(a+c). Los intervalos de confianza CI(m1) y CI(m2) se calculan al 95 %, usando la distribución binomial B(n, p) con n igual al total de muestras con las dos posiciones secuenciadas y “p” igual a la probabilidad de observar mutaciones en cada locus del par.

3.2.6. Características de las mutaciones *driver* y nuevos *driver* propuestos

Diferenciar entre las mutaciones *driver* y las *passenger* es importante para encontrar nuevos blancos terapéuticos. Se encontraron las siguientes 3 características capaces de distinguir entre un locus *driver* y uno *no-driver*.

3.2.6.1. Las mutaciones *driver* interactúan en más tipos de tumores

Se encontró un alto número de dependencias entre mutaciones *driver* en distintos tipos de cáncer, como son BRAF.V600 y KRAS.G12. Al aplicar pruebas estadísticas se obtuvo que las mutaciones *driver* tienen mayor conectividad (**Figura 15A**, más aristas en la red, prueba de Mann-Whitney p=2.17e-4) y están presentes en más tipos de cáncer (**Figura 15B**, prueba de Mann-Whitney p=2.8e-09) que las no clasificadas como *driver* (algunas serán *passenger* y otras *driver* aún no descubiertas como tal) y sorprendentemente muestran menor frecuencia mutacional (**Figura 15C**, prueba de Mann-Whitney p=2.13e-2)**.** Se esperaba que las posiciones más frecuentemente mutadas estuvieran más conectadas en un mayor número de tipos de cáncer, pero la frecuencia mutacional solo se encuentra parcialmente relacionada con la conectividad y el número de tipos de cáncer (coeficientes de Pearson: 0.54 y 0.48, prueba de Pearson p=2.2e-16 and, 7.63e-13, respectivamente), mientras que la conectividad está altamente correlacionada con el número de tipos de cáncer (coeficiente de Pearson: 0.86, prueba de Pearson p=2.2e-16).

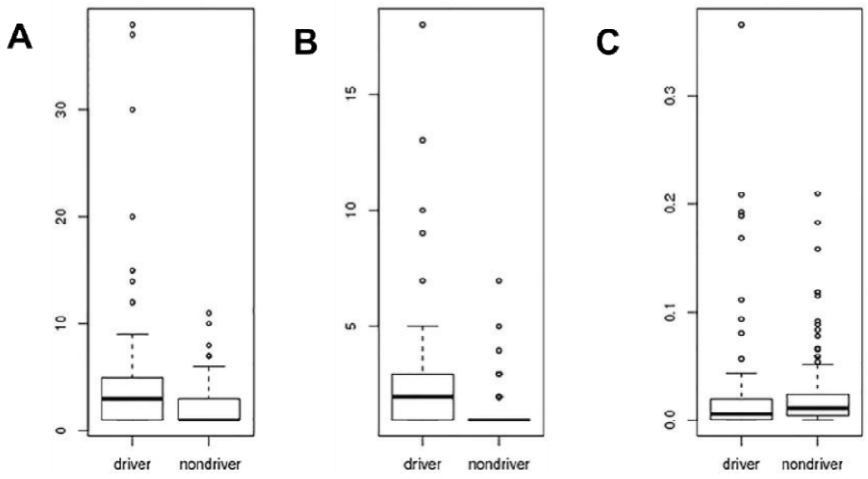


Figura 15. Distribución de mutaciones según la conectividad (A), número de tipos de cáncer (B) y frecuencia mutational (C).

3.2.6.2. Las mutaciones *driver* tienden a excluirse y las *no-driver* a co-mutar

Se clasificaron los pares de mutaciones atendiendo a la cantidad de mutaciones *driver* en el par en: las dos *no-driver*, un driver y una no-driver, y las dos *driver*. Se encontró que la composición del par (*driver* o *no-driver)* estaba asociada con el tipo de relación entre las mutaciones del par (prueba de Chi-cuadrado p<2.2e-16). Como se muestra en la **Figura 16**, la dependencia más común para pares de dos mutaciones *driver* es de exclusión en el 80.5% de los pares (128/159). En contraste, la dependencia entre los pares donde una sola mutación es *driver* tiende a ser de co-mutación en el 65.34% de los casos (115/176). Vale la pena señalar que algunas exclusiones conocidas aparecieron en nuestro análisis, como la exclusión mutua entre las mutaciones de KRAS, BRAF y NRAS en cáncer colorrectal[106].

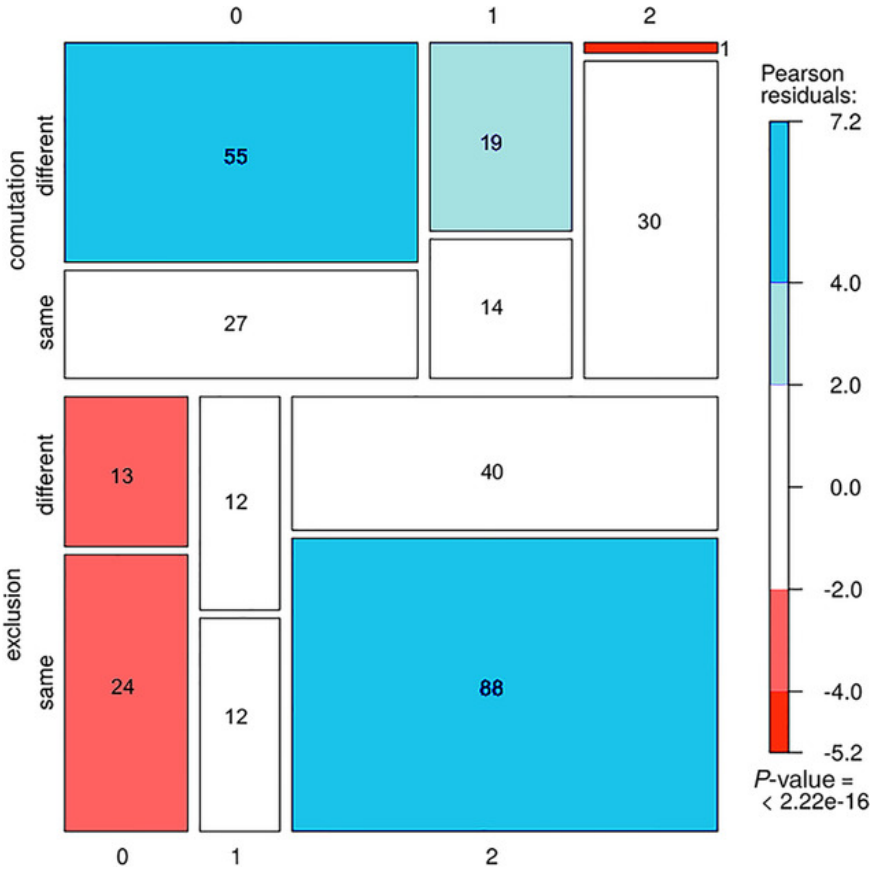


Figura 16. Clasificación de los pares de mutaciones por origen, relación y número de driver involucrados en el par.

Eje X, pares de mutaciones con 0, 1, o 2 mutaciones *driver*. Eje Y: relaciones (exclusión y co-mutación) diferencias por el origen (son o no de la misma proteína). El color indica la desviación de lo esperado bajo un modelo de independencia completa.

3.2.6.3. Pares de mutaciones *driver* tiende a ser de la misma proteína

Los únicos 31 pares de co-mutación donde ambas mutaciones son *driver* comprenden 25 posiciones de 7 proteínas en 12 tipos de cáncer. Solo un par donde ambas mutaciones son *driver* involucra a dos proteínas distintas en un cáncer, el par es KRAS.G12 + BRAF.V600. Cuando aplicamos la clasificación mostrada en la **Figura 16** se puede observar que la mayor cantidad de pares donde ambas mutaciones son *driver* tienden a estar en la misma proteína (96.77% de las co-mutaciones y 68.75% de las exclusiones) (prueba de Fisher p=7.95e-08; prueba de Chi-cuadrado p=1.13e-07).

3.2.7. Mapeo 3D de posibles *driver* en quinasas

Algunas mutaciones en quinasas poseían las 3 propiedades que caracterizan a las *driver* por lo que las proponemos como driver no descubiertos aún. Se analizan las mutaciones propuestas como *drivers* en quinasas inspeccionando su posición en un alineamiento múltiple de secuencia (MSA) de quinasas completas y se mapean en una estructura 3D de referencia. Se busca si se encuentran ubicadas en segmentos hipermutados (HS), donde se ha demostrado que las mutaciones *driver* se agrupan en varias quinasas[102]. Además se compara si la posición de la mutación propuesta es coincidente (igual columna en el alineamiento múltiple de secuencias -MSA-) con mutaciones *drivers* conocidas en posiciones equivalentes en proteínas homólogas a través de un MSA. Las mutaciones *drivers* conocidas en posición equivalente en otra quinasa y su mapeo 3D se obtienen de la base de datos del Kin-Driver[43].

3.3. Resultados

3.3.1. Red de las relaciones entre las mutaciones

Encontramos 189 pares de exclusión y 29300 de co-mutación que involucran a 492 mutaciones en 67 tipos de cáncer de 21 tejido de origen. Dado que 29155 pares de mutaciones (98.87%) son del tipo de cáncer carcinoma de células escamosas del tracto aero-digestivo superior (upper-SCC), este se estudió por separado y el próximo análisis se enfoca en el resto de los tipos de cáncer. Así obtenemos 334 pares relacionados de mutaciones, de ellos 145 son de co-mutación y 189 de exclusión, involucrando a un total de 200 mutaciones de 72 proteínas en 66 tipos de cáncer de 20 tejidos de origen. El reducido número de mutaciones resalta el hecho de que una mutación puede tener múltiples dependencias con otras y puede mutar repetidamente en distintos tipos de cáncer. Los pares se muestran en la red de la

Figura 17 y pueden ser explorados interactivamente en la web http://sdmn.leloir.org.ar/. Las mutaciones están representadas como los nodos y las aristas son las dependencias entre ellas. En la página web interactiva los usuarios pueden mirar una mutación y/o aplicar filtros para explorar determinado tumor o proteína, seleccionar las dependencias por su tipo (exclusión, co-mutación) y seleccionar las mutaciones por su tipo (*driver*, no-*driver* y *driver* sugeridos), además estos filtros se pueden combinar entre sí. También la **Figura 17** integra la probabilidad mutacional de cada mutación (representada por el tamaño de los nodos) y cuales de estas proteínas están en el *Cancer Gene Census* (CGC)[44].

Vale la pena señalar que hemos encontrado mutaciones dependientes en 72 genes de un conjunto de datos inicial de 1329 genes, lo que significa que los 1235 genes restantes tienen mutaciones que no se asocian significativamente con otros.

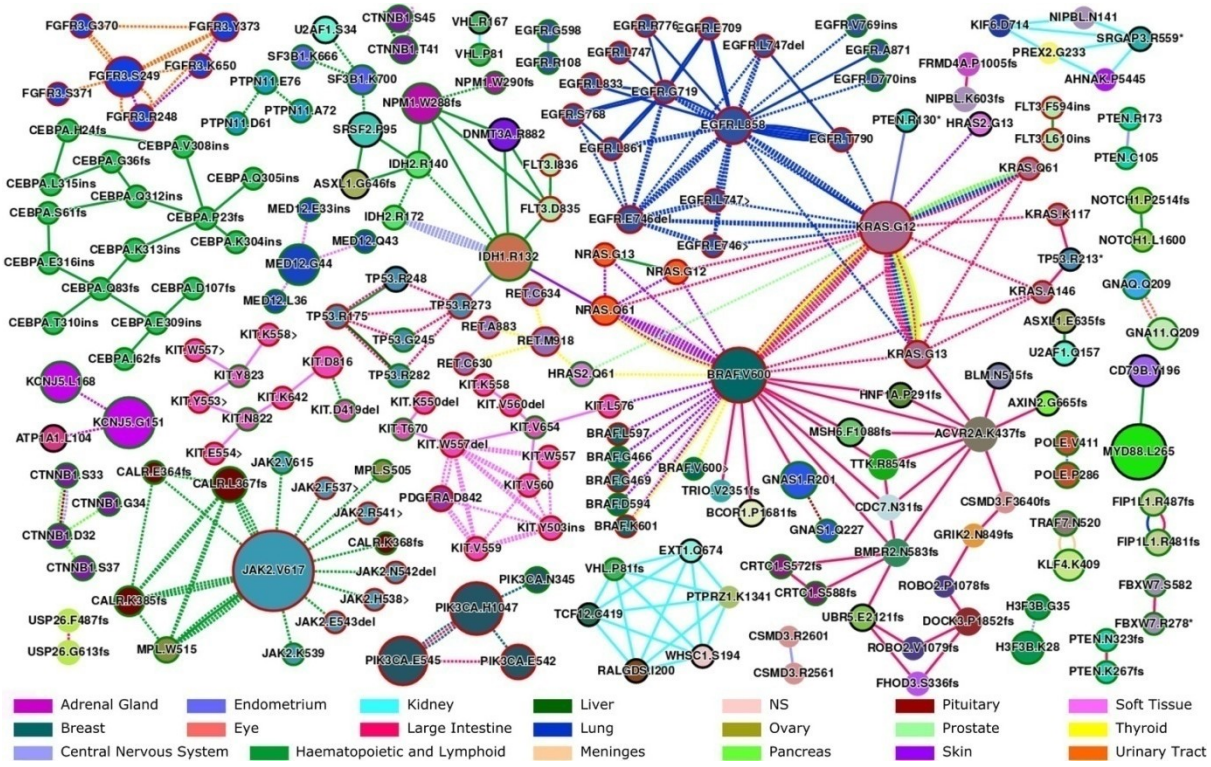


Figura 17. Red de mutaciones por tipo de cáncer.

Las mutaciones, son los círculos que representan los nodos de la red y su tamaño indica la frecuencia mutacional en COSMIC_v75; el color es por proteína (por ejemplo todas las mutaciones de BRAF están coloreados de verde oscuro). El contorno de color negro de los nodos representa que la proteína está en el CGC, rojo que está en el CGC y es *driver* y con contorno verde se representan los *driver* que sugerimos. Las aristas representan las relaciones entre pares de mutaciones, su color viene dado por el tejido de origen (por ejemplo, todas las dependencias

entre mutaciones de pulmón están coloreadas de azul oscuro); las relaciones de exclusión en línea punteada y de co-mutación en línea contínua. Web interactiva en http://sdmn.leloir.org.ar/. Dos nodos sólo están relacionados si la dependencia entre ellos es estadísticamente significativa.

3.3.2. Distribución de los pares de mutaciones

En la **Figura 17** se puede apreciar que algunos pares de mutaciones están relacionados en diferentes tipos de cáncer, son nodos conectados con más de una arista. También se puede apreciar que algunos pares de mutaciones sólo son relevantes en algunos tumores. Encontramos 45 pares repetidos en distintos tipos de cáncer, involucrando a 45 mutaciones de 20 proteínas, todas estas se encuentran reportadas en el CGC. De estos pares 33 fueron encontrados en diferentes tumores del mismo tejido de origen mientras que los otros 12 aparecen en tumores de distinto tejido de origen. En la mayoría de los 45 pares el tipo de relación es la misma es todos los tumores, pero en 6 de estos pares el tipo de dependencia es diferente entre los tumores.

3.3.3. Las mutaciones *driver* se pueden distinguir de las *no-driver*

Basado en las 3 propiedades de mutaciones *driver* obtenidas de nuestro estudio, proponemos 151 posibles nuevas mutaciones *driver* (**Anexo 2**), todos de genes presentes en el CGC. De ellos 15 están localizados en 4 proteínas quinasas a los cuales analizamos por estructura y alineamiento múltiple de secuencia. Utilizando la base de datos Kin-Driver encontramos que 9 de ellos se encuentran en regiones hiper-mutadas (-HS-) donde se ha visto que las mutaciones *driver* se agrupan[102], también 8/15 están en posiciones conocidas como *driver* en quinasas homólogas. Esto da fuerza a la sugerencia de su posible papel como *driver*.

3.3.4. Upper-SCC se encuentra mutado de una forma poco común

Upper-SCC es el nombre utilizado para designar una variedad de tipos de cáncer con diferentes etiologías, originados en el epitelio de cabeza-cuello[107]. Es uno de los tumores menos secuenciados en la base de datos de COSMIC_v75, con 12919 muestras en comparación con el adenocarcinoma colorrectal que tiene 145407 muestras, siendo este último el tipo de cáncer más secuenciado. En el estudio de la dependencia entre mutaciones encontramos que Upper-SCC tiene 29155 pares significativos siendo aproximadamente el 98.87% de todos los pares encontrados

por lo que se decidió estudiarlo por separado. Los pares significativos se pueden observar en la web http://sdmn.leloir.org.ar/ en el link inferior. Todos los pares en este tumor son del tipo co-mutación y se asocian en 292 posiciones de 42 proteínas. Solo 8 de las 42 proteins (PDE4DIP, NCOR1, HLA-A, NOTCH1, NOTCH2, BCORL1, KMT2C and SETBP1) está reportadas en el CGC, y estas proteínas forman 11061 pares de 65 posiciones en las proteínas. Hay que remarcar que no se encontraron pares de tipo exclusión y que ninguna mutación *driver* reportado con anterioridad está entre en los pares significativos de Uper-SCC.

3.4. Validación de los resultados

El uso de las nuevas tecnologías de secuenciación ha permitido obtener los perfiles mutacionales de miles de muestras de tumores, pero hasta ahora no se habían publicado estudios sistemáticos de co-mutaciones y/o exclusiones entre las mutaciones.

3.4.1. Pares conocidos de dependencias entre mutaciones

En nuestro estudio, encontramos algunos pares ya conocidos, que validan nuestros resultados, de exclusión tenemos a KRAS.G12/13 y BRAF.V600 en varios tumores[88]; y de co-mutación a EGFR.L858 y EGFR.T790 en adenocarcinoma de pulmón[108]. Sin embargo, hemos encontrado algunas dependencias controversiales ya que han sido reportadas como excluyentes en la mayoría de los tumores[109,110], mientras que para algunos tumores las hemos encontrado como co-mutación, dos ejemplos son: el par KRAS.G12 y KRAS.G13 en carcinoma anaplásico de tiroides, adenocarcinoma de próstata y en carcinoma papilar de tiroides y el par BRAF.V600 - KRAS.G12 en carcinoma anaplásico de tiroides. La heterogeneidad intratumoral puede ayudar a explicar estas coincidencias inesperadas, ya que pueden existir mutaciones mutuamente excluyentes en poblaciones subclonales dentro del mismo tumor[111], que se mezclan por secuenciación[112].

También se encontraron cliques de dependencias mutacionales en nuestro estudio. Un ejemplo es el clique de co-mutación formado por IDH1.R132, NPM1.W288fs, DNMT3A.R882 y FLT3.D835 en leucemia mieloide aguda. se han descrito que mutaciones en IDH1 - DNMT3A y DNMT3A - FLT3, se presentan juntas en un número significativo de muestras de pacientes con leucemia mieloide, y mutaciones concomitantes en la triada NPM1/DNMT3A/FLT3 han sido asociadas con una baja sobrevida en esta enfermedad[113]. Estos hallazgos experimentales respaldan nuestra hipótesis de una ventaja sinérgica de los cliques de co-mutaciones. Por otra parte,

los cliques de exclusión probablemente implican mutaciones que alteran la misma vía biológica. Una mutación en uno de ellos es suficiente para causar el desequilibrio cáncer y, por lo tanto, tiende a excluir a los demás[85].

3.4.2. Mutaciones *driver* de literatura no utilizadas en el estudio

Dadas las propiedades de las mutaciones *driver* que han emergido de este estudio propusimos 151 posibles nuevas mutaciones *driver* de genes presentes en el CGC. Uno de ellos, KIT.D419del, en efecto ha sido descrito como *driver*[114] aunque no fue considerado como tal en nuestro análisis. También se encontraron mutaciones de proteínas no quinasas que muestran las tres características que caractizan a las mutaciónes *driver*, como por ejemplo CALR.K385fs y CALR.L367fs que se encuentran en el exón 9 del gen CALR. Estas mutaciones han sido reportadas como excluyentes con otros cambios de marco de lectura en la misma proteína, y también con MPL.V515 y JAK2.V617 en neoplasias hematopoyéticas[115]. Todos los datos anteriores indican que CALR.K385fs y CALR.L367fs pueden ser *driver*, y son sugeridos como tal en nuestro estudio.

3.5. Importancia y aplicabilidad

El tratamiento del cáncer enfrenta varios desafíos, como la selección de agentes terapéuticos apropiados y la recaída a una enfermedad más agresiva después de un tratamiento inicialmente exitoso. Los pares y cliques de mutaciones que surgieron en nuestro análisis, particularmente aquellos que involucran mutaciones *driver*, podrían resultar útiles en este entorno. Dado que la mayoría de los nuevos fármacos antitumorales se dirigen específicamente a proteínas mutadas o genéticamente alteradas, la presencia de co-mutaciones sugiere que las combinaciones de terapias seleccionadas pueden ser más efectivas que una monoterapia, al menos en ciertos tumores.

Existen varios ejemplos, no solo en la literatura, sino también en la práctica clínica, en los que se ha demostrado que las terapias combinadas (ya sea con dos fármacos o con un agente multi-dirigido) son más efectivas que las monoterapias. Por ejemplo, los inhibidores de MEK y BRAF se administran habitualmente en combinación a pacientes con melanoma con mutaciones en BRAF[116]. En otro ejemplo reciente en el cáncer de pulmón de células no pequeñas, se ha informado que la combinación de dos agentes anti-EGFR con actividad diferente contra los hot-spot de EGFR (gefitinib y osimertinib) es clínicamente efectiva en pacientes que albergan una doble mutación T790M + C797S en el gen EGFR[117].

Por el contrario, las exclusiones pueden indicar que la utilidad de ciertas combinaciones de medicamentos no será beneficiosa en un porcentaje significativo de pacientes. Por ejemplo, encontramos que las mutaciones IDH1.R132 y TP53.R273 aparecen juntas en los gliomas. Las mutaciones IDH1.R132 son muy frecuentes en esta neoplasia maligna y los inhibidores de IDH1/2 mutado se encuentran actualmente en ensayos clínicos para demostrar su utilidad tanto como monoterapia o en combinación con terapias dirigidas a otras vías oncogénicas[118]. Los medicamentos cuya diana es TP53 mutado, también se están probando en ensayos clínicos[119], y nuestros resultados indican que podrían ser un socio apropiado para los inhibidores de IDH1 en un número significativo de gliomas. La coexistencia inesperada de BRAF.V600 y KRAS.G12 en el carcinoma anaplásico de tiroides, ya sea a través de una verdadera co-mutación o por heterogeneidad tumoral, indica que ambas mutaciones deben ser dianas y tratarse conjuntamente para evitar la selección de mutaciones subclonales asociadas con la resistencia[111].
La ausencia de pares también es de interés para las terapias dirigidas. Las mutaciones en PTEN que conducen a la pérdida de la expresión de proteínas se han relacionado con la resistencia a muchos agentes dirigidos, como los inhibidores de EGFR en cáncer de pulmón con EGFR mutado[120], anticuerpos anti-EGFR en cáncer colorrectal con KRAS/NRAS no mutado[121] o inhibidores de BRAF en melanoma con mutaciones en BRAF[122]. En nuestro análisis, las mutaciones PTEN no co-mutaron ni excluyeron significativamente con las mutaciones de EGFR, KRAS, NRAS o BRAF en la mayoría de los tipos de cáncer.

3.6. Limitaciones del estudio

Una limitación importante de nuestro estudio deriva del hecho de que los tumores sometidos a secuenciación de exoma completo o genoma completo son escasos, y solo se ha secuenciado un número limitado de genes en la gran mayoría de las muestras colectadas de COSMIC_v75. En consecuencia, este análisis probablemente ha perdido un número significativo de asociaciones entre mutaciones, simplemente porque ellas rara vez se han secuenciado simultáneamente. Esta limitación puede ayudar a explicar una correlación contraintuitiva encontrada durante nuestro análisis, a saber, que las posiciones más mutadas no son las más conectadas, **Figura 15C**. Por ejemplo, la mutación KCNJ5.G151 se encuentra casi tan frecuentemente en las muestras de la base de datos COSMIC como KRAS.G12, con valores de 0.18 y 0.19, respectivamente; pero

KCNJ5.G151 está significativamente en menos pares. La explicación radica en el hecho de que, si bien el *driver* omnipresente KRAS.G12 se ha secuenciado en 164511 muestras de todos los tipos de tumores, KCNJ5 solo se ha informado en 2671 muestras, la mayoría de ellas (81.2%) son en el tipo de tumor adenoma/aldosterone de las glándulas suprarrenales, un tipo de cáncer donde las mutaciones KCNJ5 son prevalentes[123].

3.7. Conclusiones del capítulo

Se realizó un análisis bioinformático exhaustivo específico por tipo de cáncer destinado a descubrir pares de co-mutación y exclusión entre las miles de mutaciones somáticas descritas en COSMIC. De este estudio obtuvimos 29489 pares de mutaciones con dependencias significativas en 21 tejidos de origen. La red de los pares se encuentra disponible en la Web http://sdmn.leloir.org.ar/. Esta red se encuentra dividida en dos redes una solo para el cáncer Upper-SCC y la otra red para el resto de los tumores. También se encontraron 3 reglas capaces de distinguir entre mutaciones *driver* y no-*driver*, que utilizamos para sugerir 151 posibles *driver*.

Los pares y redes de mutaciones asociadas que surgieron en nuestro análisis, particularmente aquellos que involucran mutaciones *driver*, podrían ser útiles no solo en estudios de biología del cáncer, sino también para la selección de terapias. Las dependencias de co-mutatación sugieren tratamientos combinados que podrían ser más efectivos que los enfoques de monoterapia. Por el contrario, las exclusiones pueden indicar que ciertas combinaciones de medicamentos probablemente no sean útiles en un porcentaje significativo de pacientes.

4. Resistencia adquirida a medicamentos

En este capítulo se muestra el estudio relacionado a las mutaciones con resistencia a medicamentos. Se analizaron las mutaciones y sus posiciones en la estructura de proteínas así como su distancia a la droga. Se obtuvieron relaciones de exclusión y co-mutación entre ellas utilizando el mismo procedimiento descrito en el capítulo de dependencia entre mutaciones.

4.1. Introducción

La resistencia a los medicamentos es un fenómeno bien conocido que se produce cuando las enfermedades se vuelven tolerantes a tratamientos farmacológicos[37]. Dado la importancia y los primeros estudios realizados sobre resistencia a medicamentos, la base de datos COSMIC incorpora las primera anotaciones sobre mutaciones resistentes a drogas en cáncer en la versión 77 en Mayo del 2016. Se trata de mutaciones somáticas que permiten que un tumor continúe creciendo a pesar de las terapias dirigidas. Las curaciones iniciales cubren 9 genes y 18 terapias farmacológicas, que detallan 226 mutaciones que impulsan la resistencia.

4.2. Materiales y Métodos

4.2.1. Procesamiento de datos

Se descargaron los datos de la versión 81 de Cosmic (disponible en mayo de 2017) relacionados con resistencia a medicamentos. Algunas muestras y mutaciones fueron eliminadas debido a inconsistencias que fueron: i) muestras duplicadas, muestras que tienen el mismo identificador ii) muestras donde la droga se especificaba como “Tyrosine kinase inhibitor - NS” ya que no se especifica la droga utilizada, pudiendo ser o no alguna de las existentes en la base de datos, iii) posiciones mutadas que tenían como notación p.? y p.?fs, porque se conoce la mutación en la proteína o es en una zona no codificante iv) proteínas presentes en menos de 5 muestras, y v) drogas para las que menos de 10 muestras presentan resistencia. El motivo principal de los filtros iv) y v) es que para encontrar un patrón debe ser generalizado y no específico de una muestra. Tras la limpieza de datos, se obtuvo un conjunto de 1675 muestras y 1817 mutaciones (119 mutaciones únicas) de 11 proteínas, 8 de las 11 proteínas son quinasas.

4.2.2. Alineamiento de estructuras y mapeo de mutaciones

Se mapean en la estructura 3D las posiciones de resistencia pertenecientes a cada proteína con su droga (en caso de estar co-cristalizada). En los casos de no encontrar la estructura con la droga se realizan alineamientos de estructuras con UCSF Chimera[124], donde un pdb es el de la proteína a analizar y el otros es una quinasa del mismo dominio pero con tiene la droga. Por otra parte, generar una estructura única con la información de todas las mutaciones de todas las quinasas, se alinearon las secuencias de las 8 quinasas y en cada posición del MSA donde alguna de las quinasas estaba mutada, contamos la cantidad de quinasas y las

muestras con esa posición mutada, estos valores se escribieron en el b-factor del aminoácido correspondiente en el pdb 3RT7 de FLT3.

4.3. Resultados

4.3.1. Descripción de la relación entre drogas y mutaciones

En la **Figura 18** se presenta una red entre las mutaciones en las muestras y las drogas asociadas a resistencia. Se puede apreciar que la proteína ABL1 cuenta con múltiples posiciones mutadas.

Al analizar las secuencias de estas proteínas y ver donde se encuentras las mutaciones obtuvimos que el 91.3% de las muestras (1529 muestras) presenta mutaciones en el dominio tirosina-quinasa, mientras que el 1.3% (21 muestras) tiene mutaciones en el dominio serina/treonina-quinasa y el 7.4% (125 muestras) en otros dominios. El 43% de las mutaciones en ABL1 (368/862 muestras) es entre los residuos 244 y 256, esta región es el conector entre el dominio SH2 y el quinasa.

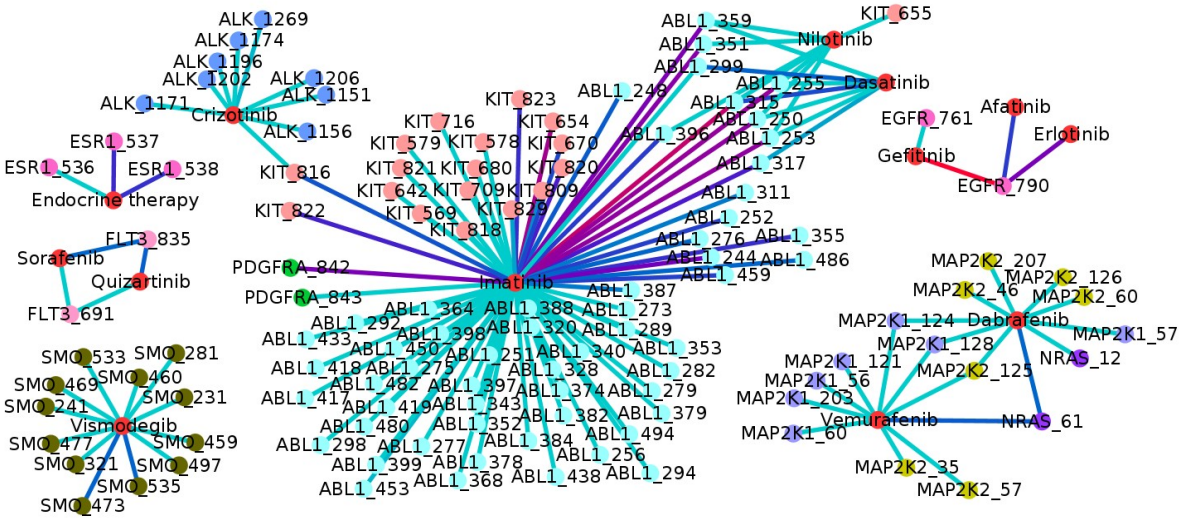


Figura 18. Red de drogas y mutaciones.

Los nodos de color rojo son drogas. Los nodos de otros colores son mutaciones, y el color es por el nombre de la proteína. Las aristas representan la conexión entre una muestra con la mutación y la droga. El color de las aristas representa la cantidad de muestras que presentan la relación, en un gradiente desde azul claro al violeta.

4.3.2. Mutaciones en las proteínas no quinasas

Tres de las proteínas no eran quinasas ellas son: SMO, ESR1 y NRAS, sus identificadores de Uniprot son: Q99835, P03372 y P01111, los dominios de PFAM donde se encuentras las mutaciones son: PF01534, PF00104 y PF00071

respectivamente. Las **Figuras 19A, B y C**, muestran las mutaciones mapeadas a la estructura de SMO, ESR1 y NRAS utilizando los PDB: 5L7I, 1GWQ y 5UHV respectivamente. Solo se pudo mapear la droga para SMO, ya que ESR1 y NRAS no contaban con estructura cristalizada con la droga, ni ninguna otra proteína de su misma familia.

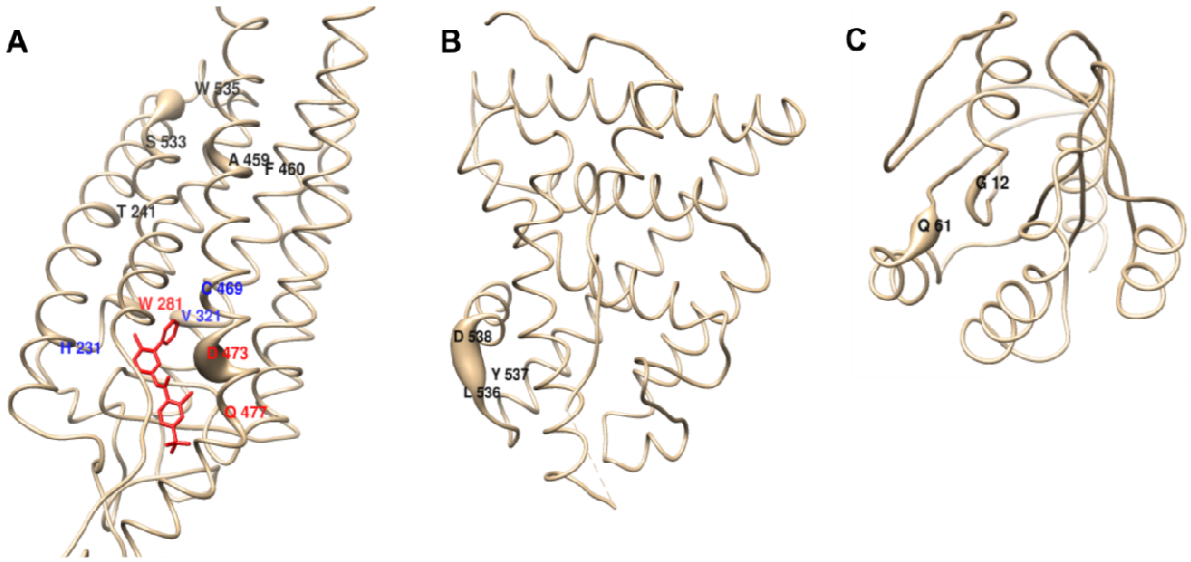


Figura 19. Posiciones asociada a resistencia en las 3 proteínas no quinasas.
A: SMO y vismodegib. **B:** ESR1 sin la droga. **C:** NRAS sin la droga. En rojo las posiciones a 6Å de la droga, en azul las posiciones a 8Å de la droga y en negro las posiciones a más de 8Å de la droga. El engrosamiento de la estructura representa la cantidad de muestras con la mutación.

4.3.3. Mutaciones en las proteínas quinasas

Se alinearon las estructuras de las 8 quinasas, se utilizaron PDB que tuvieran cristalizadas las proteínas con las drogas. Los PDB utilizados para cada proteína y las drogas se listan en la **Tabla 2**.

Proteina	Droga	PDB		Proteina	Droga	PDB
ABL1	dasatinib	2GQG[a]		FLT3	quizartinib	4RT7[a]
	imatinib	2HYY[a]			sorafenib	3WZE[b]
	nilotinib	3CS9[a]		KIT	imatinib	1T46[a]
ALK	crizotinib	5AAA[a]		MAP2K1	dabrafenib	5JRQ[b]
	afatinib	4G5J[d]			vemurafenib	5CSW[b]

EGFR	erlotinib	4HJO[d], 1M17[e]		MAP2K2	dabrafenib	5JRQ[b]
	gefitinib	4WKQ[d], 4I22[e]			vemurafenib	5CSW[b]
				PDGFRA	imatinib	5K5X[c]

Tabla 2. PDBs utilizados para el alineamiento de las proteínas con las drogas.
Donde [a] PDB de la quinasa con la droga, [b] PDB de una quinasa homóloga con la droga, [c] PDB sin la droga, [d] quinasa inactiva y [e] quinasa activa.

Se superpusieron las estructuras de las 8 quinasas con las drogas, y además se mapearon todas las posiciones mutadas al PDB 3RT7 de la proteína FLT3. En total se obtuvieron 82 posiciones mutadas en el alineamiento. En la **Figura 20A** se muestra engrosamiento en la estructura que representa la cantidad de muestras que tienen mutadas esa posición en el alineamiento (independientemente en que quinasa). En la **Figura 20B** se muestra engrosamiento en la estructura que representa la cantidad de quinasas que tienen muestras mutadas en esa posición en el alineamiento. Se puede apreciar que algunas posiciones están engrosadas en ambas figuras como son: F691 y D835.

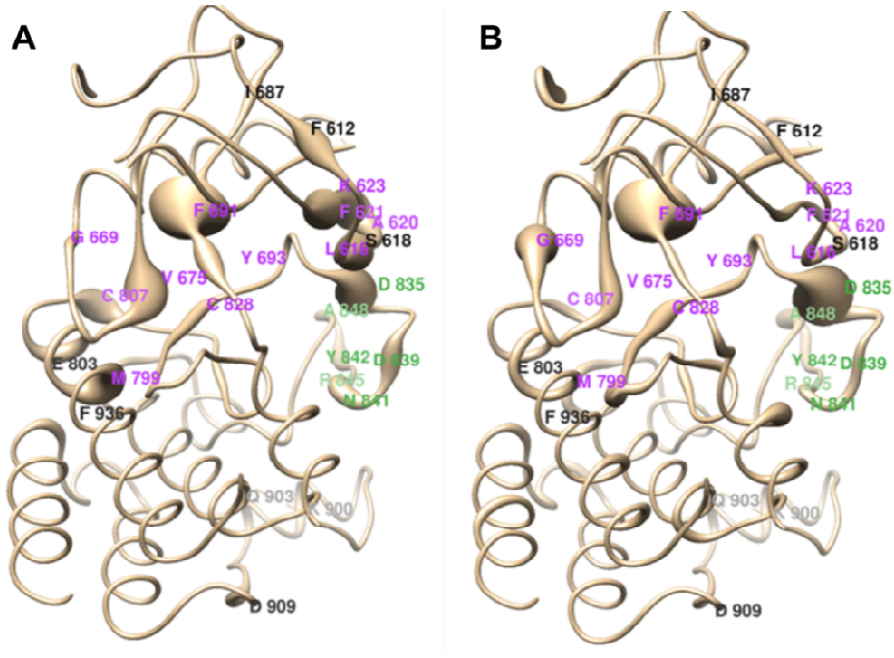


Figura 20. Mapeo de las posiciones de 8 quinasas.
A: Posiciones mapeadas respecto a la cantidad de muestras. **B:** Posiciones mapeadas respecto a la cantidad de quinasas.

4.3.3.1. Mutaciones de activación

Al analizar la relación entre las mutaciones de resistencia y las mutaciones en Kin-Driver (mutaciones *driver*), se obtienen 13 mutaciones en la intersección, otras 7 de ABL coinciden en superposición con *driver* en otras quinasas **(Figura 21A**). Obteniendo así que aproximadamente ⅓ de las mutaciones de resistencia (20/63) en quinasas podrían ser también de activación o *driver*. En las quinasas, la distribución de las mutaciones de resistencia coincide en gran medida con la de mutaciones de activación (**Figuras 21B y C**).

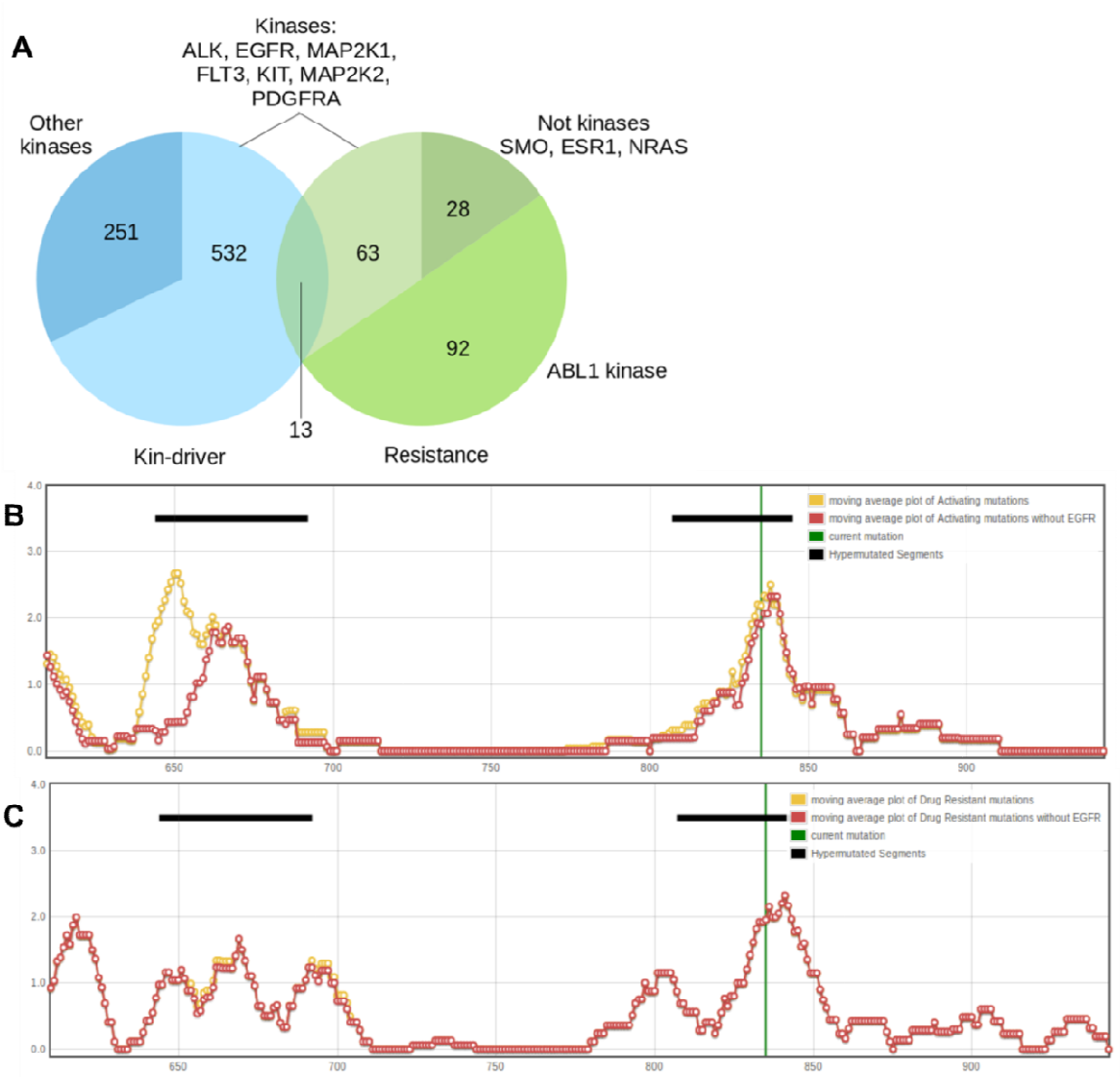


Figura 21. Mutaciones en regiones de activación de quinasas.

A: overlap entre mutaciones de resistencia y de activación. **B:** FLT3 en Kin-Driver, mutaciones en cáncer. **C:** FLT3 en Kin-Driver, mutaciones de resistencia. La línea verde indica la mutación D835Y.

4.3.4. Relaciones encontradas entre las mutaciones

De todas las muestras que presentan resistencia, recolectamos sus mutaciones de COSMIC. Utilizamos la metodología explicada en el capítulo anterior para encontrar relaciones de dependencia entre estas mutaciones. Se obtuvieron 31 pares de dependencias entre mutaciones donde 20 son exclusiones y otras 11 son co-mutaciones (**Figura 22**).

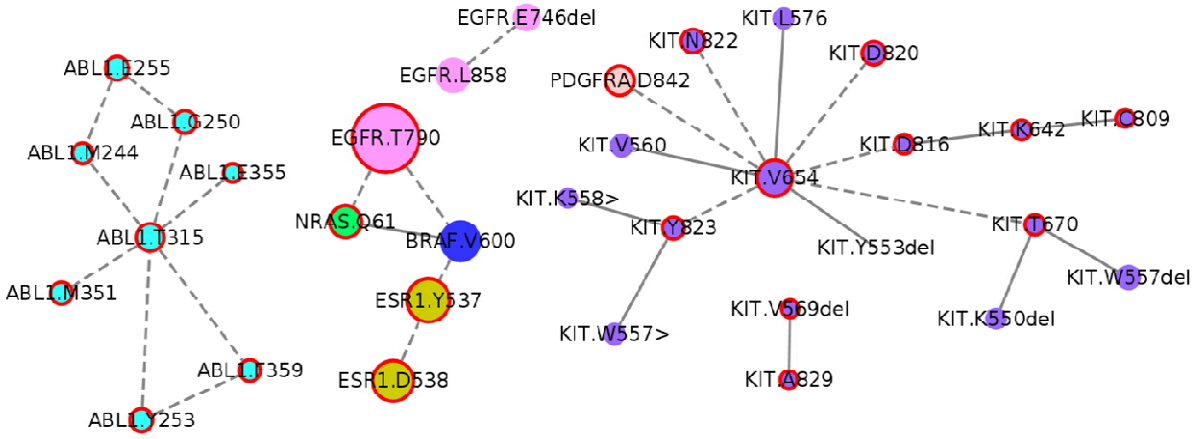


Figura 22. Relaciones encontradas entre las mutaciones.

Los nodos son las mutaciones, y su tamaño viene dado por su frecuencia mutacional; su color viene dado por la proteína. El contorno de color rojo significa que está en el CGC y es *driver*, sin contorno las *no-driver*. Las aristas representan las relaciones entre pares de mutaciones, las relaciones de exclusión en línea punteada y de co-mutación en línea contínua. Dos nodos sólo están relacionados si la dependencia entre ellos es estadísticamente significativa.

4.4. Limitaciones

La principal limitación obtenida durante el estudio fue la poca cantidad de muestras y datos de resistencia a drogas. Este inconveniente atenta contra los resultados y la credibilidad estadística.

4.5. Conclusiones del capítulo

En las quinasas, las posiciones con resistencia en más de 5 muestras (26 posiciones) por lo general se encuentran a menos de 8Å de la droga (11/26, ~42.3%) o en el lazo de activación (6/26, ~23%). Además en las quinasas estudiadas, alrededor de ⅓ de las mutaciones de resistencia también están clasificadas como mutaciones de activación.

Combinando pruebas de probabilidades mutacionales condicionales y esperadas, se encontraron parejas y redes de co-mutaciones y exclusiones, algunas de ellas en tipos particulares de cáncer y otras generalizadas. Ejemplo de esto son las mutaciones BRAF.V600 y NRAS.Q61 se presentan como co-mutaciones en pacientes con resistencia a drogas.

Sin embargo, dada la baja cantidad de datos no nos permite generalizar los resultados y hemos decidido abandonar esta línea de trabajo hasta tanto haya más datos disponibles.

5. Carga mutacional total

En el siguiente capítulo se analiza la carga mutacional total (TMB por su sigla en inglés: *Tumor Mutation Burden*) en muestras de 42 tipos de cáncer provenientes de la base de datos COSMIC. Se realizan modelos de regresión lineal específicos por tumor para predecir el TMB. Se analiza la correlación entre TMB y respuesta a inmunoterapia en dos tipos de cáncer: melanoma y carcinoma de pulmón de células no pequeñas (NSCLC). Este estudio se encuentra en proceso de publicación.

5.1. Introducción

Actualmente uno de los marcadores más estudiados para el tratamiento con inmunoterapia es el número total de mutaciones exónicas somáticas en un tumor, que se denomina TMB. Es probable que los tumores con un alto TMB generen más neoantígenos y así ser reconocidos por el sistema inmune[125–128]. Como consecuencia de este aumento de la inmunogenicidad, los tumores con TMB alto deberían mostrar mejores respuestas a la inmunoterapia.

Un número significativo de estudios retrospectivos, algunos de ellos realizados en muestras tumorales de ensayos clínicos, han encontrado una asociación entre TMB y la respuesta a inmunoterapia[129–132], aunque también se han informado algunos resultados negativos[133,134]. Además, los ensayos clínicos recientes de inmunoterapias que utilizan TMB para la selección prospectiva de pacientes han arrojado resultados contradictorios[135]. Actualmente se están empleando paneles de genes (llamados de NGS por su sigla del inglés Next Generation Sequencing) para estimar TMB, de alrededor de 1 MB (3% del exoma total). Dos ejemplos son los paneles Foundation One[136] y MSK-Impact[137], que cubren los exomas de 348 y 468 genes relacionados con el cáncer y se usan en todo tipo de tumores sólidos. Aunque estos paneles han demostrado cierto poder predictivo para el beneficio clínico de inmunoterapia[125,130,137,138], fueron diseñados para detectar mutaciones *driver* y otras alteraciones genéticas biológicamente relevantes, no para calcular TMB; y se ha informado que clasifican erróneamente un número significativo de muestras[139]. Además, los paneles son únicos e ignoran el hecho de que diferentes tumores pueden tener genes muy diferentes mutados. Hasta la fecha, nunca se han aplicado herramientas bioinformáticas para un diseño racional de paneles de genes para extrapolar con precisión el TMB.

5.2. Materiales y Métodos

La **Figura 23**, muestra el esquema de trabajo de este estudio. Se encuentra dividido en 3 etapas. La primera etapa, destinada al filtrado de datos y selección de genes y exones a utilizar por cada estrategia. La segunda etapa, relacionada con la creación de paneles y sus modelos asociados se descompone en 2 partes, la creación de los paneles/modelos y la selección de paneles/modelos como consensos de los iniciales. La tercera etapa es la validación tanto interna como externa. En esta última etapa también se seleccionan un conjunto de paneles y sus modelos asociados atendiendo a la cantidad de megabases a secuenciar para evaluarlos con datos de

inmunoterapia. Estos últimos paneles/modelos son seleccionados como representativos para predecir el TMB.

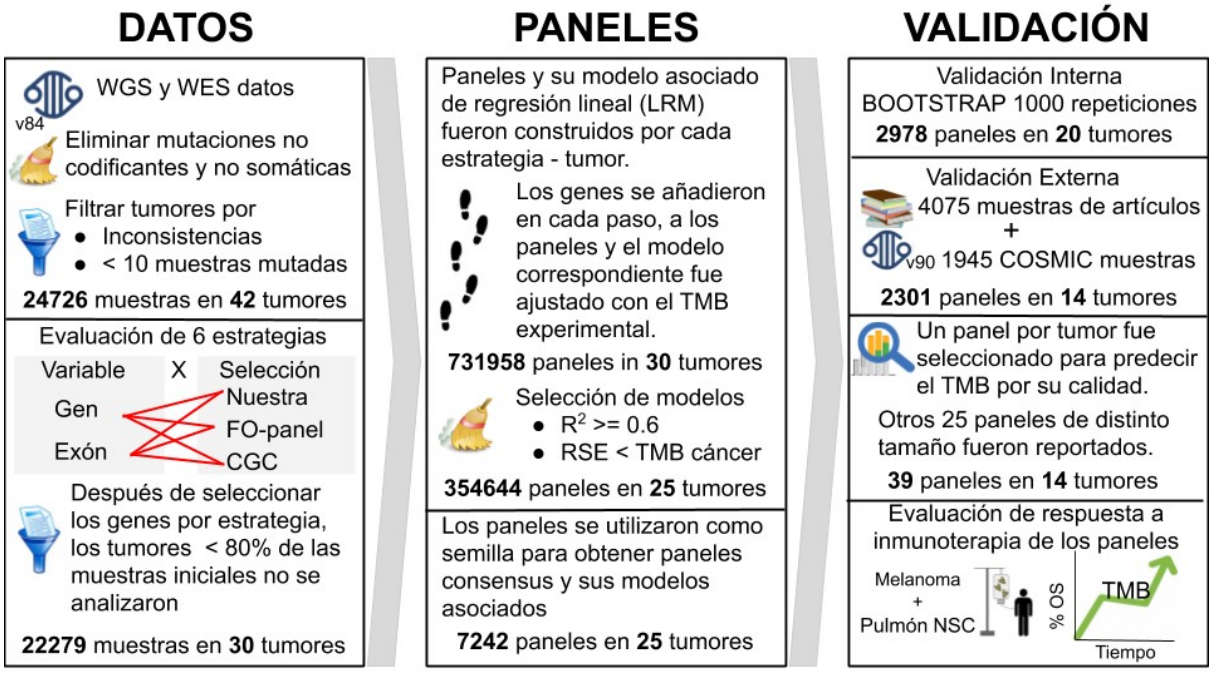


Figura 23. Esquema de trabajo para predicción de TMB.

5.2.1. Procesamiento de datos

Se dividió el conjunto de datos en: de entrenamiento, de validación externa y de relación con la inmunoterapia

5.2.1.1. Conjunto de datos de entrenamiento

Descargamos las muestras que cuentan con secuenciación del genoma completo (WGS) y exón completo (WES) de la versión 84 de COSMIC (Cosmic_v84)[44] para el ensamblaje humano grch38. El set de datos consta de 25533 muestras tumorales clasificadas en tipos de cáncer según el subtipo de sitio primario, histología primaria e histología. Luego, excluimos: i) tumores benignos, ii) muestras con mutaciones sin anotación de coordenadas genómicas, iii) muestras con mutaciones no somáticas (etiquetadas como *Variant of unknown origin* y *Not specified*), iv) muestras con mutaciones exclusivamente en regiones no codificantes y v) se excluyeron los tipos de cáncer con menos de 10 muestras.

Luego de la limpieza de los datos, contamos con 24726 muestras distribuidas en 42 tipos de cáncer para el desarrollo del método.

5.2.1.2. Conjunto de datos de validación externa

El conjunto de datos de validación externa contiene las nuevas muestras WGS y WES agregadas desde la versión 84 (datos usados para entrenamiento) hasta la versión 90 de COSMIC (Cosmic_v90) (última versión al momento del análisis), siendo 3144 muestras. Adicionalmente se recuperaron 4773 muestras con datos WGS y WES de 133 artículos disponibles públicamente en la literatura. Todas las coordenadas genómicas de las muestras se mapean al genoma humano grch38 y se filtraron de acuerdo con el procedimiento descrito para el conjunto de datos de entrenamiento. En general, el conjunto de datos de validación externa contiene 7917 muestras de 40 tipos de cáncer. No se encontraron muestras de los cánceres paratiroideo y pituitario.

5.2.1.3. Conjunto de datos de respuesta a inmunoterapia

Entre los datos de validación externa hay datos de respuesta a inmunoterapia de 174 muestras de melanoma[140,141] y 35 de NSCLC[125]. De las muestras de melanoma contamos con los datos de supervivencia total (OS por *overall survival*) y supervivencia libre de progresión (PFS por *free progression survival*) y respuesta clinica a inmunoterapia (CB por *clinical benefit*). De los datos de pulmón solo contamos con datos de PFS y CB.

5.2.2. Análisis de los genes más mutados

Los genes se ordenaron por el número de muestras mutadas en cada tipo de cáncer. Después se seleccionaron los 10 genes más mutados por tipo de cáncer, obteniendo un total de 157 genes distintos entre todos los tumores. Algunos de estos genes son TP53, TTN, MUC16, KRAS, PIK3CA, SYNE1, APC and OBSCN (**Figura 24**). Entre ellos, TTN y MUC16 tienen 34350 y 14507 residuos respectivamente y se describieron en la literatura como genes muy frecuentemente mutados[142–145]. Marouf C et. al[146], afirmó que el papel de TTN como gen canceroso es actualmente una predicción matemática y requerirá una evaluación biológica directa, ya que su mutabilidad podría ser debida simplemente a su longitud. En estudios previos se penalizaron los genes por su longitud y así ponderaron su rol[147]. Por estas razones, decidimos también aplicar una penalización a los genes por su longitud.

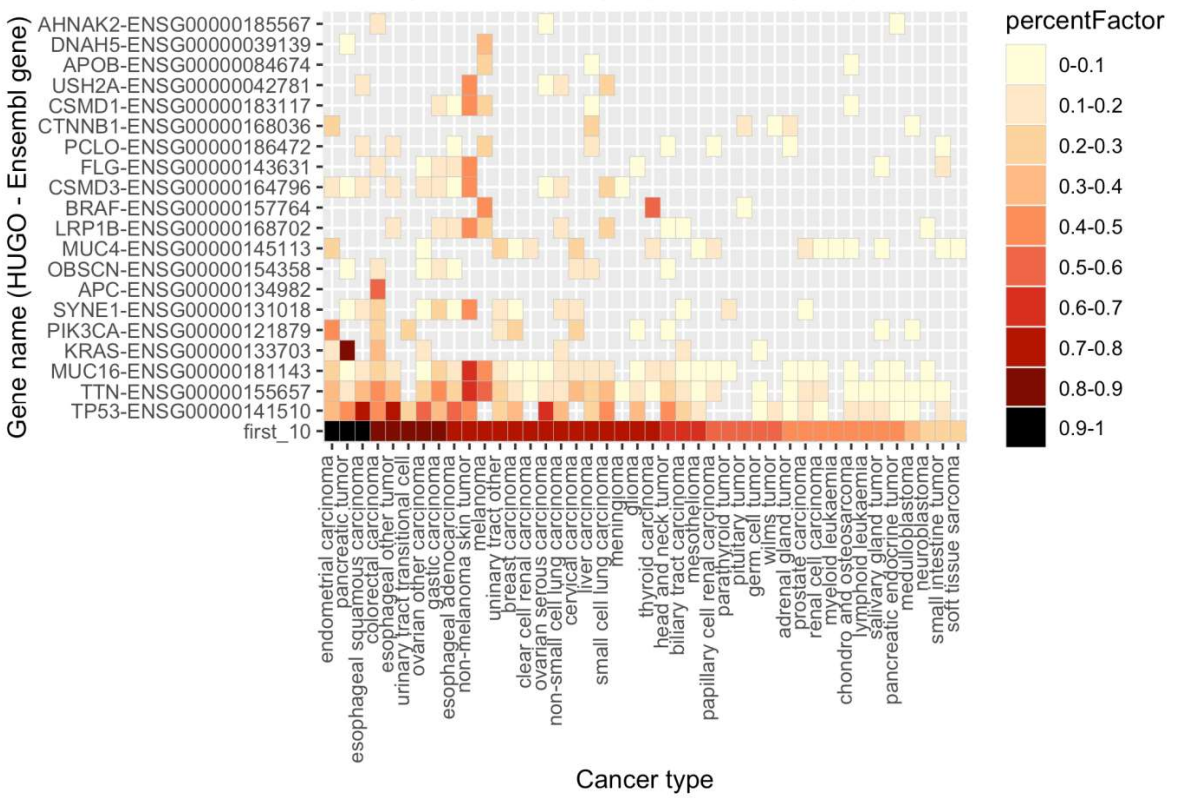


Figura 24. Genes más mutados por tumor.

Se muestran los primeros 20 genes más mutados (de un total de 157) en el top 10 por tipo de cáncer.

5.2.2.1. Penalizar los genes por su longitud

Para penalizar los genes por su longitud, se descargó el ensamblaje del genoma humano grch38 que contiene 20291 genes codificantes ("datos gtf" de ensembl, versión grch38.91, https://www.ensembl.org/index.html). Para cada gen, se calculó el número de bases codificantes. La distribución de longitud de los genes se muestra en la **Tabla 3** y la **Figura 25** .

Min.	25% (1er Q)	Mediana (50%)	Promedio	75% (3er Q)	95%	99%	99.95%	Max.
8	828	1341	1780	2139	4559.5	8460	18928.24	114595

Tabla 3. Pares de bases codificantes en los 20291 genes del genoma humano.

Algunos genes tienen una alta cantidad de pares de bases (bp) codificantes. Por ejemplo: TTN, MUC16 y DST con 114595, 43538 y 30580 pb respectivamente. En estadística, en toda distribución los datos ubicados a más de 3 veces la distancia intercuartil (IQ), se consideran valores atípicos extremos[148]. En el genoma humano,

los genes más largos que este umbral (2139 + 3 * (2139 - 828) = 6072 pb) son 499 (2,46% del genoma), se muestran en color rojo en la **Figura 25**.

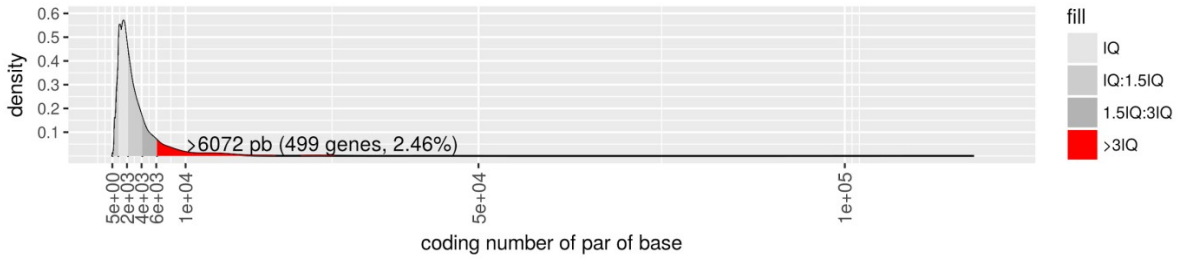


Figura 25. Distribución de genes atendiendo a su longitud.
En rojo el conjunto de genes que se encuentran a más de 3 veces la distancia intercuartil, representada con más de 6072 pb y representan 499 genes del genoma humano.

Para evitar el sesgo debido a la longitud del gen, el número de muestras mutadas por gen por tipo de cáncer se pondera como se muestra en la **Ecuación 3**. Por ejemplo, el gen ENSG00000143631 -FLG- tiene 12186 bp que codifican y 379 muestras mutadas en el cáncer colorrectal, aplicando la ecuación, quedarían 188 muestras (= 379 / (12186/6072)). Es decir, es como si tuviera 188 muestras mutadas en nuestro análisis.

$$\boldsymbol{muestras_{penalizadas(G,C)} = \frac{muestras\ en\ C}{max(1, \frac{pb\ codificantes\ de\ G}{6072})}}$$

Ecuación 4. Cantidad de muestras en el gen por tipo de cáncer después de la penalización utilizando la longitud del gen de pares de bases.
Donde “G” es el gen y “C” es el tipo de cáncer. No se penalizan los genes con menos de 6072 pb.

La **Figura 26** muestra los primeros 10 genes por tipo de cáncer después de la penalización, los más comunes son TP53, KRAS, PIK3CA, BRAF, APC, etc. Después de la penalización, los genes más largos no se encuentran entre los 10 genes principalmente mutados, al igual que en Hua Tan et.al[147].

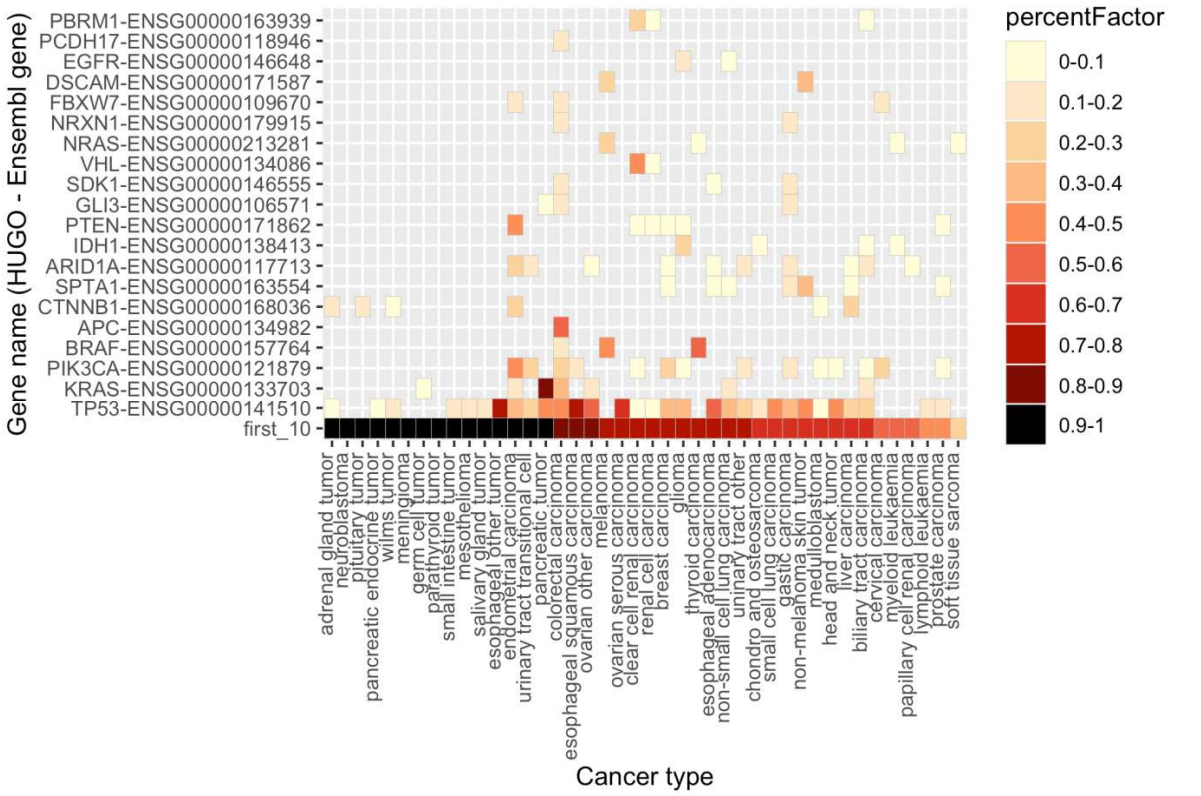


Figura 26. Genes más mutados por tumor después de la penalización por longitud.

Se muestran los primeros 20 genes más mutados (de un total de 196) en el top 10 por tipo de cáncer.

5.2.3. Estrategias a comparar

El TMB se definió como el número total de mutaciones en la parte exónica en una muestra dada. Nuestro objetivo fue seleccionar paneles específicos para cada tipo de cáncer, con un número limitado de genes o exones, que permita predecir el TMB usando modelos matemáticos apropiados. Decidimos construir paneles con 3 conjuntos de genes, atendiendo a su procedencia: nuestra selección, genes del Cancer Gene Census[149] (CGC) y genes del panel de FoundationOne[136] (FO-panel). De esta manera construiremos paneles con los genes que emerjan del análisis de datos (*our-gene* selection) que *a priori* no sabemos cuales serán ni la longitud en pb del panel. Otros paneles serán construidos a partir de los 719 genes del Census (CGC) y con los 314 genes del panel FO de aproximadamente 1.0Mb de longitud. Además generamos paneles utilizando todos los exones de estos 3 conjuntos de genes como variables independientes. Obteniendo así 6 estrategias para comparar, combinación entre 2 tipos de variables (genes y exones) y 3 tipos de conjunto de datos: nuestra selección, CGC y FO-panel.

5.2.3.1. Nuestra estrategia

Para obtener el conjunto de genes que formarán parte de nuestra estrategia de genes (*our-gene*), primero se excluyeron los 11 genes más largos (0.05% de los genes, tienen más de 18928 pb), ya que algunos de ellos han sido excluidos en estudios anteriores. Luego se eliminan algunos de los genes poco frecuentemente mutados por tipo de cáncer (genes que contienen menos del 1% de la muestras mutadas). Por lo que entonces la selección *our-gene* tiene los genes donde la cantidad de muestras penalizadas (obtenidas con la **Ecuación 3**) sea mayor a 5 o al menos un 10% de la muestras iniciales en dicho tumor.

Algo similar se realiza para obtener el conjunto de exones para la estrategia *our-exon*. El conjunto *our-exon* inicialmente contiene todos los exones de los genes de la selección *our-gene*. Para penalizar las muestras mutadas en el exón por tumor, se usó la **Ecuación 4**. Esta ecuación representa el número más pequeño de muestras mutadas (penalizadas por la longitud del gen) que deberían tener en el exón por su longitud. Por último la estrategia de *our-exon* solo contiene los exones donde la cantidad de muestras mutadas sea mínimo 2 y además sea mayor a la cantidad mínima penalizada del exón.

$$\boldsymbol{muestras\_penalizadas(E,G,C)} = \frac{\boldsymbol{muestras\_penalizadas(G,C)}}{\frac{\boldsymbol{pb\ codificantes\ de\ G}}{\boldsymbol{pb\ codificantes\ de\ E}}}$$

Ecuación 5. El menor número de muestras mutadas en el exón por tipo de cáncer después de la penalización.

Donde “E” es el exón, “G” es el gen del exón, “C” es el tipo de cáncer.

5.2.3.2. Estrategias usando los genes del CGC y FO-panel

Se descargó la lista de genes del CGC de la versión Cosmic_v84, obteniendo 719 genes, siendo este conjunto la estrategia *CGC-gene*. La estrategia resultante de los exones presentes en *GCG-gene* se llamó *CGC-exon*. El panel de Foundation One realiza la secuencia completa de 315 genes. Uno de los genes (TERC) no tiene parte codificante, por lo que se usaron 314 genes como la estrategia llamada *FO-panel-gene*. La estrategia llamada *FO-panel-exon* contiene los exones involucrados en estos 314 genes.

5.2.4. Paneles y modelos de regresión lineal

A cada combinación entre tipo de cáncer y estrategia que cumplieran el umbral de al menos el 80% de las muestras iniciales, se le calcularon modelos de regresión lineal. Luego, a partir de los modelos, se hicieron modelos de consenso para obtener posibles subconjuntos que participaran como nucleadores. Para finalizar, a los modelos consenso se les realiza una validación interna y externa, obteniendo así los mejores modelos.

5.2.4.1. Modelos de regresión lineal

A cada combinación entre tipo de cáncer y estrategia que cumplieran el umbral de al menos el 80% de las muestras iniciales, se le aplicó el algoritmo *ForwardStep* para seleccionar paneles y generar el modelo de regresión lineal asociado al panel. En este caso, en el modelo de regresión lineal, la variable dependiente es el TMB (lo que se desea estimar) y las variables independientes son los paneles de genes/exones. Para medir la calidad de los modelos obtenidos se utilizaron 4 evaluadores, de ellos 3 son algunos de los más utilizados para medir la calidad de regresiones lineales: R^2, R^2 ajustado y RSE; el cuarto fue la suma de la penalización de todos los genes o exones involucrados en el modelo.

La **Figura 27**, es un pseudocódigo del algoritmo. En cada paso del algoritmo, se utiliza un panel como semilla, el cual es extendido y genera múltiples nuevos paneles, los cuales se diferencian entre ellos en una sola variable (genes o exones) que se les incorporó, el panel semilla inicial es un panel vacío que no contiene ninguna variable. Cada panel semilla se extiende para generar nuevos paneles. A cada nuevo panel se les realiza su correspondiente modelo de regresión lineal. Haciendo uso de los coeficientes obtenidos para cada variable en los modelos se seleccionan los modelos factibles, que son los modelos donde todos los coeficientes de las variables tiene significación estadística para la regresión. Debido a que la cantidad de modelos puede ser muy grande, se seleccionan como máximo los primeros 100 mejores modelos mejores que el modelo que les dió origen (panel semilla) con los 4 evaluadores y los paneles asociados a estos modelos son utilizados como semillas para el siguiente paso del algoritmo. Tras terminar de expandir los modelos de un paso, se verifica si hay modelos guardados para el siguiente paso, de ser así continúa el algoritmo. El algoritmo termina cuando no se puede agregar más variables al panel semilla, sin mejorar la calidad del modelo asociado.

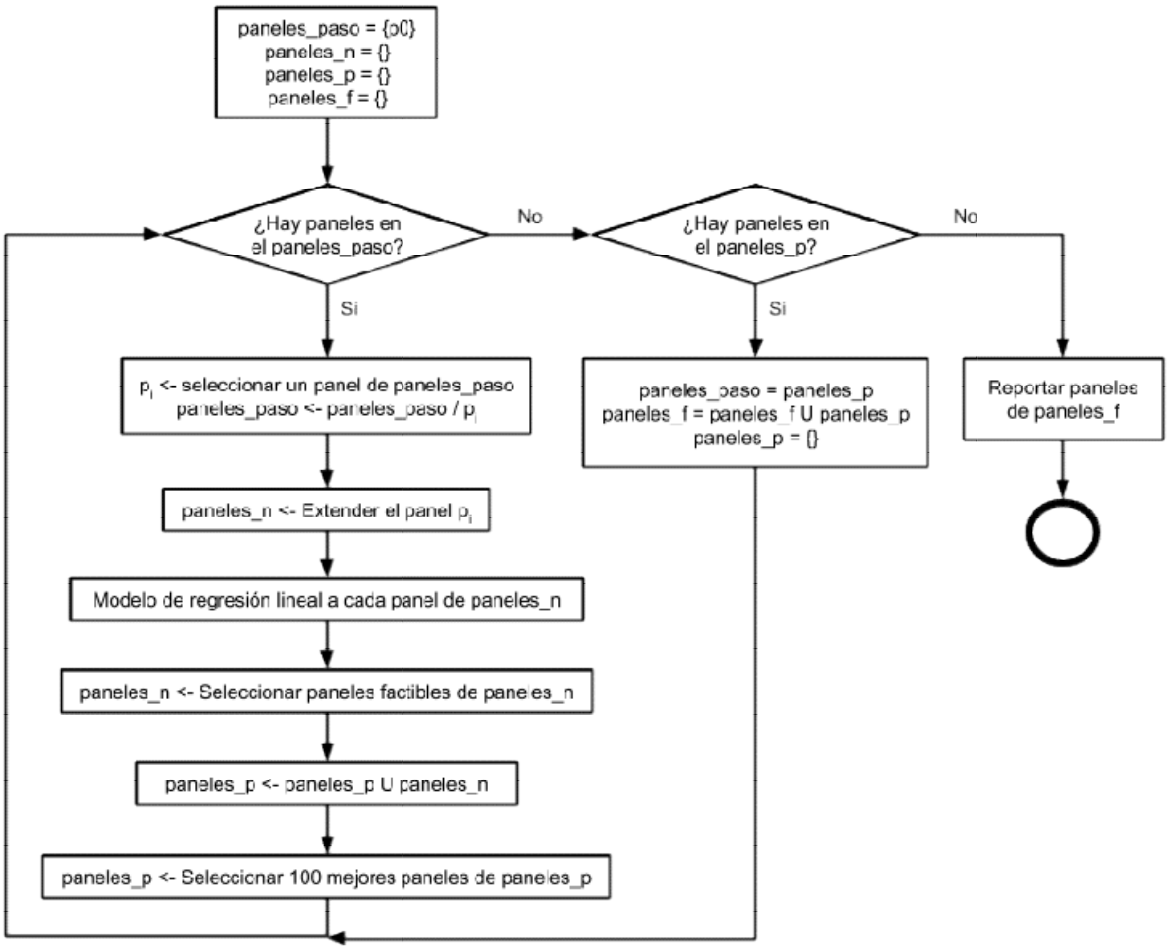


Figura 27. Pseudocódigo del algoritmo ForwardStep utilizado para generar los paneles y modelos de regresión lineal asociados.

Al aplicar este algoritmo a cada tipo de cáncer y estrategia que cumplieran el umbral, se generó un total de 371958 modelos entre los 30 tipos de cáncer y las 6 estrategias. Al ser modelos provenientes de todos los pasos del algoritmo algunos no cumplen con los requisitos de considerarse aceptables. Se eliminaron los modelos que tienen R2 < 0.6 y RSE > mediana de TMB en el cáncer, quedando 354644 modelos de 25 tipos de cáncer.

5.2.4.2. Paneles y modelos consenso

Los paneles y sus modelos asociados considerados como aceptables tienen muchos genes/exones en común por lo que se decidió realizar modelos de consenso. Con los modelos consensuados se eliminan genes/exones poco frecuentes en los modelos que hacen que sean muy específicos. Los modelos de consenso se generaron utilizando como semilla los modelos anteriores, utilizando el pseudocódigo de la **Figura 28**. Se generaron un total de 7242 modelos entre los 25 tipos de cáncer.

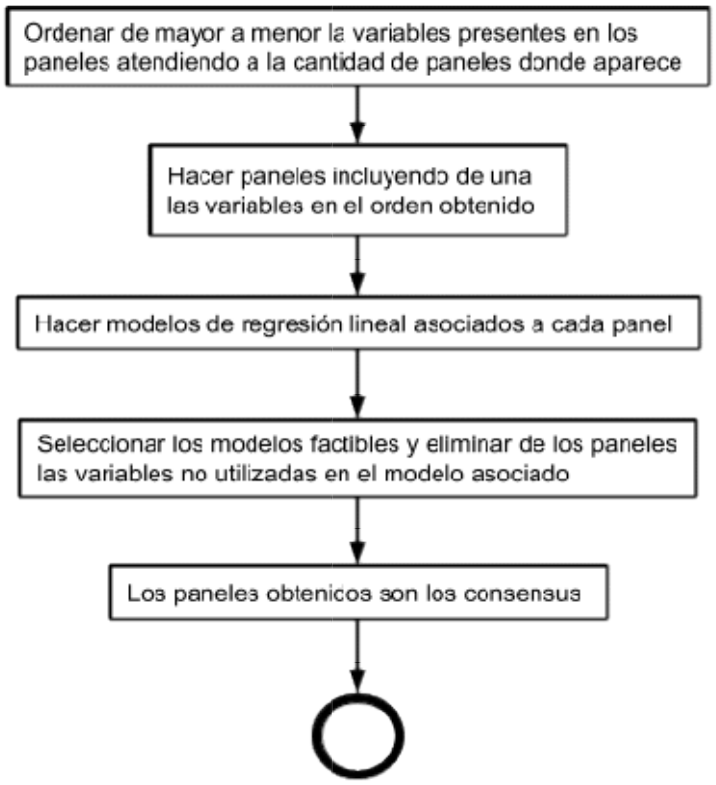


Figura 28. Pseudocódigo para generar los paneles consensus y modelos de regresión lineal asociados.

A estos modelos consensuados se les realizó una validación interna que consistió en 1000 repeticiones de validación o *bootstrap*. Además se les realizó una validación externa con datos provenientes de una nueva versión de COSMIC y de literatura. Se consideran modelos aceptables a los modelos que tienen las métricas R^2 y RSE de la validación interna similares a dichas métricas obtenidas por los modelos. Las reglas y la cantidad de modelos y tipos de cáncer que la pasan se describen en la **Tabla 4**. Terminamos con un total de 2301 modelos de 14 tipos de cáncer que pasaron tanto la validación interna como externa con una calidad aceptable.

Condiciones para aceptar modelos		**Modelos Consensuados**	**Tipos de Cáncer**
		7242	25
i)	R^2 modelo >= 0.6	6988	25
ii)	RSE modelo =< mediana de TMB	6459	25
iii)	R^2 validación interna >= 0.6 & R^2 validación interna >= R^2 modelo - 0.1	5556	22
iv	RSE validación externa =< mediana de TMB &	2978	20

)	RSE validación externa =< 1.25*RSE model		
v)	R^2 validación externa >= 0.6	2301	14

Tabla 4. Condiciones para aceptar los modelos consenso, y la cantidad de modelos y tipos de cáncer que cumplen las condiciones.

5.3. Resultados

5.3.1. Carga mutacional en distintos tumores

La **Figura 29** muestra la distribución de TMB en nuestro conjunto de datos de entrenamiento de 24726 muestras tumorales en 42 tipos diferentes de cáncer. Se observaron diferencias notables en la mediana de TMB, que muestra una variación >100 veces (rango 3 - 502). Se puede apreciar que los tipos de cáncer con mayor media de TMB son: los de piel (melanoma y no melanoma), a los que le siguen: carcinoma colorrectal, carcinoma de esófago, carcinoma de cérvix, carcinoma gástrico y tumores del tracto urinario, todos con más de 80 mutaciones como mediana entre sus muestras. Aunque algunos tipos de tumores como: carcinoma de piel, melanoma, colorrectal o carcinoma de hígado tienen una mediana de más de 30 genes mutados comunes en todas las muestras, la mayoría de ellos varían de 1 a 10 (**Figura 29**, datos en color azúl).

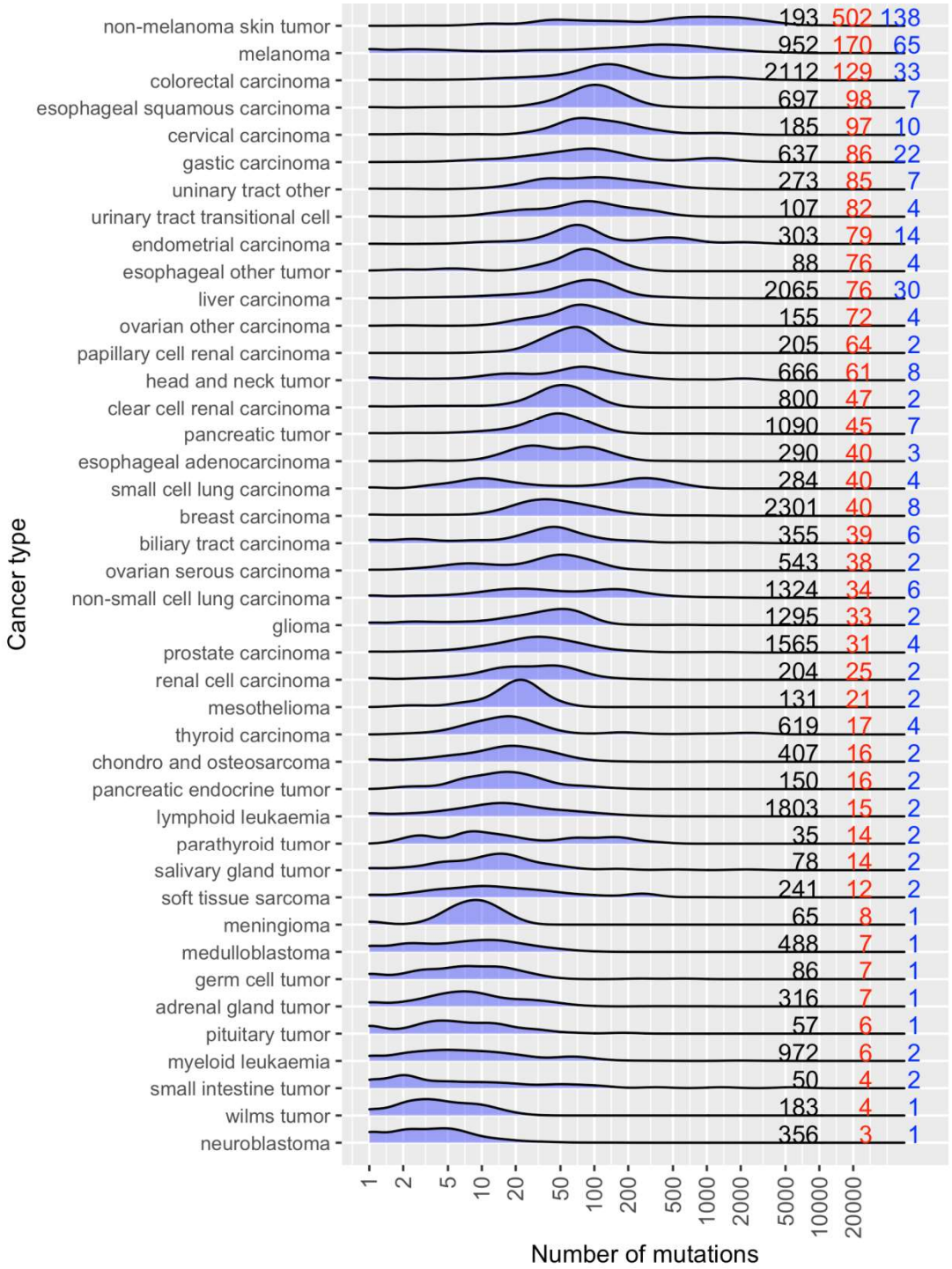


Figura 29. Distribución de TMB en 42 tipos de cáncer en las 24726 muestras del conjunto de datos de entrenamiento.

El eje X corresponde al TMB, el eje Y a la frecuencia de las muestras con mutaciones. Los números negros a la derecha del gráfico indican el número de muestras por tipo de tumor incluidas en el análisis, mientras que la mediana de los TMB se presentan en rojo y la mediana de genes mutados comunes entre las muestras, en azúl. Los tipos de cáncer están ordenados por la mediana de TMB mediana.

5.3.2. Genes penalizados

Del top 10 de genes sin penalizar, 57 dejan de estar en el top 10 después de la penalización, y otros 95 genes son incluidos en el top 10 después de la penalización (**Figura 30**). En los 57 genes eliminados del top10 por la penalización de su longitud hay 8 de los 11 genes más largos: TTN, MUC16, DST, MACF1, OBSCN, SYNE1, NEB y ADGRV1. Entre los 100 genes que se mantuvieron comunes entre ambos top 10, se encuentran TP53, KRAS, PIK3CA, BRAF, APC, CTNNB1, IDH1 y EGFR. Entre los 95 genes agregados al top 10 luego de penalizar los genes por su longitud se encuentran: FBXW7, FAM135B, AXIN1, PTPRT, DCC y KEAP1.

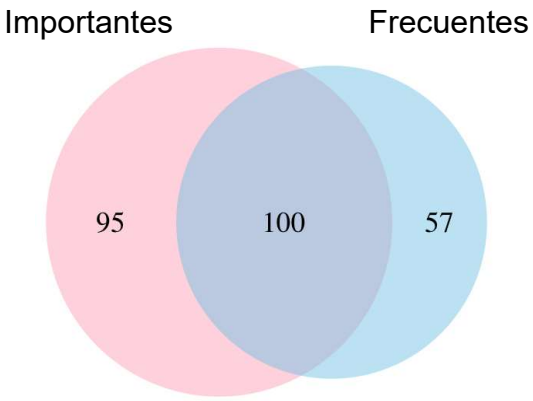


Figura 30. Diagrama de venn entre el top 10 de genes antes y después de la penalización.

5.3.3. Tumores afectados por la selección de genes/exones

Después de aplicar todas las estrategias de selección, algunos de los tipos de cáncer pierden un número considerable de muestras, considerando como umbral el 80% de las muestras iniciales. La mayoría de las muestras eliminadas en todas las estrategias se encuentran en los tipos de cáncer con menor mediana de TMB, Se obtuvieron 30 tipos de cáncer que sobrepasan el umbral en al menos una de las estrategias. De ellos, 21 sobrepasan el umbral en todas las estrategias. En la **Figura 31** se muestran los datos de las 4 estrategias.

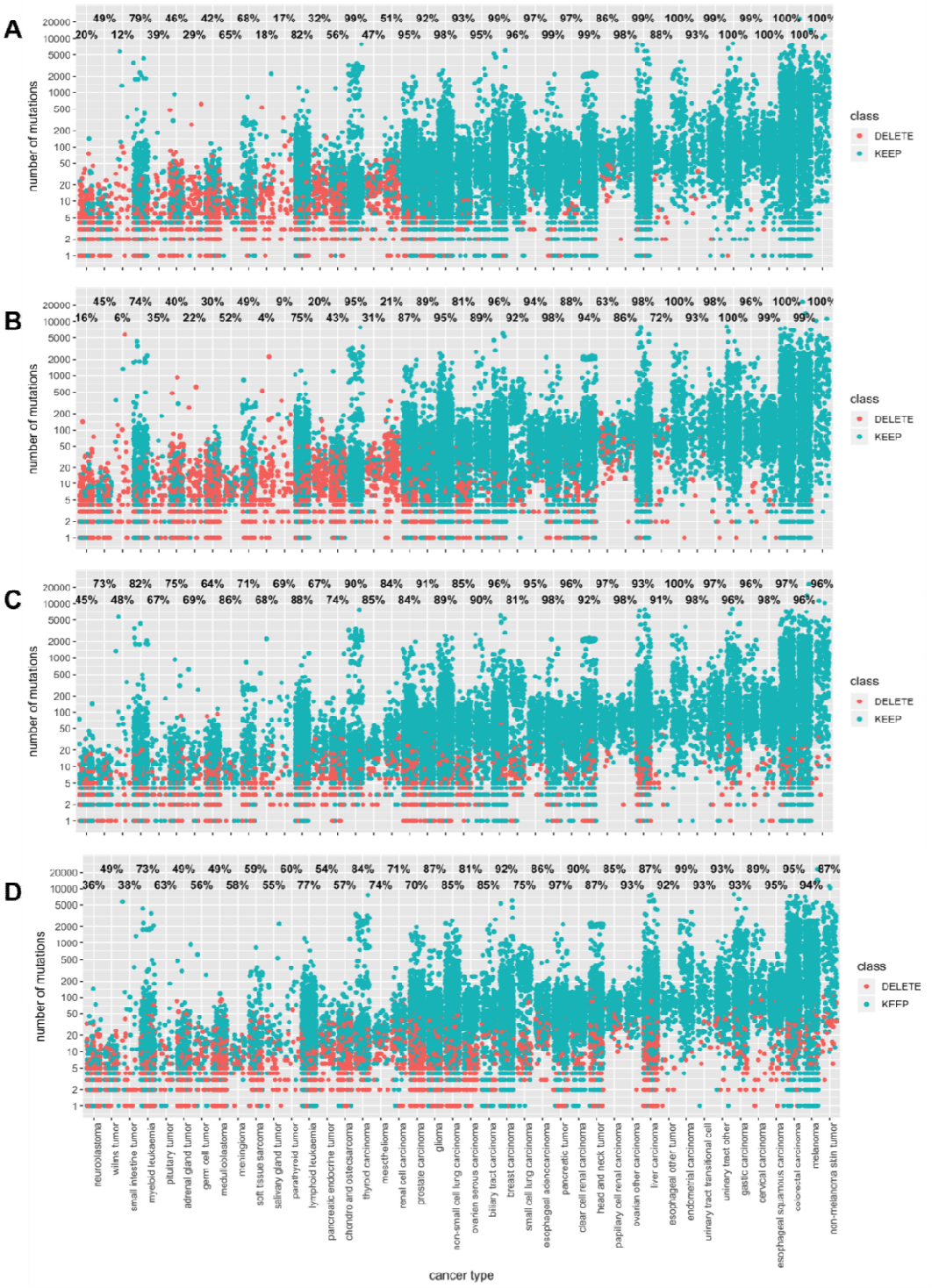


Figura 31. Carga mutacional por tipo de cáncer en las 4 estrategias.

A: Estrategia *our-gene*. **B:** Estrategia *our-exon*. **C:** Estrategia *CGC-gene/exon*. **D:** Estrategia *FO-panel-gene/exon*. Cada punto es una muestra, en rojo se definen las muestras eliminadas y en azul las que quedaron tras la selección por la estrategia. El eje "X" es el tipo de cáncer y el eje "Y" es la carga mutacional total. El porcentaje de arriba representa el porcentaje de muestras que quedaron seleccionadas en el tipo de cáncer.

5.3.4. Comparación de los modelos obtenidos

Para 28 tipos de cáncer, no pudimos generar paneles y modelos para la predicción de TMB. En algunos tipos de tumores (n = 12), el número de muestras fue insuficiente (menos del 80% de las muestras iniciales) tras la selección de genes y exones por las estrategias. En otros tipos de cáncer (n = 5) no se generaron paneles y sus modelos asociados aceptables. Finalmente, los modelos de consenso para algunos tipos de cáncer no pasaron las validaciones internas (n = 5) o externas (n = 6) debido a un alto error (RSE) o baja correlación (R^2) con el TMB experimental. Por lo tanto, se puede predecir con calidad el TMB en 14 tipos de cáncer.

En los 14 tipos de cáncer se obtuvieron modelos con la estrategia *our-gene*, mientras que solo 8, 3, 3 y 2 obtuvieron modelos con las estrategias *CGC-gene*, *FO-panel-gene*, *our-exon* y *CGC-exon*, respectivamente. De los 8 tipos de cáncer para los que se obtuvieron paneles con la estrategia *CGC-gene*, uno de ellos solo generó un solo modelo. Para comparar las estrategias en los 7 tipos de cáncer que tienen al menos 2 modelos y estrategias en común, utilizaremos los valores de calidad R^2 y RSE obtenidos de los modelos, su validación interna y externa. La comparación se realizó con la prueba estadística de Kruskal-Wallis (**Anexo 3A**), obteniendo que hay diferencias significativas entre las estrategias en cada uno de los tipos de cáncer. Al evaluar las diferencias significativas con la prueba *post-hoc* de Dunn (**Anexo 3B**), se obtiene que los modelos generados con la estrategia *our-gene* tiene los mejores resultados en los valores de calidad estudiados. Algo importante al diseñar paneles es conocer la cantidad de pares de bases a secuenciar, en la **Figura 32A-B** se muestran el R^2 y RSE respecto a la cantidad de pares de bases a secuenciar de los modelos obtenidos para los 14 tipos de cáncer en las estrategias.

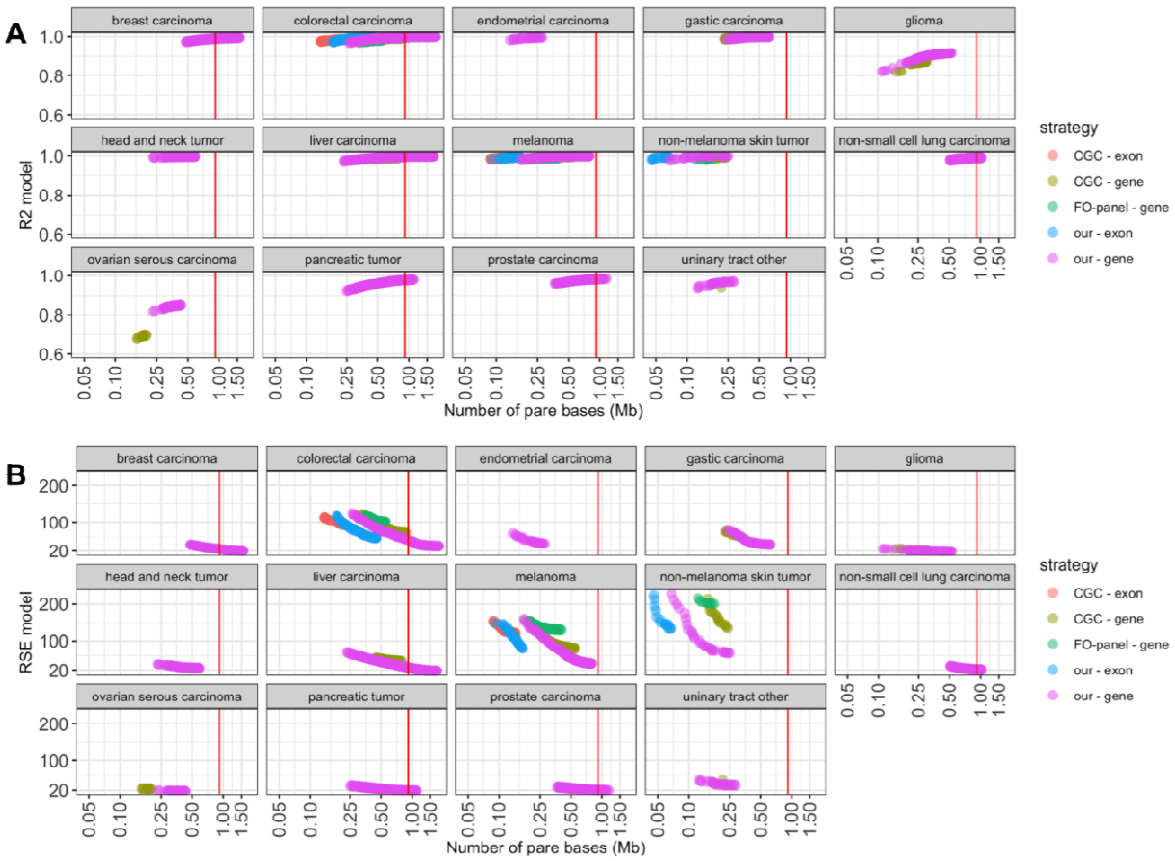


Figura 32. R2 y RSE de los modelos vs longitud en Mb en los 14 tipos de cáncer.

A-B: datos de R^2 y RSE respectivamente. La línea roja muestra el número de Mb del *FO-panel* para tomar como referencia.

5.3.5. Genes de nuestros paneles

Comparamos los genes de los paneles consenso estrategia *our-gene* por tipo de cáncer. Sorprendentemente, el número de genes comunes entre diferentes tumores varió de 0 a 40% y en algunos casos, como el tracto urinario o el carcinoma de piel no-melanoma, nunca alcanzó el 10% (**Figura 33A**). El número total de genes utilizados por nuestra selección en todos los paneles de consenso es 1239. A continuación, comparamos este conjunto con los 314 genes del FO-panel y los 719 genes del CGC. La mayoría de los genes en nuestros paneles no están contenidos en el FO-panel ni en el CGC (**Figura 33B**). Al discriminar por tipo de cáncer, la intersección siempre fue inferior al 20% (**Figura 33C**).

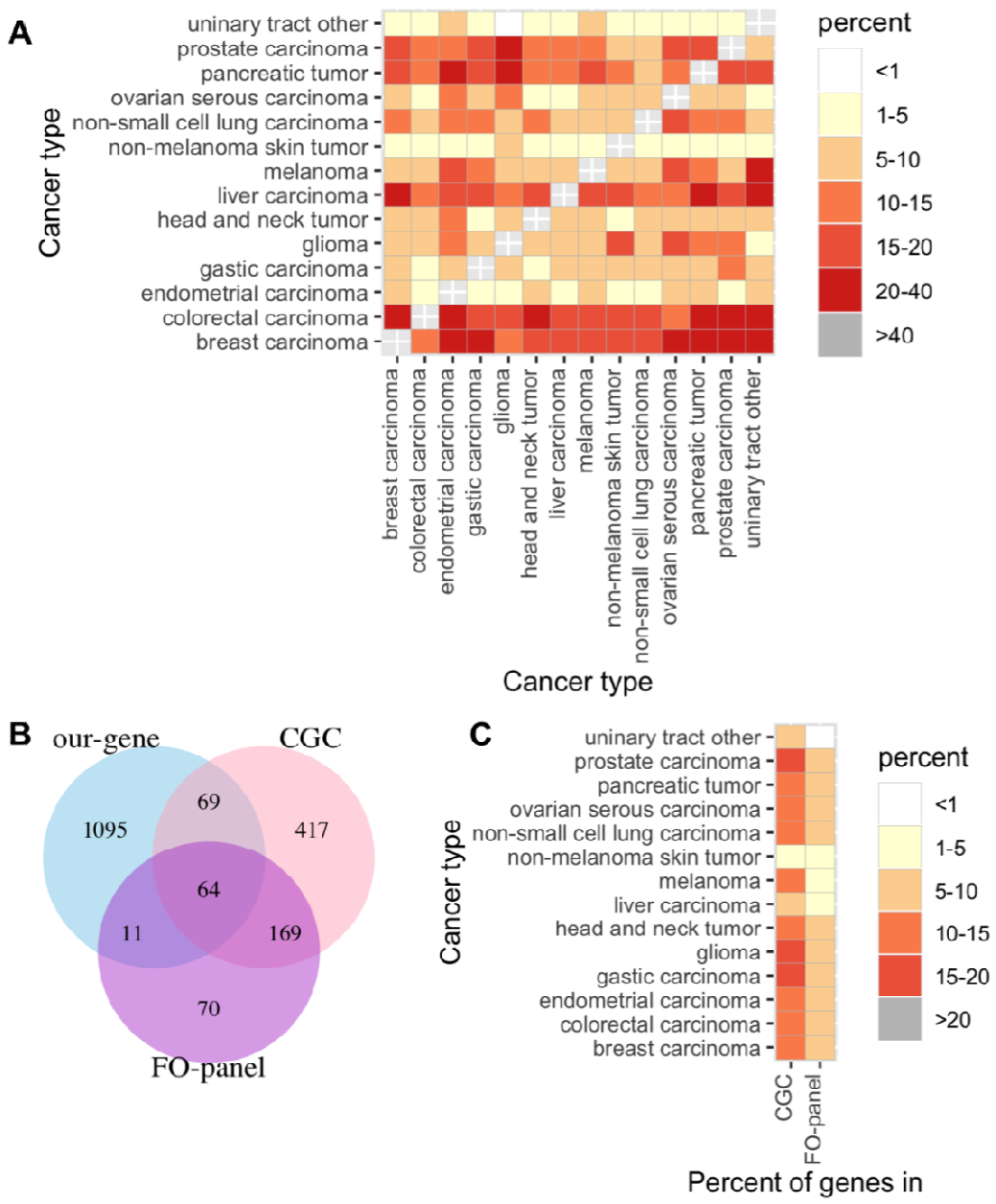


Figura 33. Intersección de los genes presentes en los modelos consensus generados con la estrategia our-gene.

A: Porcentaje de genes comunes entre los tipos de cáncer, calculados como el número de genes comunes entre los cánceres dividido por el número de genes en cada tipo de cáncer (eje X). **B:** diagrama de venns-euler de genes sumados para todos los tipos de cáncer que emergen de los modelos con la estrategia *our-gene*, genes de CGC y FO-panel. **C:** Porcentaje de genes en común con la estrategia *our-gene* y los genes de CGC y FO-panel por tipo de cáncer.

5.3.6. Modelos sugeridos por MB a secuenciar

La cantidad de paneles resultante de la estrategia *our-gene* son un total de 1446 (**Tabla 5**). Todos estos modelos cumplen las validaciones internas y externas y se ha demostrado que los paneles obtenidos con la selección de genes *our-gene* dan mejores resultados que los de las otras estrategias. Por lo anterior se decide hacer un subconjunto de estos modelos para su utilización en la clínica atendiendo a las MB a secuenciar.

Tumores	Paneles	Genes	MB
Carcinoma de seno	161	130 (48 - 210)	1.05 (0.48 - 1.56)
Glioma	68	63 (21 - 101)	0.34 (0.11 - 0.52)
Carcinoma de endometrio	25	28 (16 - 40)	0.21 (0.14 - 0.28)
Carcinoma colorectal	280	173 (33 - 323)	1.06 (0.26 - 1.81)
Tumor en hígado	250	150 (26 - 278)	1.30 (0.23 - 1.74)
Carcinoma de células no pequeñas de pulmón	107	130 (77 - 185)	0.79 (0.51 -1.02)
Carcinoma de ovario seroso	26	52 (34 - 67)	0.35 (0.23 - 0.43)
Carcinoma de páncreas	131	19 (17 - 31)	0.66 (0.25 - 1.11)
Carcinoma de próstata	108	103 (49 - 170)	0.76 (0.37 - 1.17)
Carcinoma de piel no-melanoma	31	24 (9 - 40)	0.14 (0.07 - 0.25)
Melanoma	110	70 (15 - 126)	0.51 (0.17 - 0.79)
Carcinoma gástrico	68	63 (30 - 99)	0.41 (0.24 - 0.62)
Carcinoma de cabeza y cuello	58	67 (36 - 96)	0.44 (0.23 - 0.59)
Tumores de tracto urinario no-transicional	23	30 (19 - 41)	0.20 (0.13 - 0.28)

Tabla 5. Número de paneles, genes y megabases (mediana y rango) con la selección our-gene selection la predicción de TMB en los 14 tipos de cáncer.
La selección se realiza por tipo de cáncer. Los modelos fueron ordenados teniendo en cuenta los valores de R^2 y RSE del modelo, validación interna y externa.

Luego, por tipo de cáncer se seleccionó el mejor modelo en cada uno de los siguientes rangos de Mb a secuenciar: sin importar los Mb, [1.0 - 1.2], [0.8 - 1.0], [0.6 - 0.8], [0.4 - 0.6] y con menos de 0.4 Mb. Resultando para uso clínico 39 paneles y sus modelos asociados en los 14 tipos de cáncer. De algunos tipos de cáncer se tiene varios paneles como es para tumores de seno, colorrectal e hígado,

con 5, 5 y 6 paneles respectivamente; mientras que otros solo cuentan con un panel, como es el caso del tumor en endometrio.

5.3.7. Correlación entre TMB predicho e inmunoterapia

Para evaluar la correlación entre TMB predicho con los paneles de la selección *our-gene* y la respuesta a inmunoterapia, se utilizaron dos estudios: uno de melanoma y uno de NSCLC. Se utilizaron el mejor panel de melanoma (126 genes, 0.79 Mb) y de NSCLC (175 genes, 0.97 Mb) para predecir el TMB en dichos tumores. Como se esperaba los TMB predichos muestran una excelente correlación con el TMB experimental (R^2=0.96 y 0.84 para melanoma y NSCLC. **Figuras 34A y B**). La respuesta a inmunoterapia, se evaluó con el TMB predicho, utilizando 3 valores de número de mutaciones de corte para discriminar entre TMB "alto" y "bajo": 100, 150 y 200 mutaciones,. Los mejores resultados se obtuvieron con el umbral de 150 mutaciones en ambos tumores. La tasa de respuesta de pacientes con melanoma y >150 mutaciones predichas fue del 51%, respecto al 18% en aquellos con <150 (prueba de z-score p <0.05). Además, los pacientes con alto TMB tenían una mediana de supervivencia global (OS) de 730 días, significativamente más largo que 303 días de pacientes con bajo TMB (HR = 0.66; IC95% = 0.44–0.93; P = 0.02) (**Figura 34C**). En el caso de la cohorte NSCLC, el 56% de los pacientes con alto TMB mostraron respuestas parciales, en comparación con el 17% en pacientes con bajo TMB (prueba de z-score p <0.05). A pesar de ser pocos pacientes, la mediana de FPS de aquellos con >150 mutaciones fue de 435 días, significativamente más larga que 105 días para pacientes con bajo TMB (HR = 0.27; IC95% = 0.10–0.57; P = 0.002) (**Figura 34D**). La calidad del TMB predicho respecto al beneficio clínico de la inmunoterapia se muestra en la **Figura 34E-F**. Las áreas bajo la curva reportan 0.67 para el melanoma y 0.70 para NSCLC, valores muy similares a los obtenidos con el TMB experimental (0.67 y 0.76, respectivamente).

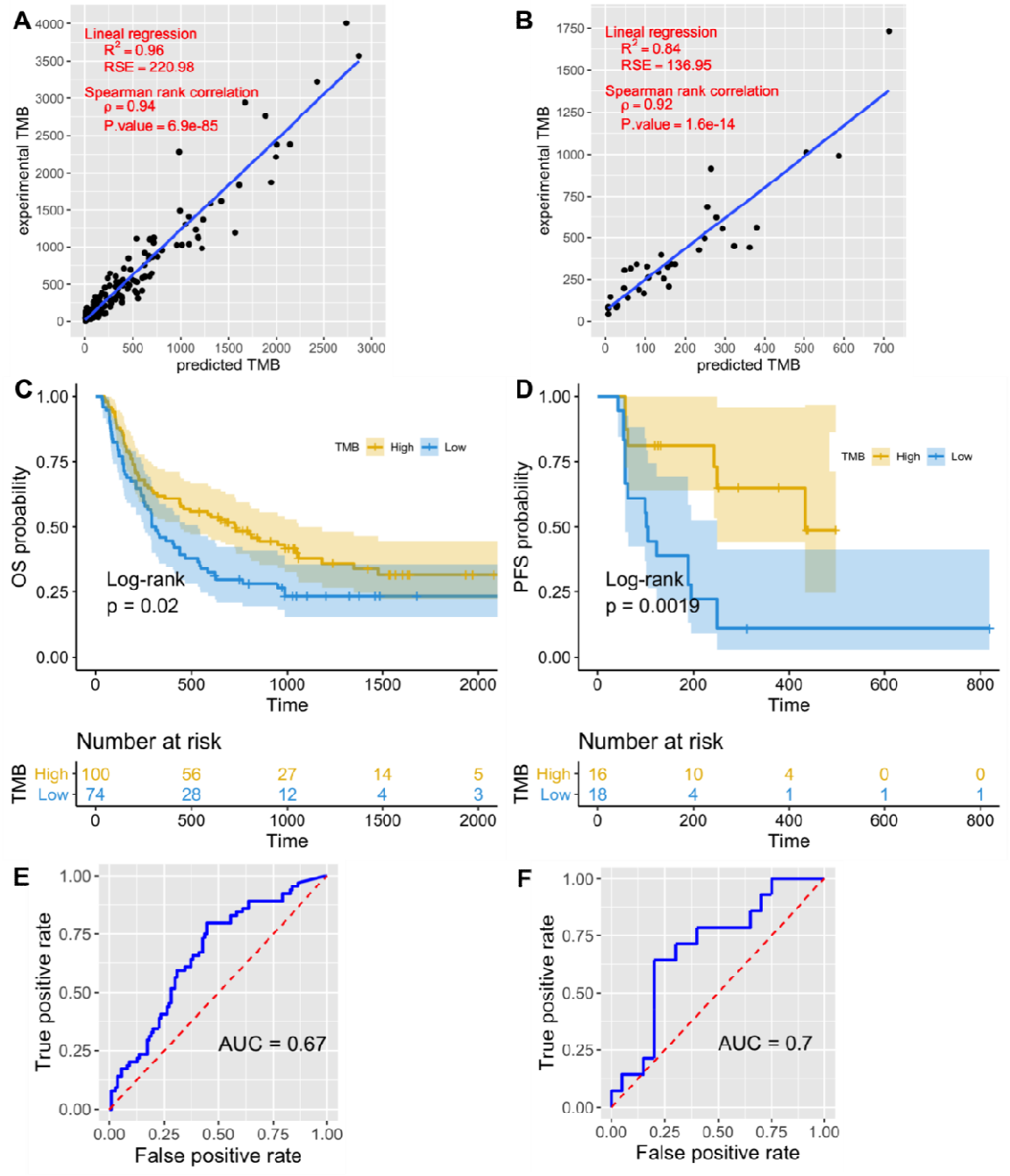


Figura 34. Análisis estadístico relacionado con TMB: la predicción de TMB se realizó con el mejor modelo.

A y B: correlación entre TMB predicho y experimental para melanoma y NSCLC respectivamente (2 muestras con >6000 mutaciones no se muestran). **C y D:** Kaplan-Meier de "OS" de pacientes con melanoma y "PFS" en NSCLC respectivamente. **E y F:** Curva ROC entre TMB predicho y respuesta a la inmunoterapia en melanoma y NSCLC respectivamente.

5.4. Importancia y aplicabilidad

Proveer paneles de genes y sus modelos asociados para predecir con precisión el TMB, y así guiar a la clínica en la terapia inmunológica a los pacientes. Es importante destacar que nuestros paneles predicen mejor el TMB que los paneles corrientemente usados en la clínica y aun mejor que con paneles que incluyeran todos los genes implicados en cáncer (CGC).

Proveer paneles específicos por tipo de tumor, ya que los distintos tipos de tumor utilizan distintos mecanismos para tomar ventaja del sistema, lo que significa que los distintos tumores no presentan las mismas mutaciones.

Tener varios paneles con calidad aceptable dependiendo de la cantidad de Mb a secuenciar permite adecuarnos a costos de secuenciación, pero siempre teniendo en cuenta que la precisión de los modelos se ve afectada mientras menor cantidad de Mb secuenciamos.

5.5. Limitaciones del estudio

Algunos tipos de cáncer presentan muy pocas muestras, lo que influye en la posibilidad de obtener buenos modelos para predecir el TMB en estos tumores.

Los modelos generados a partir de exones serían mucho menos costosos, pero para obtener buenos modelos con exones comparables a la calidad de los modelos obtenidos con genes, necesitamos aumentar considerablemente el número de muestras ya que aumentan la cantidad de variables a incorporar a los modelos.

5.6. Conclusiones del capítulo

Se identificaron paneles y sus modelos asociados para predecir con precisión el TMB en 14 tipos de cáncer.

El resultado más importante es que los paneles generados a partir de los datos (*our-genes*) son más precisos y tienen menos error que los paneles generados con los genes de FO y el CGC.

Los paneles entre los distintos tipos de cáncer comparten muy pocos genes en común, lo que evidencia la heterogeneidad de genes mutados y por consiguiente la importancia de tener paneles cáncer específicos.

Nuestros paneles tienen un rango de 0.24 a 1.5 Mb a comparación de las 1.2 Mb de uno de los paneles de NGS comerciales utilizados (se usan actualmente paneles de hasta 3.3 Mb). Son generalmente menores a los de FO en Mb a secuenciar y siendo para algunos tumores un valor tan bajo como 0.24 Mb.

6. Colaboraciones

Durante el período junio 2016 a abril 2021 en el que se encuentra enmarcada la tesis, participé en 4 estadías de investigación relacionadas al proyecto IDPfun (por Intrisically Disordered Proteins Function) sobre proteínas desordenadas. Todas las estadías se realizaron en el laboratorio de biología estructural del Laboratorio de Biología Molecular de Europa (EMBL) ubicado en Heidelberg Alemania, liderado por el Dr. Toby J. Gibson. En el marco de estas estadías se generaron importantes colaboraciones obteniendo 3 artículos publicados (o en proceso de publicación) relacionados al proyecto IDPfun y otro trabajo relacionado con la proteína desordenada ACE2 relacionada al COVID-19. En el siguiente capítulo se esbozan las colaboraciones realizadas.

6.1. Introducción a proteínas desordenadas

La estructura experimental de casi la mitad de las proteínas o regiones de proteínas nunca se ha podido determinar y son inaccesibles para el modelado de homología. Esta fracción del proteoma sin similitud detectable con ninguna estructura conocida se considera el proteoma oscuro[150]. La mayor parte de su oscuridad se debe al hecho de que estas proteínas y regiones están intrínsecamente desordenadas. Además, se sabe que el 38% de los residuos en todas las proteínas humanas no pueden asignarse a una familia Pfam, una base de datos de familias de proteínas reunidas por similitud de secuencia[151].

Tres de los estudios que se listan en este capítulo forman parte de un proyecto internacional llamado IDPfun relacionado a las estadías de investigación. IDPfun es un consorcio internacional cuyo objetivo es ampliar nuestro conocimiento sobre las funciones de las proteínas intrínsecamente desordenadas (IDP). A partir de las herramientas y bases de datos computacionales de última generación disponibles, que han sido desarrolladas principalmente por participantes de IDPfun, su objetivo es impulsar un nuevo nivel de caracterización de IDP.

Este proyecto en el que participa el grupo de la Dra. Marino-Buslje, otros 2 laboratorios Argentinos y 5 laboratorios Europeos (entre los que se encuentra el EMBL) ha recibido financiamiento del programa de investigación e innovación Horizon 2020 de la Unión Europea en virtud del acuerdo de subvención número 778247, relativo a la beca Marie Skłodowska-Curie de intercambio de personal de investigación e innovación (RISE).

Adicionalmente, durante el período en que se enmarca la tesis ocurrió la pandemia de COVID-19, mientras realizaba una de las estadías de investigación en el grupo del Dr. Toby Gibson. Haciendo uso de los datos y el conocimiento en proteínas desordenadas se realizó un estudio sobre la proteína desordenada ACE2 que se encuentra relacionada al COVID-19, siendo este el cuarto estudio resultado de la colaboración.

6.2. Ontología de proteínas desordenadas

6.2.1. Introducción

Se han realizado varias categorizaciones funcionales de las proteínas desordenadas[152–155]. En la publicación de Dunker se introdujeron 28 términos funcionales[153]. Peter Tompa sugirió cinco términos generales para clasificar las funciones moleculares de las regiones con desorden[154]. Este esquema se extendió ligeramente desde la publicación original y ahora forma la base de la anotación funcional en la base de datos DisProt versión 7.0[156]. Para el 2018 en marco del proyecto para el estudio de proteínas desordenadas (IDPfun) una de las tareas fue transformar el esquema en una ontología.

6.2.2. Implementación de la nueva ontología

En el esquema inicial existían términos que se encontraban definidos o tenían un equivalente en otras ontologías existentes, como en la ontología de genes (GO)[157,158], y en la de métodos experimentales, donde se definían con mayor rigor. Además el formato en que se presentaba el esquema (un excel) no era compatible con los formatos de descripción de ontologías mediante estándares OWL y OBOformat.

El software libre Protégé[159] (https://protege.stanford.edu/) permite implementar ontologías con lenguaje estandarizado. Además este software permite un manejo (adición, eliminación o modificación) de los términos y relaciones en la ontología de forma rápida, fácil e intuitiva. Se decidió implementar la ontología con ayuda de este software ya que esto ayudaría a mejorar el diseño de la ontología y posteriormente su incorporación al sistema de ontologías. La **Figura 35** muestra una representación de la ontología de proteínas desordenadas en el software protégé. Esta ontología todavía se encuentra en un paso incipiente y necesita constantes mejoras para adecuarnos a las necesidades de anotación en las regiones desordenadas. Durante esta tarea se requirió de mucho estudio de la bibliografía sobre proteínas desordenadas y de charlas con expertos en el tema para definir las correcciones.

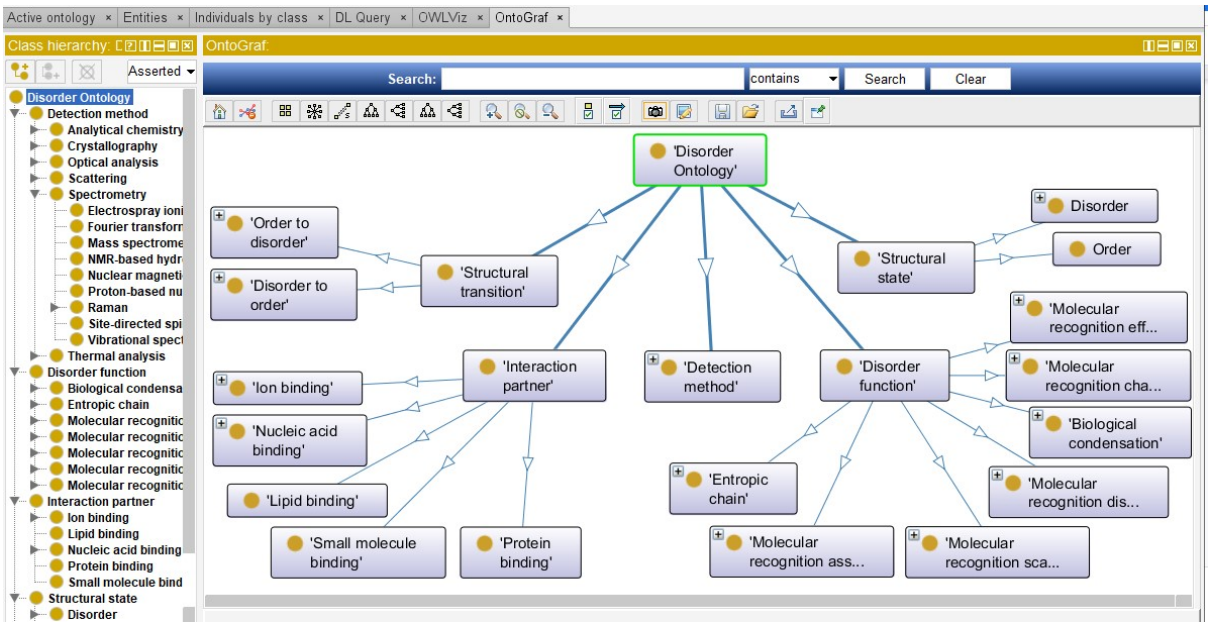


Figura 35. Ontología de proteínas desordenadas en el software protégé

6.2.3. Resultados

La ontología es utilizada como uno de los elementos de la base de datos DisProt, la que ha tenido una actualización en enero del 2020 para llegar a la versión 8.0[160].

6.3. Transferencia de anotaciones por homología

6.3.1. Introducción

La base de datos de proteínas desordenadas (DisProt, URL: https://disprot.org) es el repositorio principal de datos relacionados al desorden que se centra en proteínas o regiones desordenadas con verificación experimental. La versión 8.0 (2019-09) es el resultado del esfuerzo de más de 60 expertos, recopila 1390 proteínas y un total de 13169 anotaciones de desorden verificadas manual y experimentalmente[160].

A pesar de la gran ventaja de ser una base de datos curada manualmente que asegura que cada anotación tenga un soporte experimental, contiene un número relativamente pequeño de proteínas. La anotación de desorden es un proceso que requiere mucho trabajo y mucho tiempo, destacando la importancia de extender las anotaciones de desorden documentadas de proteínas conocidas a proteínas con secuencias similares y estado estructural desconocido. La expansión de este conocimiento a proteínas homólogas mejorará considerablemente el número de proteínas con una información muy valiosa para comprender su función.

El objetivo principal de este trabajo es transferir anotaciones de regiones desordenadas y términos de ontología de proteínas desordenadas de DisProt a proteínas ortólogas, con el supuesto de que las proteínas ortólogas tienen una función similar. Para ello se analizarán varios métodos de alineamiento de secuencias (MSAs). Además, se calcularán los puntajes de calidad de alineamiento completo, así como los puntajes de calidad parcial (región) para poder transferir una anotación a una región particular.

6.3.2. Materiales y Métodos

La **Figura 36** describe brevemente el esquema de trabajo como sigue: de cada entrada de DisProt, se buscan las proteínas ortólogas en las bases de datos de ortólogos. Se hacen alineamientos de las secuencias ortólogas con 3 métodos y se les evalúa la calidad. Se comparan 3 puntuaciones de calidad correspondientes a la alineación completa (longitud de proteína completa) y 2 regiones identificables: las regiones desordenadas y las regiones con términos de ontología. Después de evaluar la calidad del alineamiento de una región en particular, es posible definir si es lo suficientemente bueno como para transferir una característica (región desordenada o término de la ontología).

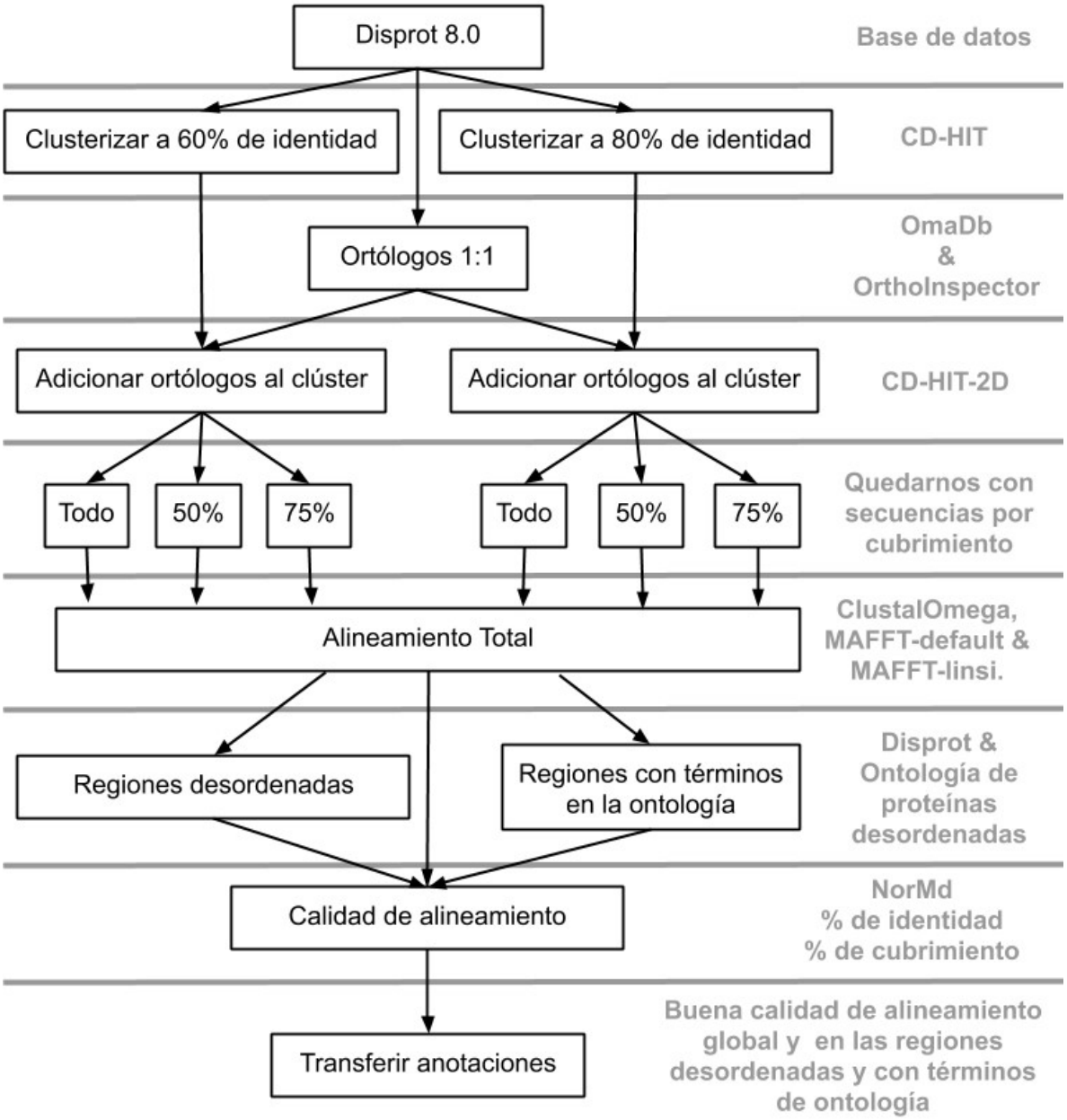


Figura 36. Diagrama general de flujo de trabajo para este proyecto.

6.3.2.1. Procesamiento de datos

DisProt 8.0 (versión de 2019-09) es una base de datos de proteínas desordenadas[160]. Tiene 1390 proteínas y un total de 13169 anotaciones para estas proteínas. Algunas anotaciones en la base de datos están relacionadas con el dominio pfam, el compañero de interacción y la información experimental sobre la región desordenada y los términos de ontología de la región. Se descargan los datos y se seleccionan las proteínas siguiendo las reglas: proteínas canónicas, proteínas completas (no fragmentos) y secuencias que no tienen indefiniciones

(aminoácidos "X"). El conjunto de proteínas que cumplen los filtros está formado por 1359 proteínas.

6.3.2.2. Agrupar proteínas por porcentaje de identidad

Se agrupan las 1359 proteínas por su porcentaje de identidad con el software CD-HIT[161,162]. El software devuelve una proteína de referencia por cada grupo. El número de grupos obtenidos con distintos porcentajes de identidad se muestra en la **Tabla 6**. Se seleccionan los grupos con 60% de identidad y 80% para los estudios posteriores.

% iden.	40	50	**60**	70	**80**	90	100
Grupos	1166	1212	**1241**	1263	**1286**	1305	1247

Tabla 6. Cantidad de grupos obtenidos con cd-hit con diferentes porcentajes de identidad (rango 40% - 100%).

Los valores rojos son los parámetros utilizados desde ahora hasta el final del trabajo.

6.3.2.3. Recopilación de proteínas ortólogas

Para realizar alineamientos y transferir anotaciones a proteínas homólogas, se necesita aumentar el número de proteínas en cada grupo, por lo que el siguiente paso fue buscar proteínas ortólogas. Para ello, por cada proteína de referencia se buscan proteínas ortólogas de tipo 1:1 (para disminuir la posibilidad de agregar parálogos en los alineamientos). Programáticamente se consultaron las 1359 entradas de DisProt (con su id de UNIPROT) en las bases de datos de ortólogos OmaDB[163] (https://omabrowser.org) y OrthoInspector[164] (https://www.lbgi.fr/orthoinspectorv3/). Consideramos ortólogos no válidos a las proteínas ortólogas con uniprot obsoleto, fragmentos o con aminoácido "X", por ellos fueron eliminados, detalle de los filtros en la **Tabla 7**. Un total de 1098 de las 1359 secuencias tienen ortólogos tipo 1:1 entre las dos bases de datos. La **Figura 37A** muestra la cantidad de proteínas de disprot que fueron encontradas en ambas bases de datos, mientras que la **Figura 37B** muestra los ortólogos encontrados en ambas bases de datos.

		OmaDb	OrthoInsp.	Total
Proteínas de DisProt no encontradas		303	284	
Proteínas de DisProt encontradas	Sin ortólogos	34	5	
	Sin ortólogos tipo 1:1	10	19	
	Con ortólogos no válidos	2	2	
	Con ortólogos válidos tipo 1:1	1010	1049	1098
Ortólogos válidos encontrados		123202	245743	311935

Tabla 7. Las proteínas DisProt y sus ortólogos se encuentran en las bases de datos OmaDb y OrthoInspector.

Ortólogos no válidos son los que uniprot obsoleto, fragmentos o con aminoácido "X", los demás se consideran válidos.

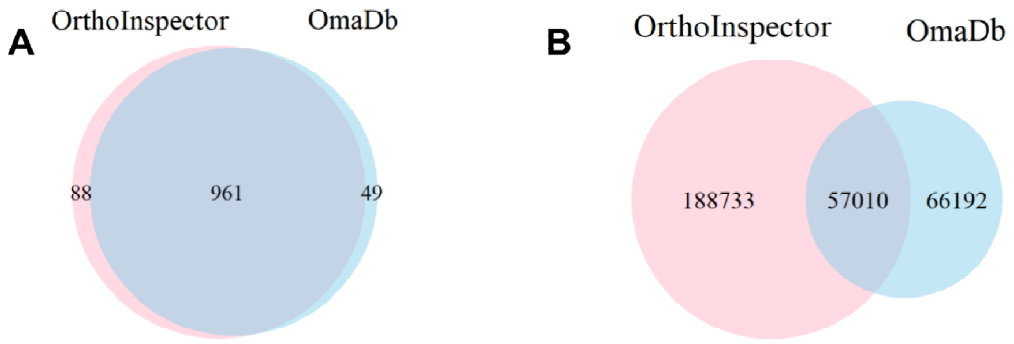


Figura 37. Diagrama de Venn entre las bases de datos de ortólogos.

A: Proteínas DisProt que tienen ortólogos de tipo 1:1. **B:** Proteínas ortólogas encontradas de tipo 1:1.

6.3.2.4. Incorporación de proteínas ortólogas a los grupos

La incorporación de las proteínas ortólogas a los grupos de 60% y 80% de identidad se realiza con el software CD-HIT-2D. Este software necesita como entrada dos bases de datos, la primera es una base de datos con las proteínas de referencia (db1), en nuestro caso esta tiene las proteínas representativas de cada grupo al 60% y 80% de identidad. La segunda (db2) tiene todos los ortólogos y las proteínas DisProt no seleccionadas como representante de *clúster*. Después de la incorporación de los ortólogos a cada grupo, se consideran los conjuntos: todas las proteínas sin diferenciarlas por su cubrimiento (% de longitud de la proteína de referencia que puede ser alineado con su homóloga), secuencias con al menos 50% y 75% de cubrimiento respecto a la proteína representativa del grupo. De esta

forma, se obtuvieron seis conjuntos de datos por proteína de referencia, lo que resultó en 5803 grupos de proteínas para alinear (**Tabla 8**).

% iden.	Cubrimiento	Grupos >= 2 secuencias	Secuencias en grupos con >=2 secuencias	Proteínas en Swiss-prot
60%	Sin diferenciar	989	53497	6334
	Al menos 50%	983	53021	6323
	Al menos 75%	980	51231	6300
80%	Sin diferenciar	954	29069	4085
	Al menos 50%	950	28851	4079
	Al menos 75%	947	28045	4063
Total de alineamientos		5803		

Tabla 8. Número de alineamientos para cada grupo en diferentes porcentajes de identidad y cubrimiento.

6.3.2.5. Alineamientos con 3 métodos y su calidad

Se utilizaron 3 métodos de alineamiento: Clustal Omega[165,166] con parámetros predeterminados, MAFFT con parámetros predeterminados[167] y MAFFT con parámetros "linsi"[168] (desarrollado para alinear proteínas desordenadas). Se obtuvo un total de 17402 alineamientos múltiples de secuencia (MSA) (5803 x 3 = 17409, pero 7 alineamientos no pudieron ejecutarse con los parámetros de mafft-linsi). Se usó el software NorMD[169] versión 1.3 para calificar la calidad de los alineamientos. Para cada uno de los alineamientos se mide la calidad en tres regiones: completo, regiones desordenadas y regiones con términos de ontología, de acuerdo con la proteína representativa. Los alineamientos con una puntuación inferior a 0.6 en NorMD son considerados malos[169].

6.3.3. Resultados

6.3.3.1. Calidad de los alineamientos

La **Figura 39** muestra los puntajes de las 18 estrategias (métodos de alineaciones x porcentaje de identidad x porcentaje de cubrimiento = 3 x 2 x 3) en las 3 regiones de alineamiento.

Al comparar estadísticamente la calidad de los alineamientos (prueba de Kruskal-Wallis) en las 18 estrategias en las 3 distintas regiones alineadas se obtuvieron diferencias significativas entre ellas. El análisis de las diferencias por pares entre las estrategias se realizó con la prueba de Dunn. El **Anexo 4** muestra los p valores de

ambas pruebas estadísticas. Como se esperaba, los alineamientos de secuencias con más del 80% de identidad son diferentes y mejores que los alineamientos con una identidad del 60% en todos los métodos y en todas las regiones. Un resultado notable es que los tres métodos de alineamiento no son estadísticamente diferentes en ninguna de las regiones (**Figura 38A-C**). Otro resultado interesante es que no hay diferencias significativas entre los alineamientos con 50% y 75% de cubrimiento, en ninguna de las tres regiones (regiones anotadas de alineamiento completo, desorden y ontología).

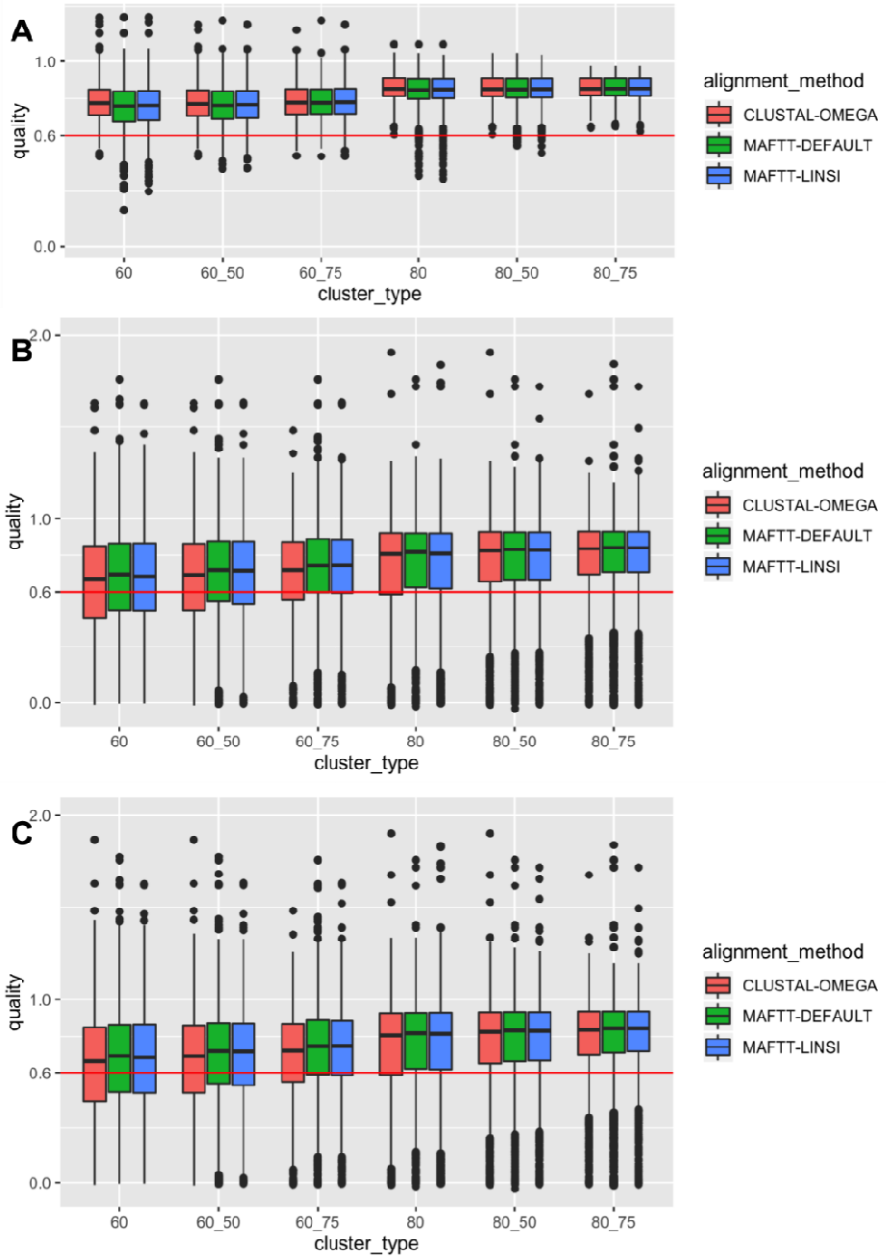


Figura 38. Calidad de los alineamientos en las 6 estrategias y los 3 métodos de alineamiento.

Donde **A, B y C:** valores de calidad del alineamiento: completo, en regiones desordenadas y en regiones con términos en la ontología, respectivamente. Para claridad las puntuaciones 5, 14, 14 no se muestran en las figuras A, B y C, respectivamente.

6.3.3.2. Validación en grupos con más de una proteína de Disprot

Se aprovecha el hecho de que algunos grupos tienen más de una entrada DisProt, para probar la hipótesis de que dentro de un alineamiento las regiones anotadas de diferentes proteínas están alineadas y se pueden transferir.

La superposición en cada grupo entre las regiones desordenadas de la proteína de referencia y las regiones de las demás proteínas (no-referencia) se calculó para todos los grupos que tienen más de una proteína de DisProt (utilizando los alineamientos realizados con MAFFT y parámetros por defecto) para ambos conjuntos de grupos (60% y 80% de identidad). En ambos conjuntos de grupos, la superposición de cada región se calcula como el porcentaje de aminoácidos en la región de la proteína de referencia presente en la proteína de prueba, y se divide en conjuntos (de 1 a 100). Ejemplo, si una región desordenada de referencia se superpone con una no-referencia en el 13%, pertenece al conjunto de "10-20". Hay casos en los que la región alineada de una proteína DisProt no-referencia dada no está anotada en DisProt en consecuencia, el porcentaje de superposición es 0 (incluso cuando la identidad entre regiones es muy alta).

Se calcula una puntuación de similitud e identidad entre ambas regiones desordenadas superpuestas. El número total de regiones superpuestas es 142 y 85 para los conjuntos de identidad del 60% y 80% respectivamente. La distribución de la superposición secuencial de regiones desordenadas entre la referencia y otras regiones equivalentes de entrada DisProt se muestra en la **Figura 39**. En ambos conjuntos de grupos (60% y 80% de identidad), la mayoría (más del 50%) de los pares de regiones superpuestas tienen al menos un 50% de cubrimiento (58% y 61% respectivamente).

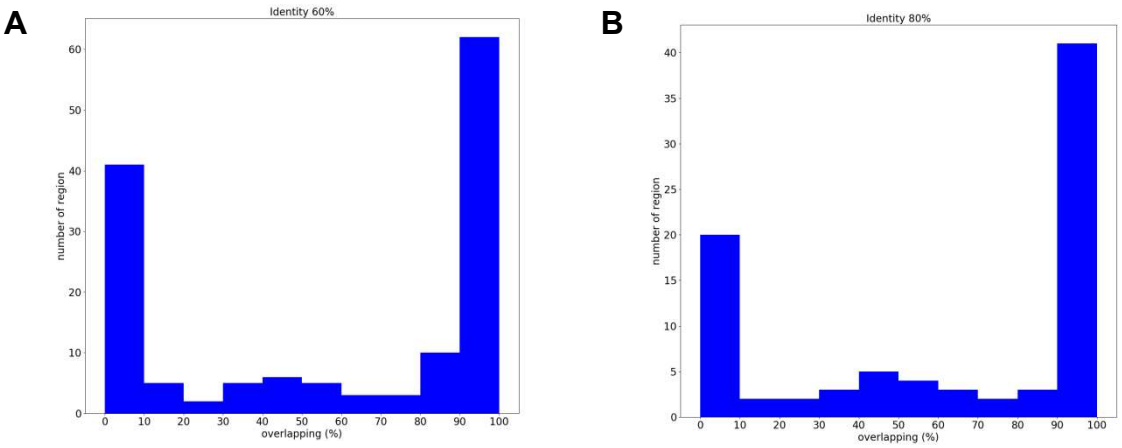


Figura 39. Distribución de zonas superpuestas entre la proteína de referencia y la región equivalente en proteínas no de referencia.

Donde **A**: al 60% de identidad (142 regiones superpuestas). **B**: al 80% de identidad (85 regiones superpuestas).

La superposición entre las regiones de la proteína de referencia y no-referencia puede subestimarse debido a las diferencias en la anotación de las dos entradas DisProt. Las regiones 100% idénticas pueden tener una superposición del 0% si la región no se anotó en la proteína de consulta. Lo mismo se aplica a diferentes porcentajes de superposición para regiones que no están anotadas en exactamente las mismas posiciones entre las dos proteínas. La **Tabla 9** presenta algunos ejemplos de estos alineamientos con diferentes valores de superposición y 100% de identidad.

Tipo	Fragmento Alineamiento	%superposición
Referencia	MSTNPKPQRKTKRNTNRRPQDVKFPGGGQI	
No Referencia	MSTNPKPQRKTKRNTNRRPQDVKFPGGGQI	**100**
No Referencia	MSTNPKPQRKTKRNTNRRPQDVKFPGGGQI	**100**
No Referencia	MSTNPKPQRKTKRNTNRRPQDVKFPGGGQI	**0**
No Referencia	MSTNPKPQRKTKRNTNRRPQDVKFPGGGQI	**50**
No Referencia	MSTNPKPQRKTKRNTN------KFPGGGQI	**50**

Tabla 9. Ejemplos de alineamientos entre una proteína de referencia con proteínas no de referencia pero ambas de disprot.

Se muestran los diferentes porcentajes de superposición y 100% de identidad. En color naranja la región desordenada anotada en la proteína de referencia. En verde las regiones desordenadas anotadas en las proteínas que no son las de referencia.

Los valores de las métricas de identidad y similitud por pares entre todas las regiones desordenadas se representan en la **Figura 40** para los MSA de identidad del 60% y 80%. Para suavizar el efecto de las diferencias de anotación, también investigamos los puntajes de identidad y similitud de las regiones desordenadas que se superponen al menos al 50% con la proteína de referencia (50% y 80% de cubrimiento respectivamente). Como se esperaba al comparar las regiones de ambos grupos de identidad global (60% y 80%) en tres niveles de regiones superpuestas (todas, 50% y 80%), obtenemos diferencias estadísticas significativas tanto en identidad como en similitud de las regiones desordenadas (valor de prueba de Kruskal-Wallis 2.8e-09 y 1.2e-08 respectivamente). Para evaluar las diferencias encontradas se usó el conjunto con del clúster con 80% de identidad global y el 80% de las regiones superpuestas (este conjunto tiene la media más alta en identidad y similitud) como control para la comparación por pares con la prueba de Dunn. La

comparación por pares muestra diferencias estadísticas entre el conjunto de control y todos los conjuntos al 60% de identidad global en ambas métricas, pero sorprendentemente, no mostró diferencias estadísticas entre conjuntos con 80% de identidad global sin importar el nivel de superposición en esas métricas.

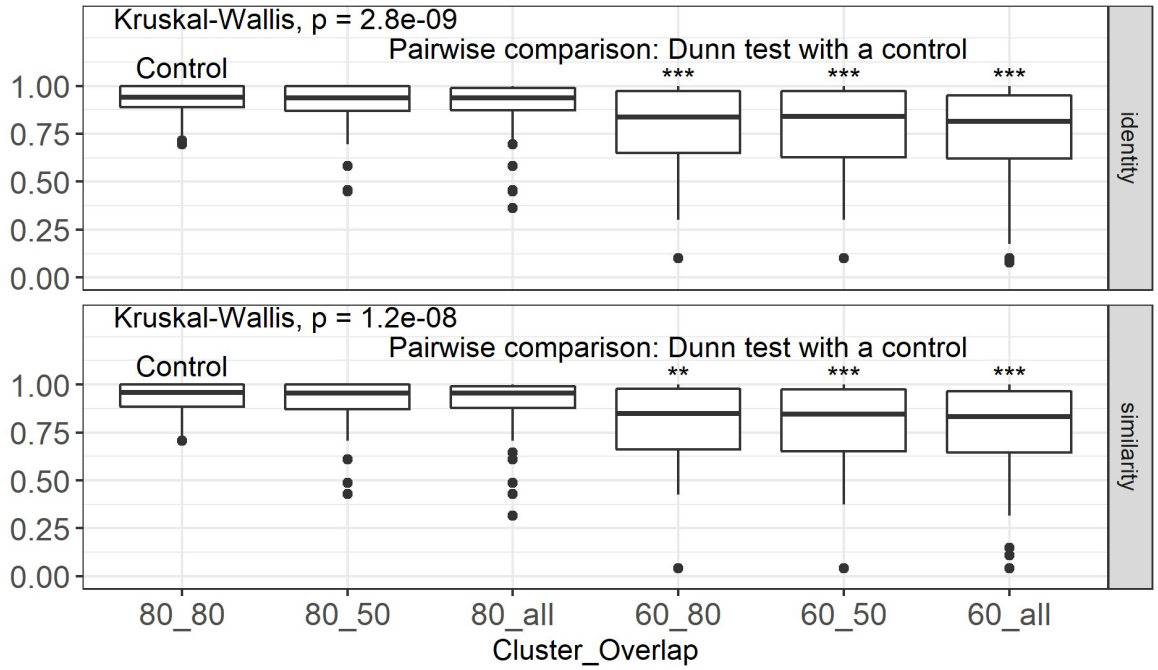


Figura 40. Puntajes de identidad y similitud entre pares de regiones desordenadas anotadas de proteínas DisProt de referencia y no referencia. Donde el eje x son los conjuntos definidos por cluster a 60 y 80% de identidad con el porcentaje de solapamiento entre las anotaciones de Disprot (todas, >50%, >80%).

La razón principal es que en el subconjunto de más del 50% de superposición, muchos pares de comparación excluidos, aunque no se superponen, son muy idénticos (100% varias veces) debido al hecho de que la región de proteína no de referencia no está anotada en DisProt.

6.3.3.3. Posibles anotaciones a transferir

Del MSA realizado con Clustal-omega y 80% de identidad global y 75% de cobertura global, evaluamos la identidad y la cobertura de las regiones desordenadas y las regiones con términos de ontología. En total contamos con 45292 y 36805 provenientes de regiones desordenadas y regiones con términos de ontología respectivamente. Teniendo en cuenta los resultados anteriores, decidimos seleccionar como posibles regiones para transferir la regiones desordenadas y de los términos de ontología cuando el porcentaje de identidad y cobertura es igual o

mayor al 80%. Dado esto, podemos transferir 37257 regiones con términos de ontología y 30172 regiones desordenadas (**Figura 41**). Esas regiones están formadas 23214 y 23449 por proteínas homólogas en regiones desordenadas y regiones con términos de ontología respectivamente, referentes a 913 y 915 proteínas de referencia de Disprot respectivamente.

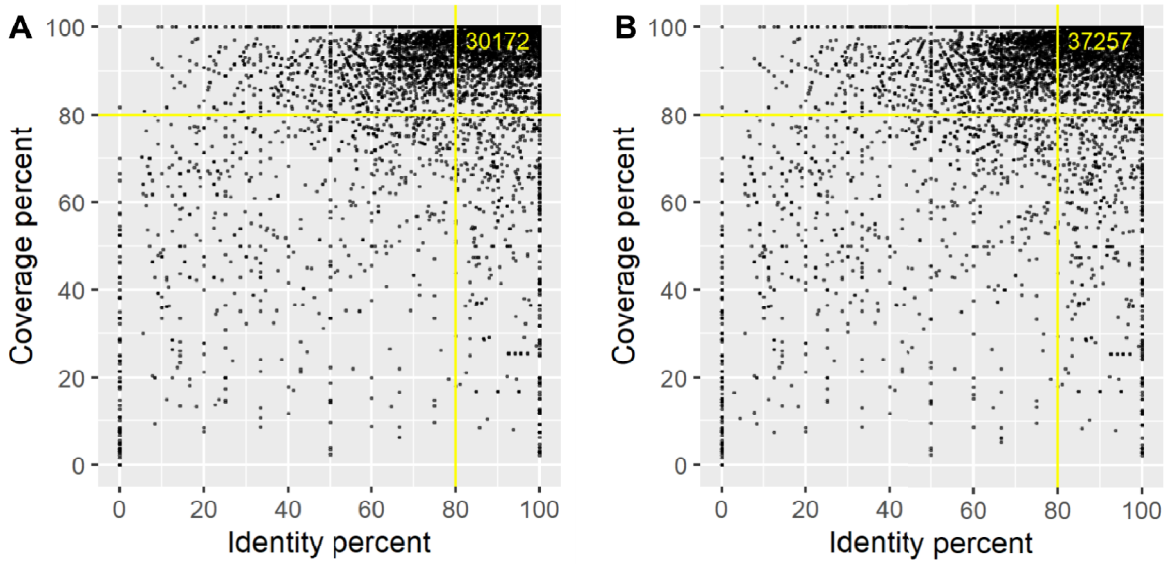


Figura 41. Porcentajes de identidad y cobertura en alineamientos por pares entre una proteína de disprot como referencia y una proteína.

Los alineamientos provienen de grupos al 80% de la identidad global y al 75% de la cobertura global. Las líneas amarillas representan el límite en el 80% de la identidad y cobertura. El número en amarillo es el total de regiones igual o mayor a 80% de identidad y cobertura. **A:** Porcentajes de identidad y cobertura en regiones desordenadas. **B:** Porcentajes de identidad y cobertura en regiones con términos de ontología.

6.3.4. Importancia del estudio

Se puede observar que la mayoría de los alineamientos dan un puntaje de identidad y similitud superior a 0.6 (alrededor de 0.8 en promedio) en los ortólogos de recopilación de MSA con más del 60% de identidad y superior a 0.8 (alrededor de 0.9 en promedio) para aquellas secuencias de recopilación de MSA de más del 80% idéntico a la proteína de referencia.

Teniendo en cuenta estos resultados, podemos transferir las anotaciones de la región desordenada a 23214 proteínas de las 51231 encontradas como ortólogos y

las anotaciones de término de ontología a 23449 proteínas. Con esta transferencia la base de datos de Disprot se incrementa en 16 veces la cantidad de proteínas con anotaciones de desorden.

6.4. PED, servidor de estructuras desordenadas

6.4.1. Introducción

Enmarcado en el proyecto de IDPfun, otra de las tareas fue almacenar información estructural de las proteínas desordenadas relacionada a las posibles conformaciones de una proteína dependiendo del ensamblado. La base de datos PED existe con este objetivo desde 2013[170], en su versión 3.0 contaba con 14 proteínas incorporadas y 24 ensamblados de ellas, cada ensamblado varía el número de conformaciones entre 3 y 5000.

Otras personas del proyecto han tenido como tarea detectar errores en los archivos almacenados y programar algunos tipos de reportes sobre los ensamblados. La base de datos cuenta con una estructura obsoleta y difícil de incorporar nuevas estructuras. Se decidió realizar una nueva versión de la base de datos. Se realizaron nuevos códigos implementados en lenguaje Python ya que es el lenguaje en que se programaría la REST API para el acceso programáticamente al servidor. Además se programamó i) validación del formulario y de los archivo PDB, ii) recopilación de mutaciones, modificaciones, moléculas y residuos faltantes en el PDB, iii) cálculo de métricas con los softwares Molprobity y DSSP y otros cálculos con las coordenadas de los átomos del PDB, iv) Creación de gráficos a partir de los datos de las métricas, v) creación de un informe PDF, similar al que se genera tras depositar una nueva entrada en el Protein Data Bank, y vi) el mapeo de las entradas antiguas a la nueva versión.

6.4.2. Materiales y métodos

6.4.2.1. Pipeline del servidor para depositar una nueva entrada

Los códigos utilizados en la deposición de una nueva entrada se agruparon en 4 archivos ejecutables: T1, T2, T3 y T4. La T1 relacionada con la validación y recolección de datos. La T2 relacionada con los cálculos de las distintas métricas utilizando Molprobity, DSSP y las coordenadas de los átomos. La T3 dedicada a graficar todas las figuras necesarias a mostrar en el servidor y en el informe final. La T4 realiza el informe final en formato PDF donde se muestran datos colectados en el formulario, los resultados de validación y recolección de datos del PDB, además de mostrar gráficos de las métricas calculadas. El pipeline donde se utilizan estas tareas se muestra en la **Figura 42**.

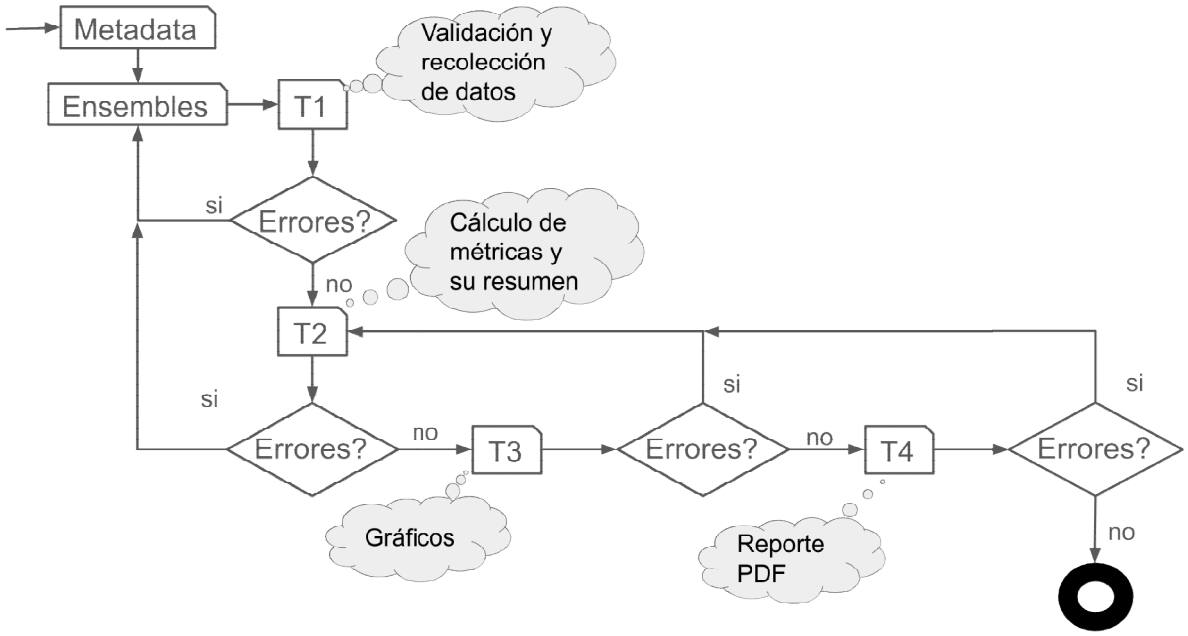


Figura 42. Pipeline utilizado en el servidor para depositar una nueva entrada. Donde T1, T2, T3 y T4 son los archivos ejecutables descritos.

6.4.2.2. Validaciones de los datos a depositar

La validación de datos siempre es el primer paso de cada programa, pues de esta forma se garantiza no tener errores al ejecutarlo. En el servidor PED la validación se realiza a los metadatos ingresados en el formulario, a cada archivo PDB para que cumpla el formato estándar y si una nueva entrada tiene más de un ensamblado, todos los ensamblados tienen que cumplir las mismas características. La validación forma parte de la T1. Si algún error es reportado por la validación no se puede proceder a la T2 como se muestra en la **Figura 42**. A continuación se listan las validaciones programadas para el servidor:

Validación de los metadatos entrados por el formulario, entre ellos se verifica: i) que al menos exista un archivo PDB por cada entrada, ii) especificar al menos una construcción de cadena, iii) cada cadena tiene que tener un nombre único y es un carácter en mayúscula entre A y Z, iv) es obligatorio especificar Uniprot accession, posición inicial y final para cada fragmento en el constructor de cadena, excepto en el caso de tags, likers y secuencias no reportadas un UniProt donde es obligatorio especificar la secuencia válida y v) las posición de inicio y fin son números enteros entre 1 y la longitud de la secuencia (especificada o de la secuencia del Uniprot accession)

Validación del formato del archivo PDB o mmCIF, entre ellos se verifica: i) Errores reportados al parsear el archivo con biopython, ii) coincidencia entre los

nombres de cadena en metadatos y el archivo PDB, iii) todos los modelos tienen que tener las mismas cadenas, secuencia de aminoácidos, modificaciones de aminoácidos, moléculas y el mismo numerado de residuos en cada cadena entre los modelos para ser considerados en la misma entrada de la base de datos.

Validación en caso de varios ensamblados en la entrada, se verifica que todos los modelos de ensamblados de la misma entrada tengan la misma lista de residuos faltantes, residuos modificados, residuos mutados y de moléculas.

6.4.2.3. Recolección de datos

La recolección de datos provenientes del archivo PDB forma parte de la T1 y son usadas durante las validaciones. Los datos a recolectar son: residuos faltantes, residuos mutados con su tipo de mutación (C, D or I para cambio, deleción o inserción respectivamente), residuos modificados y otras moléculas (consideradas como todos los códigos de 3 letras que no forman parte de los 20 aminoácidos estándar, los 6 aminoácidos no estándar (PYL, SEC, ASX, GLX, XLE, and XXX), y los códigos de las modificaciones.

6.4.2.4. Modificaciones en los residuos

Los siguientes pasos fueron ejecutados para obtener la lista de modificaciones posibles a los residuos: i) se descargó el mapeo de tres letras utilizado en el PDB (ftp://ftp.wwpdb.org/pub/pdb/data/monomers/components.cif), ii) se seleccionaron los códigos de 3 letras que en el tipo tuvieran escrito “peptide”, y tuvieran asociado un aminoácido en el código de una letras, iii) Se descargó la lista de mapeo utilizado por ASTRAL (http://scop.berkeley.edu/astral/raf/ver=1.69), iv) se evaluaron las inconsistencias entre ambos mapeos (hubo 10 inconsistencias) y v) se adicionaron las modificaciones por estados de protonación utilizadas en dinámica molecular.. Se obtuvieron los 883 códigos para las modificaciones.

6.4.2.5. Cálculo de Métricas y su resumen

El cálculo de métricas forma parte de la T2. Se utilizaron los softwares Molprobity y DSSP, y además se utilizaron las coordenadas de los átomos para otras métricas. También se calcularon resúmenes de los datos. Con **Molprobity**, algunos de los datos que se obtuvieron son: Clashscore, ángulos phi y psi atípicos, permitidos y favorecidos por Ramachandran y evaluación de los rotámeros en valores atípicos, permitidos y favorecidos. Con **DSSP**: ASA Global, ASA y ASA relativa por residuo, estructura secundaria por la definición de DSSP y por la de ángulos phi y psi[171], y

composición de la estructura secundaria entre los distintos modelos. Con **otros cálculos, utilizando las coordenadas de los átomos de los carbonos alfa**, obtuvimos: Distancia Máxima, RMSD y Radio de giro.

6.4.2.6. Mapeo de entradas antiguas al nuevo formato

Durante el mapeo de las entradas antiguas se hizo necesario corregir algunos pequeños inconvenientes como fueron: i) poner como nombre de cadena “A” en los PDBs donde el nombre de cadena estaba vació, ii) juntar los modelos de un mismo ensamblado ya que anteriormente cada modelo era un fichero y iii) traducir los metadatos de las entradas antiguas en el nuevo formato, parte de esto se pudo realizar programáticamente pero otras cosas se tuvieron que traducir manualmente.

6.4.2.7. Deposición de nuevas entradas

Se colectaron del PDB, las estructuras que tuvieran más de 20 conformaciones y presentaban más de 10 Å de RMSD. Además se anotaron las proteínas de Disprot que tenían datos de SAXDB. También se escribió a distintos colaboradores relacionados a proteínas desordenadas para que introdujera los datos de conformaciones estructurales de proteínas desordenadas.

6.4.2.8. Reporte final en PDF

Al finalizar la deposición de un archivo pdb en esta base de datos se genera un reporte que puede ser accedido posteriormente por cualquier usuario. El reporte en formato PDF tiene los datos introducidos en el formulario de deposición, los datos de la validación, datos recolectados de los ensamblados, resúmenes de las métricas y gráficos representando los datos. Las tres primeras secciones del documento describen a la entrada utilizando los metadatos, validación y recolección de datos. La otras secciones se encuentran divididas por cada ensamblado proporcionando el resumen de métricas y las gráficos. La **Figura 43** es un ejemplo del reporte.

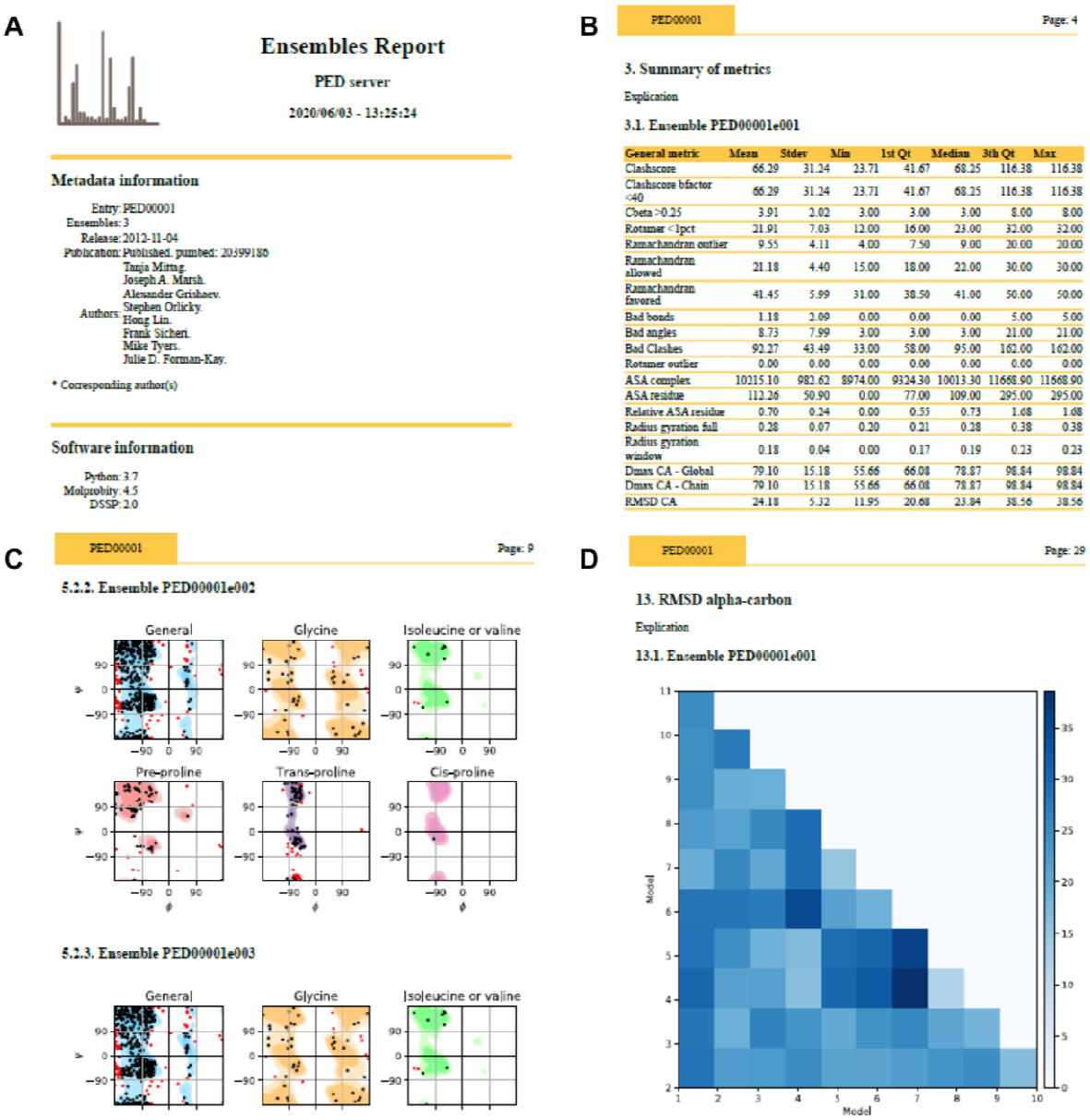


Figura 43. Ejemplo de reporte al depositar un pdb.

A: página de presentación. **B:** Listado con las métricas y sus detalles estadísticos. **C:** Ramachandran análisis. **D:** RMSD entre modelos.

6.4.3. Resultados

Al corregir las entradas con la solución propuesta el nuevo servidor cuenta inicialmente con 24 entradas provenientes de las entradas antiguas. Además se incorporan nuevas entradas provenientes del PDB, SAXDB y de colaboradores. Contando con un total de 152 entradas. Las entradas representan 215 ensamblados de 74 proteínas obteniendo un más de 290 000 conformaciones estructurales. El servidor es accesible mediante la web https://proteinensemble.org/.

6.5. Posibles SLIMs dianas para el tratamiento contra el COVID-19

Durante el período de la tesis nos vimos inmersos en la pandemia del COVID-19. Debido a esto y los conocimientos en proteínas desordenadas, participé en una colaboración sobre los motivos lineales en proteínas desordenadas relacionadas a esta enfermedad. Las proteínas incluidas en el estudios fueron Spike del COVID-19, ACE2 y las integrinas como receptor y posible correceptor en las células huéspedes respectivamente. Este estudio se encuentra aceptado para su publicación en la revista *Science Signaling* en diciembre del 2020.

La pandemia de COVID-19 es causada por el coronavirus 2 del síndrome respiratorio agudo severo (SARS-CoV-2). Ha infectado a más de dos millones de personas y ha causado alrededor de 140,000 muertes en todo el mundo a mediados de abril de 2020. El SARS-CoV-2, como el SARS-CoV[172], utiliza la enzima convertidora de angiotensina 2 (ACE2) como receptor[173–175] para unirse a las células huésped. El dominio de unión al receptor SARS-CoV-2 (RBD) de la proteína Spike interactúa con ACE2 para la entrada celular. Sin embargo, las células alveolares tipo II (AT2), los objetivos principales del SARS-CoV-2 en el pulmón[176], expresan una cantidad relativamente baja de ACE2, lo que apunta a la existencia de correceptores a los que se dirige por el virus en paralelo. Uno de esos candidatos son las integrinas que se unen a una gran variedad de ligandos que albergan un motivo de secuencia RGD, ya que un análisis reciente del RBD identificó un motivo RGD posiblemente funcional[177].

Se realizó una amplia búsqueda bibliográfica sobre el tema. Se hicieron alineamientos múltiples de secuencias de la proteína *Spike* de coronavirus, integrinas y ACE2 de homólogas humanas.

Como resultado principal del estudio fue obtener SLIMs (*Short Lineal Motif*) candidatos en las proteínas involucradas en la unión, entrada y replicación viral. El análisis de SLiMs candidatos en ACE2 e integrinas sugiere que el SARS-CoV-2 secuestra ambos receptores, cooptando sus SLiMs para impulsar la unión, entrada y replicación viral. Esto crea una oportunidad para drogar estas interacciones, o los procesos que controlan a través de Terapias Dirigidas por el Host, para prevenir la entrada viral.

Se presentó evidencia a nivel de secuencia para SLiMs en β integrinas y ACE2 con el potencial de funcionar en la unión viral, entrada y replicación para SARS-CoV-2. En general, la colección de motivos candidatos en este sistema sugiere que se podría explorar una gama de terapias dirigidas al huésped, incluida la inhibición de

RGD, la inhibición de la tirosina quinasa, la inhibición de la endocitosis y la inhibición y / o activación de la autofagia.

6.6. Conclusiones del capítulo

En el marco del proyecto IDPfun participé en 4 tareas:

- Implementación de la ontología de proteínas desordenadas en un formato estándar y modificarla haciendo uso de la literatura y de los expertos en el tema. Esta ontología se implementó en la versión 8.0 de la base de datos de Disprot.
- Análisis de la transferencia de anotaciones de desorden y términos de ontología a otras proteínas utilizando las anotaciones de proteínas desordenadas de Disprot. Se obtuvo que se pueden transferir las anotaciones a más de 23000 proteínas ortólogas resultantes de más de 600 proteínas presentes en Disprot.
- Mejora de la base de datos PED, dedicada a almacenar estructuras de proteínas desordenadas. La nueva versión (4.0) mejora su accesibilidad y la calidad de las estructuras almacenadas.
- Estudio de SLIMs en proteínas involucradas en la infección y replicación del virus del COVID-19. La proteína analizada del virus fue Spike. Las proteínas analizadas dentro de las células humanas fueron las integrinas y ACE2.

7. Conclusiones

Durante el estudio de **dependencia entre mutaciones** se encontraron 189 relaciones de exclusión y 29300 de co-mutación.

El tipo de cáncer Upper-SCC mostró un comportamiento raro, tanto en número como en tipo de dependencias, con 98.7% (29155/29300) de las mutaciones de co-mutación, casi la totalidad de las conmutaciones encontradas, por lo que se analizó separado de los demás.

En los demás tipos de cáncer se pudo apreciar que las relaciones de exclusión y/o co-mutación entre las mutaciones son diferentes por tipo de tumor. Por ejemplo el par de exclusión entre KRAS.G12 y BRAF.V600 en adenocarcinoma colorrectal, no es un par de mutaciones dependientes en adenocarcinoma de pulmón. Entre las relaciones encontradas, algunas son de distinto tipo dependiendo del tumor, por ejemplo el par KRAS.G12 y KRAS.G13 se excluyen en tumores de células no pequeñas de pulmón mientras que co-mutan en adenocarcinoma de próstata. Algunas de las relaciones descritas en la literatura se encontraron con nuestro análisis demostrando así que el protocolo utilizado es capaz de reconocerlas. Un ejemplo es la relación entre EGFR.T790 y EGFR.L858 en tumores de pulmón, siendo una causal de la otra tras la resistencia a medicamentos. También obtuvimos 3 características que nos ayudan a diferenciar las mutaciones *driver* de las *no-driver*, i) las mutaciones *driver* interactúan en más tipos de tumores, ii) las mutaciones *driver* tienden a excluirse y las *no-driver* a co-mutar y iii) los pares de mutaciones *driver* tienden a ser de la misma proteína. Apoyándonos en las tres características anteriores sugerimos 151 nuevos *drivers*.

La red de dependencias entre mutaciones se puede visualizar en http://sdmn.leloir.org.ar, es una red interactiva donde las relaciones y las mutaciones se pueden filtrar atendiendo al tejido de origen, el nombre de la proteína, tipo de relación (exclusión o co-mutación) y el tipo de la mutación (*driver*, *driver* sugerido, o si está en el CGC). Estas dependencias encontradas pueden ayudarnos a determinar si una terapia multidirigida tendría un buen efecto o no; la presencia de co-mutaciones sugiere que las combinaciones de terapias pueden ser más efectivas que una monoterapia, al menos en ciertos tumores; mientras que las exclusiones pueden indicar que la combinación de medicamentos no será beneficiosa en un porcentaje significativo de pacientes.

Durante el estudio de las **mutaciones consideradas de resistencia a medicamentos** aunque son pocos datos (mutaciones, medicamentos y pacientes)

mapeamos las mutaciones a las estructuras. Identificamos que gran parte de las mutaciones ocurren cerca del sitio de unión a la droga a menos de 6Å. En quinasas las mutaciones de resistencia ocurren también en el lazo de activación. Además analizamos las relaciones utilizando el mismo protocolo de la dependencia entre mutaciones y encontramos 20 exclusiones y 11 co-mutaciones. En este estudio encontramos la relación de co-mutación entre NRAS.Q61 y BRAF.V600 mientras que el estudio de todo COSMIC es del tipo exclusión en cánceres de piel y tiroides. Al tener pocos datos no podemos generalizar los resultados de este estudio, aunque sería interesante tener en cuenta las relaciones encontradas a la hora de suministrar al paciente una nueva terapia.

Durante el estudio de **paneles de genes/exones para predecir el TMB** encontramos paneles de genes y su modelo matemático asociado capaces de predecir con precisión el TMB en 14 tipos de cáncer. En total se obtuvieron 2301 paneles, aunque aconsejamos uno (el óptimo) por tipo de cáncer. El resto podrían ser igualmente utilizados para la predicción de TMB en caso que haya un compromiso con la cantidad de bases a secuenciar. Los paneles con la selección de genes "*our-gene"* predicen TMB con mayor precisión que los obtenidos con los genes de CGC y FO-panel según las métricas R^2 y RSE, tanto del modelo como de su validación interna y externa. De los genes presentes en nuestros paneles, un total de 1239, obtuvimos que 10% (133/1239) están entre los genes de CGC y 6% (75/1239) están entre los genes de FO-panel, evidenciando que los genes necesarios para predecir el TMB tienen baja coincidencia con los genes que están involucrados en cáncer (CGC) o los que son diana (FO-panel). También se constató que menos del 40% de los genes de nuestro panel se comparten entre los distintos tipos de cáncer, evidenciando que un único panel para todos los tipos de cáncer no es adecuado, sino paneles individuales para cada uno. Además estos modelos obtenidos con nuestra técnica por lo general necesita secuenciar menos pares de bases que el panel de FO-panel implicando un menor costo para el estudio por parte del hospital y del paciente.

Analizamos la correlación entre el TMB predicho por nuestros paneles y la sobrevida de los pacientes y, en coincidencia con el TMB calculado experimentalmente, correlaciona con la inmunoterapia, por lo que estos paneles podrían ser utilizados como un marcador a la hora de decidir si la inmunoterapia sería beneficiosa o no para el paciente.

De los 2301 paneles y modelos factibles seleccionamos 39, 14 de los cuales son el mejor panel por cada tipo de cáncer y los otros 25 paneles pueden utilizarse dependiendo de la cantidad de pares de bases que se desee secuenciar. Estos paneles ayudarían para determinar el beneficio clínico de los pacientes a la inmunoterapia.

Durante la tesis también participé en **colaboraciones sobre proteínas desordenadas** bajo el proyecto IDPfun. En la primera colaboración se obtuvo una versión 1.0 de la ontología de proteínas desordenadas en formato estándar de ontologías la cual se utiliza en la base de datos de DIsprot. En la segunda colaboración tratamos de transferir las anotaciones de desorden y términos de ontología ya asignados a la proteínas en Disprot a sus proteínas ortólogas, obteniendo así que a más de 23000 proteínas se les puede transferir anotaciones y términos de la ontología. Con la tercera colaboración creamos la versión 4.0 de PED (base de datos de estructuras de proteínas desordenadas) que tiene 15 veces más información que la versión anterior. La última colaboración se enmarca durante la pandemia de COVID-19 donde trabajamos en los motivos cortos lineales (SLIMs) en las proteínas desordenadas relacionadas al virus, tanto del huésped (proteína Spike) como del hospedador humano (ACE2 e integrinas), estos SLIMs podrían ser posibles diana terapéuticas. Todas estas colaboraciones ayudan a incrementar los conocimientos sobre las proteínas desordenadas.

8. Referencias Bibliográficas

1. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer* 924–935 (2006) doi:10.1038/nrc2013.
2. Chin, L., Hahn, W. C., Getz, G. & Meyerson, M. Making sense of cancer genomic data. *Genes Dev.* 534–555 (2011) doi:10.1101/gad.2017311.
3. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* 646–674 (2011) doi:10.1016/j.cell.2011.02.013.
4. Institute for Health Metrics and Evaluation (IHME). Data Visualizations. http://www.healthdata.org/results/data-visualizations (2018).
5. Global Health Data Exchange (GHDx). Global Burden of Disease. https://vizhub.healthdata.org/gbd-compare (2018).
6. GBD Compare | IHME Viz Hub. http://vizhub.healthdata.org/gbd-compare.
7. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
8. International Agency for Research on Cancer, Lyon, France. Global Cancer Observatory. https://gco.iarc.fr.
9. Cancer today. http://gco.iarc.fr/today/home.
10. Cancer tomorrow. http://gco.iarc.fr/tomorrow/home.
11. Wang, J.-J., Lei, K.-F. & Han, F. Tumor microenvironment: recent advances in various cancer treatments. *Eur. Rev. Med. Pharmacol. Sci.* **22**, 3855–3864 (2018).
12. Røsland, G. V. & Engelsen, A. S. T. Novel Points of Attack for Targeted Cancer Therapy. *Basic Clin. Pharmacol. Toxicol.* **116**, 9–18 (2015).
13. D'Amico, A. V. Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer. *JAMA* **280**, 969 (1998).
14. Lee, Y. T., Tan, Y. J. & Oon, C. E. Molecular targeted therapy: Treating

cancer with specificity. *Eur. J. Pharmacol.* **834**, 188–196 (2018).

15. Saijo, N. Progress in Cancer Chemotherapy with Special Stress on Molecular-targeted Therapy. *Jpn. J. Clin. Oncol.* **40**, 855–862 (2010).

16. Padma, V. V. An overview of targeted cancer therapy. *BioMedicine* **5**, 19 (2015).

17. Tsai, M.-J., Chang, W.-A., Huang, M.-S. & Kuo, P.-L. Tumor Microenvironment: A New Treatment Target for Cancer. *ISRN Biochem.* **2014**, 1–8 (2014).

18. Riley, R. S., June, C. H., Langer, R. & Mitchell, M. J. Delivery technologies for cancer immunotherapy. *Nat. Rev. Drug Discov.* **18**, 175–196 (2019).

19. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).

20. Sharma, P. & Allison, J. P. The future of immune checkpoint therapy. *Science* **348**, 56–61 (2015).

21. Topalian, S. L., Drake, C. G. & Pardoll, D. M. Immune Checkpoint Blockade: A Common Denominator Approach to Cancer Therapy. *Cancer Cell* **27**, 450–461 (2015).

22. Alsaab, H. O. *et al.* PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. *Front. Pharmacol.* **8**, 561 (2017).

23. Munn, D. H. & Bronte, V. Immune suppressive mechanisms in the tumor microenvironment. *Curr. Opin. Immunol.* **39**, 1–6 (2016).

24. Blank, C. *et al.* Blockade of PD-L1 (B7-H1) augments human tumor-specific T cell responses in vitro. *Int. J. Cancer* **119**, 317–327 (2006).

25. Webb, E. S. *et al.* Immune checkpoint inhibitors in cancer therapy. *J. Biomed. Res.* **32**, 317–326 (2018).

26. Steven, A., Fisher, S. A. & Robinson, B. W. Immunotherapy for lung cancer. *Respirology* **21**, 821–833 (2016).

27. Ellis, P. M., Vella, E. T. & Ung, Y. C. Immune Checkpoint Inhibitors for Patients With Advanced Non-Small-Cell Lung Cancer: A Systematic Review. *Clin. Lung Cancer* **18**, 444-459.e1 (2017).

28. Hodi, F. S. *et al.* Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *N. Engl. J. Med.* **363**, 711–723 (2010).

29. Brahmer, J. *et al.* Nivolumab versus Docetaxel in Advanced Squamous-Cell Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **373**, 123–135 (2015).

30. Borghaei, H. *et al.* Nivolumab versus Docetaxel in Advanced Nonsquamous Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **373**, 1627–1639 (2015).

31. Bellmunt, J. *et al.* Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma. *N. Engl. J. Med.* **376**, 1015–1026 (2017).

32. Ansell, S. M. *et al.* PD-1 Blockade with Nivolumab in Relapsed or Refractory Hodgkin's Lymphoma. *N. Engl. J. Med.* **372**, 311–319 (2015).

33. Motzer, R. J. *et al.* Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. *N. Engl. J. Med.* **373**, 1803–1813 (2015).

34. Ribas, A. & Wolchok, J. D. Cancer immunotherapy using checkpoint blockade. *Science* **359**, 1350–1355 (2018).

35. Suresh, K., Naidoo, J., Lin, C. T. & Danoff, S. Immune Checkpoint Immunotherapy for Non-Small Cell Lung Cancer. *Chest* **154**, 1416–1423 (2018).

36. Havel, J. J., Chowell, D. & Chan, T. A. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat. Rev. Cancer* **19**, 133–150 (2019).

37. Housman, G. *et al.* Drug Resistance in Cancer: An Overview. *Cancers* **6**, 1769–1792 (2014).

38. Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer.

Nature **575**, 299–309 (2019).

39. How Cancers Evolve Drug Resistance. *The Scientist Magazine*® https://www.the-scientist.com/features/how-cancers-evolve-drug-resistance-31742.

40. Lodish, H. *et al.* Mutations: Types and Causes. in *Molecular Cell Biology* (W. H. Freeman, 2000).

41. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* 789–799 (2004) doi:10.1038/nm1087.

42. Iengar, P. An analysis of substitution, deletion and insertion mutations in cancer genes. *Nucleic Acids Res.* **40**, 6401–6413 (2012).

43. Simonetti, F. L., Tornador, C., Nabau-Moreto, N., Molina-Vila, M. A. & Marino-Buslje, C. Kin-Driver: a database of driver mutations in protein kinases. *Database* bau104–bau104 (2014) doi:10.1093/database/bau104.

44. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* D941–D947 (2019) doi:10.1093/nar/gky1015.

45. International Cancer Genome Consortium. https://icgc.org/content/icgc-home-0.

46. Welcome | ICGC Data Portal. https://dcc.icgc.org/.

47. The Cancer Genome Atlas Program - National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (2018).

48. Greenblatt, M. S., Bennett, W. P., Hollstein, M. & Harris, C. C. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res.* **54**, 4855–4878 (1994).

49. Pfeifer, G. P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21**, 7435–7451 (2002).

50. Pfeifer, G. P., You, Y.-H. & Besaratinia, A. Mutations induced by ultraviolet light. *Mutat. Res. Mol. Mech. Mutagen.* **571**, 19–31 (2005).

51. Mace, K. Aflatoxin B1-induced DNA adduct formation and p53 mutations in CYP450- expressing human liver cell lines. *Carcinogenesis* **18**, 1291–1297 (1997).

52. Nedelko, T., Arlt, V. M., Phillips, D. H. & Hollstein, M. *TP53* mutation signature supports involvement of aristolochic acid in the aetiology of endemic nephropathy-associated tumours. *Int. J. Cancer* **124**, 987–990 (2009).

53. Metzker, M. L. Sequencing technologies — the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).

54. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).

55. R Foundation for Statistical Computing, R. C. T. *R: A Language and Environment for Statistical Computing*. (R Core Team, 2018).

56. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

57. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

58. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

59. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. (2019).

60. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

61. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical

computing devoted to biological sequences retrieval and analysis. in *Structural approaches to sequence evolution: molecules, networks, populations* (eds. Bastolla, U., Porto, M., Roman, H. E. & Vendruscolo, M.) 207–232 (Springer, 2007).

62. Hubisz, M., Pollard, K. & Siepel, A. *rphast: Interface to 'PHAST' Software for Comparative Genomics*. (2018).

63. Kuhn, M. *et al. caret: Classification and Regression Training*. (CRAN R project, 2018).

64. Therneau, T. M. *A Package for Survival Analysis in S*. (2015).

65. Therneau, T. M. & Grambsch, P. M. *Modeling survival data: extending the Cox model*. (Springer, 2000).

66. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).

67. Wilke, C. O. *ggridges: Ridgeline Plots in 'ggplot2'*. (2020).

68. Chen, H. *VennDiagram: Generate High-Resolution Venn and Euler Plots*. (2018).

69. Kassambara, A., Kosinski, M. & Biecek, P. *survminer: Drawing Survival Curves using 'ggplot2'*. (2019).

70. Sachs, M. C. **plotROC** : A Tool for Plotting ROC Curves. *J. Stat. Softw.* **79**, (2017).

71. Couture-Beil, A. rjson: JSON for R. (2018).

72. Wickham, H. httr: Tools for Working with URLs and HTTP. (2019).

73. Rossum, G. van, Drake, F. L. & Van Rossum, G. *The Python language reference*. (Python Software Foundation, 2010).

74. Bassi, S. *Python for bioinformatics*. (CRC Press, 2010).

75. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

76. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

77. Caswell, T. A. *et al. matplotlib/matplotlib v3.1.3*. (Zenodo, 2020). doi:10.5281/ZENODO.3633844.

78. Barabási, A.-L. & Pósfai, M. *Network science*. (Cambridge University Press, 2016).

79. Walpole, R. E. *Probabilidad y estadística para ingeniería y ciencias*. (Pearson, 2012).

80. Ochoa, S., Martínez□Pérez, E., Zea, D. J., Molina□Vila, M. A. & Marino□Buslje, C. Comutation and exclusion analysis in human tumors: A tool for cancer biology studies and for rational selection of multitargeted therapeutic approaches. *Hum. Mutat.* 413–425 (2019) doi:10.1002/humu.23705.

81. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).

82. Compagno, M. *et al.* Mutations of multiple genes cause deregulation of NF-κB in diffuse large B-cell lymphoma. *Nature* **459**, 717–721 (2009).

83. Jones, S. *et al.* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science* **321**, 1801–1806 (2008).

84. Kim, Y.-A., Madan, S. & Przytycka, T. M. WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* btw242 (2016) doi:10.1093/bioinformatics/btw242.

85. Thomas, R. K. *et al.* High-throughput oncogene mutation profiling in human cancer. *Nat. Genet.* 347–351 (2007) doi:10.1038/ng1975.

86. Teschendorff, A. E. & Caldas, C. The breast cancer somatic 'muta-ome': tackling the complexity. *Breast Cancer Res.* **11**, 301 (2009).

87. Bommi-Reddy, A. *et al.* Kinase requirements in human cells: III. Altered

kinase requirements in VHL-/- cancer cells detected in a pilot synthetic lethal screen. *Proc. Natl. Acad. Sci.* **105**, 16484–16489 (2008).

88. Pratilas, C. A., Xing, F. & Solit, D. B. Targeting Oncogenic BRAF in Human Cancer. in *Therapeutic Kinase Inhibitors* (eds. Mellinghoff, I. K. & Sawyers, C. L.) 83–98 (Springer Berlin Heidelberg, 2010). doi:10.1007/82_2011_162.

89. Babur, Ö. *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* **16**, 45 (2015).

90. Cui, Q. A Network of Cancer Genes with Co-Occurring and Anti-Co-Occurring Mutations. *PLoS ONE* **5**, e13180 (2010).

91. Kim, Y.-A., Cho, D.-Y., Dao, P. & Przytycka, T. M. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* **31**, i284–i292 (2015).

92. Rubio-Perez, C. *et al.* In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell* **27**, 382–396 (2015).

93. Yeang, C.-H., McCormick, F. & Levine, A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J.* **22**, 2605–2622 (2008).

94. Cai, C. Q. *et al.* Epidermal growth factor receptor activation in prostate cancer by three novel missense mutations. *Oncogene* **27**, 3201–3210 (2008).

95. Nussinov, R. & Tsai, C.-J. 'Latent drivers' expand the cancer mutational landscape. *Curr. Opin. Struct. Biol.* **32**, 25–32 (2015).

96. Heinrich, M. C. *et al.* Kinase Mutations and Imatinib Response in Patients With Metastatic Gastrointestinal Stromal Tumor. *J. Clin. Oncol.* **21**, 4342–4349 (2003).

97. Wang, X., Goldstein, D., Crowe, P. & Yang, J.-L. Next-generation EGFR/HER tyrosine kinase inhibitors for the treatment of patients with non-small-cell lung

cancer harboring EGFR mutations: a review of the evidence. *OncoTargets Ther.* **Volume 9**, 5461–5473 (2016).

98. Gençler, B. & Gönül, M. Cutaneous Side Effects of BRAF Inhibitors in Advanced Melanoma: Review of the Literature. *Dermatol. Res. Pract.* **2016**, 1–6 (2016).

99. Tian, T., Olson, S., Whitacre, J. M. & Harding, A. The origins of cancer robustness and evolvability. *Integr Biol* 17–30 (2011) doi:10.1039/C0IB00046A.

100. Simeone, E. *et al.* Combination Treatment of Patients with BRAF-Mutant Melanoma: A New Standard of Care. *BioDrugs* **31**, 51–61 (2017).

101. Mokhtari, R. B. *et al.* Combination therapy in combating cancer. *Oncotarget* **8**, (2017).

102. Molina-Vila, M. A. *et al.* Activating Mutations Cluster in the “Molecular Brake” Regions of Protein Kinases and Do Not Associate with Conserved or Catalytic Residues. *Hum. Mutat.* 318–328 (2014) doi:10.1002/humu.22493.

103. Moura, M. M., Cavaco, B. M. & Leite, V. RAS proto-oncogene in medullary thyroid carcinoma. *Endocr. Relat. Cancer* R235–R252 (2015) doi:10.1530/ERC-15-0070.

104. Pao, W. & Girard, N. New driver mutations in non-small-cell lung cancer. *Lancet Oncol.* 175–180 (2011) doi:10.1016/S1470-2045(10)70087-5.

105. Olivier, M., Hollstein, M. & Hainaut, P. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harb. Perspect. Biol.* a001008–a001008 (2010) doi:10.1101/cshperspect.a001008.

106. De Roock, W. *et al.* Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol.* 753–762 (2010) doi:10.1016/S1470-2045(10)70130-3.

107. Stelow, E. B. & Mills, S. E. Squamous Cell Carcinoma Variants of the Upper Aerodigestive Tract. *Pathol. Patterns Rev.* S96–S109 (2005) doi:10.1309/CR5JXUY3J2YGTC1D.

108. Chen, Z. *et al.* EGFR somatic doublets in lung cancer are frequent and generally arise from a pair of driver mutations uncommonly seen as singlet mutations: one-third of doublets occur at five pairs of amino acids. *Oncogene* 4336–4343 (2008) doi:10.1038/onc.2008.71.

109. Cho, N.-Y. *et al.* BRAF andKRAS mutations in prostatic adenocarcinoma. *Int. J. Cancer* 1858–1862 (2006) doi:10.1002/ijc.22071.

110. Park, J. Y. *et al.* BRAF and RAS Mutations in Follicular Variants of Papillary Thyroid Carcinoma. *Endocr. Pathol.* 69–76 (2013) doi:10.1007/s12022-013-9244-0.

111. Schmitt, M. W., Loeb, L. A. & Salk, J. J. The influence of subclonal resistance mutations on targeted cancer therapy. *Nat. Rev. Clin. Oncol.* 335–347 (2016) doi:10.1038/nrclinonc.2015.175.

112. Zou, M. *et al.* Concomitant RAS, RET/PTC, or BRAF Mutations in Advanced Stage of Papillary Thyroid Carcinoma. *Thyroid* 1256–1266 (2014) doi:10.1089/thy.2013.0610.

113. Wang, M., Yang, C., Zhang, L. & Schaar, D. G. Molecular Mutations and Their Cooccurrences in Cytogenetically Normal Acute Myeloid Leukemia. *Stem Cells Int.* 1–11 (2017) doi:10.1155/2017/6962379.

114. Hartmann, K. *et al.* Novel Germline Mutation of KIT Associated With Familial Gastrointestinal Stromal Tumors and Mastocytosis. *Gastroenterology* 1042–1046 (2005) doi:10.1053/j.gastro.2005.06.060.

115. Rumi, E. *et al.* JAK2 or CALR mutation status defines subtypes of essential thrombocythemia with substantially different clinical course and outcomes. *Blood*

1544–1551 (2014) doi:10.1182/blood-2013-11-539098.

116. Smalley, K. S. M. & Flaherty, K. T. Integrating BRAF/MEK inhibitors into combination therapy for melanoma. *Br. J. Cancer* 431–435 (2009) doi:10.1038/sj.bjc.6604891.

117. Arulananda, S. *et al.* Combination Osimertinib and Gefitinib in C797S and T790M EGFR-Mutated Non–Small Cell Lung Cancer. *J. Thorac. Oncol.* 1728–1732 (2017) doi:10.1016/j.jtho.2017.08.006.

118. Dang, L., Yen, K. & Attar, E. C. IDH mutations in cancer and progress toward development of targeted therapeutics. *Ann. Oncol.* 599–608 (2016) doi:10.1093/annonc/mdw013.

119. Duffy, M. J. *et al.* p53 as a target for the treatment of cancer. *Cancer Treat. Rev.* 1153–1160 (2014) doi:10.1016/j.ctrv.2014.10.004.

120. Pérez-Ramírez, C., Cañadas-Garre, M., Molina, M. Á., Faus-Dáder, M. J. & Calleja-Hernández, M. Á. PTEN and PI3K/AKT in non-small-cell lung cancer. *Pharmacogenomics* 1843–1862 (2015) doi:10.2217/pgs.15.122.

121. Siena, S., Sartore-Bianchi, A., Di Nicolantonio, F., Balfour, J. & Bardelli, A. Biomarkers Predicting Clinical Outcome of Epidermal Growth Factor Receptor–Targeted Therapy in Metastatic Colorectal Cancer. *JNCI J. Natl. Cancer Inst.* 1308–1324 (2009) doi:10.1093/jnci/djp280.

122. Paraiso, K. H. T. *et al.* PTEN Loss Confers BRAF Inhibitor Resistance to Melanoma Cells through the Suppression of BIM Expression. *Cancer Res.* 2750–2760 (2011) doi:10.1158/0008-5472.CAN-10-2954.

123. Monticone, S. *et al.* Effect of KCNJ5 Mutations on Gene Expression in Aldosterone-Producing Adenomas and Adrenocortical Cells. *J. Clin. Endocrinol. Metab.* E1567–E1572 (2012) doi:10.1210/jc.2011-3132.

124. Pettersen, E. F. *et al.* UCSF Chimera?A visualization system for exploratory

research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

125. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 124–128 (2015) doi:10.1126/science.aaa1348.

126. Alspach, E. *et al.* MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature* **574**, 696–701 (2019).

127. Gubin, M. M., Artyomov, M. N., Mardis, E. R. & Schreiber, R. D. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.* **125**, 3413–3421 (2015).

128. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell* **160**, 48–61 (2015).

129. Chan, T. A. *et al.* Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* **30**, 44–56 (2019).

130. Hellmann, M. D. *et al.* Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *N. Engl. J. Med.* **378**, 2093–2104 (2018).

131. Wang, F. *et al.* Safety, efficacy and tumor mutational burden as a biomarker of overall survival benefit in chemo-refractory gastric cancer treated with toripalimab, a PD-1 antibody in phase Ib/II clinical trial NCT02915432. *Ann. Oncol.* **30**, 1479–1486 (2019).

132. Ready, N. *et al.* First-Line Nivolumab Plus Ipilimumab in Advanced Non-Small-Cell Lung Cancer (CheckMate 568): Outcomes by Programmed Death Ligand 1 and Tumor Mutational Burden as Biomarkers. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **37**, 992–1000 (2019).

133. Mishima, S. *et al.* Clinicopathological and molecular features of responders to

nivolumab for patients with advanced gastric cancer. *J. Immunother. Cancer* **7**, 24 (2019).

134. Morrison, C. *et al.* Predicting response to checkpoint inhibitors in melanoma beyond PD-L1 and mutational burden. *J. Immunother. Cancer* **6**, 32 (2018).

135. Fizazi, K. *et al.* LBA15_PRA phase III trial of empiric chemotherapy with cisplatin and gemcitabine or systemic treatment tailored by molecular gene expression analysis in patients with carcinomas of an unknown primary (CUP) site (GEFCAPI 04). *Ann. Oncol.* **30**, mdz394 (2019).

136. FoundationOne. FoundationOne Panel 315 genes. https://www.foundationmedicineasia.com/dam/assets/pdf/FOne_Current_Gene_List.pdf.

137. Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).

138. Balar, A. V. *et al.* Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. *Lancet Lond. Engl.* **389**, 67–76 (2017).

139. Budczies, J. *et al.* Optimizing panel-based tumor mutational burden (TMB) measurement. *Ann. Oncol.* **30**, 1496–1506 (2019).

140. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* 2189–2199 (2014) doi:10.1056/NEJMoa1406498.

141. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).

142. Kim, N., Hong, Y., Kwon, D. & Yoon, S. Somatic mutaome profile in human cancer tissues. *Genomics Inform.* **11**, 239–244 (2013).

143. Wolff, R. K. *et al.* Mutation analysis of adenomas and carcinomas of the colon: Early and late drivers. *Genes. Chromosomes Cancer* **57**, 366–376 (2018).

144. Wu, Y. *et al.* Orchestrating a biomarker panel with lncRNAs and mRNAs for predicting survival in pancreatic ductal adenocarcinoma. *J. Cell. Biochem.* **119**, 7696–7706 (2018).

145. Zhang, Y.-H. *et al.* Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets. *Oncotarget* **8**, 87494–87511 (2017).

146. Marouf, C. *et al.* Analysis of functional germline variants in APOBEC3 and driver genes on breast cancer risk in Moroccan study population. *BMC Cancer* **16**, 165 (2016).

147. Tan, H., Bao, J. & Zhou, X. Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. *Sci. Rep.* **5**, 12566 (2015).

148. Tukey, J. W. *Exploratory data analysis*. (Reading, Mass. : Addison-Wesley Pub. Co., 1977).

149. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

150. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.* **112**, 15898–15903 (2015).

151. Mistry, J. *et al.* The challenge of increasing Pfam coverage of the human proteome. *Database* **2013**, (2013).

152. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533 (2002).

153. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582 (2002).

154. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **579**, 3346–3354 (2005).

155. van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**, 6589–6631 (2014).

156. Piovesan, D. *et al.* DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227 (2017).

157. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

158. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

159. Musen, M. A. The protégé project: a look back and a look forward. *AI Matters* **1**, 4–12 (2015).

160. Hatos, A. *et al.* DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* gkz975 (2019) doi:10.1093/nar/gkz975.

161. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

162. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

163. Altenhoff, A. M. *et al.* The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.* **46**, D477–D485 (2018).

164. Nevers, Y. *et al.* OrthoInspector 3.0: open portal for comparative genomics. *Nucleic Acids Res.* **47**, D411–D418 (2019).

165. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences: Clustal Omega for Many Protein Sequences. *Protein Sci.* **27**, 135–145 (2018).

166. Sievers, F. *et al.* Fast, scalable generation of high□quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

167. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–

780 (2013).

168. Katoh, K. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

169. Thompson, J. D., Plewniak, F., Ripp, R., Thierry, J.-C. & Poch, O. Towards a reliable objective function for multiple sequence alignments 1 1Edited by J. Karn. *J. Mol. Biol.* **314**, 937–951 (2001).

170. Varadi, M. *et al.* pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* **42**, D326–D335 (2014).

171. Ozenne, V. *et al.* Mapping the Potential Energy Landscape of Intrinsically Disordered Proteins at Amino Acid Resolution. *J. Am. Chem. Soc.* **134**, 15138–15148 (2012).

172. Li, W. *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).

173. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).

174. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).

175. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

176. Zou, X. *et al.* Single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to 2019-nCoV infection. *Front. Med.* **14**, 185–192 (2020).

177. Sigrist, C. J., Bridge, A. & Le Mercier, P. A potential role for integrins in host cell entry by SARS-CoV-2. *Antiviral Res.* **177**, 104759 (2020).

9. Anexos

9.1. Anexo 1: Mutaciones *driver* de Kin-Driver y literatura

ALK.F1174	ALK.F1245	ALK.G1128	ALK.I1171	ALK.M1166
ALK.R1275	ALK.Y1278	BRAF.E586	BRAF.F595	BRAF.G464
BRAF.G469	BRAF.K499	BRAF.K601	BRAF.L485	BRAF.L597
BRAF.Q257	BRAF.S467	BRAF.T599	BRAF.V600	CHEK2.I251
CHEK2.R180	CHEK2.S428	CHEK2.Y327	CHEK2.Y390	CSF1R.L301
CSF1R.Y571	CSF1R.Y969	EGFR.A839	EGFR.D855	EGFR.E709
EGFR.E746>	EGFR.E746del	EGFR.E749	EGFR.E865	EGFR.F856
EGFR.G696	EGFR.G719	EGFR.G779	EGFR.G857	EGFR.G863
EGFR.K745del	EGFR.K806	EGFR.K846	EGFR.L747	EGFR.L747>
EGFR.L747del	EGFR.L814	EGFR.L833	EGFR.L838	EGFR.L858
EGFR.L861	EGFR.N700	EGFR.P699	EGFR.Q812	EGFR.R748del
EGFR.R776	EGFR.R836	EGFR.S720	EGFR.S752del	EGFR.S768
EGFR.T751>	EGFR.T790	EGFR.V689	EGFR.V765	EGFR.V843
EPHA2.R721	EPHA5.G582	ERBB2.A775ins	ERBB2.G776	ERBB2.G776>
ERBB2.M774ins	ERBB2.N857	ERBB2.P780ins	FGFR2.E565	FGFR2.K641
FGFR2.K659	FGFR2.N549	FGFR2.S252	FGFR2.S372	FGFR2.Y375
FGFR3.A391	FGFR3.G370	FGFR3.G697	FGFR3.K650	FGFR3.R248
FGFR3.S249	FGFR3.S371	FGFR3.Y373	FLT3.D586ins	FLT3.D593ins
FLT3.D600ins	FLT3.D835	FLT3.D839	FLT3.E596ins	FLT3.E598ins
FLT3.E604ins	FLT3.E608ins	FLT3.E611ins	FLT3.F590	FLT3.F590>
FLT3.F590ins	FLT3.F594	FLT3.F594ins	FLT3.F605ins	FLT3.F612ins
FLT3.G613ins	FLT3.I836	FLT3.I836del	FLT3.K602ins	FLT3.K614ins
FLT3.K663	FLT3.L601ins	FLT3.L610ins	FLT3.N587ins	FLT3.N609ins
FLT3.N676	FLT3.N841	FLT3.P606ins	FLT3.Q580ins	FLT3.R595ins
FLT3.R607ins	FLT3.T582ins	FLT3.V579	FLT3.V592	FLT3.V592ins
FLT3.W603ins	FLT3.Y591	FLT3.Y591ins	FLT3.Y597ins	FLT3.Y599ins
HRAS.G12	HRAS.G12ins	HRAS.G13	HRAS.K117	HRAS.Q61
JAK1.A634	JAK1.V658	JAK2.E543del	JAK2.F537>	JAK2.H538>
JAK2.I540>	JAK2.N542del	JAK2.R541>	JAK2.R564	JAK2.R683
JAK2.T875	JAK2.V617	JAK3.A572	JAK3.I87	JAK3.P132
JAK3.Q501	JAK3.R657	JAK3.V715	JAK3.V722	KIT.D579del
KIT.D579ins	KIT.D816	KIT.E554>	KIT.E554del	KIT.E583ins
KIT.F591ins	KIT.H580ins	KIT.K550>	KIT.K550del	KIT.K558
KIT.K558>	KIT.K558del	KIT.K581ins	KIT.K642	KIT.L576
KIT.L576del	KIT.L589ins	KIT.M552>	KIT.M552del	KIT.N564del
KIT.N567del	KIT.P551>	KIT.P551del	KIT.P585ins	KIT.Q556>
KIT.Q556del	KIT.R586ins	KIT.V555del	KIT.V559	KIT.V559del
KIT.V560	KIT.V560del	KIT.V569del	KIT.W557	KIT.W557>
KIT.W557del	KIT.Y503ins	KIT.Y553>	KIT.Y553del	KIT.Y570del
KRAS.A146	KRAS.E63	KRAS.E63del	KRAS.G12	KRAS.G12fs
KRAS.G12ins	KRAS.G13	KRAS.G13>	KRAS.G13ins	KRAS.K117
KRAS.Q61	MAP2K1.F53	MAP2K1.Y130	MAP2K2.F57	MET.D1246

MET.M1268	MET.N1118	MET.T1010	MET.V1110	MET.Y1248
MET.Y1253	NRAS.A18	NRAS.G12	NRAS.G13	NRAS.G138
NRAS.Q61	NRAS.Q61>	PDGFRA.D842	PDGFRA.D842>	PDGFRA.D842del
PDGFRA.H845>	PDGFRA.I843>	PDGFRA.I843del	PDGFRA.M844del	PDGFRA.N659
PDGFRA.R560del	PDGFRA.S566>	PDGFRA.V561	PDGFRA.Y849	PIK3CA.E542
PIK3CA.E545	PIK3CA.H1047	PRKCG.P524	PRKCG.R659	RAF1.L613
RAF1.P261	RAF1.S257	RAF1.S259	RAF1.V263	RET.A883
RET.C609	RET.C618	RET.C620	RET.C630	RET.C634
RET.C634ins	RET.E632>	RET.E632del	RET.E768	RET.M918
RET.S891	RET.V804	RET.Y791	STK11.D176	STK11.D194
STK11.E98del	STK11.F354	STK11.G163	STK11.G171	STK11.G215
STK11.G242	STK11.I177	STK11.K78	STK11.L182del	STK11.N181
STK11.P324	STK11.Q123	STK11.R304	STK11.R86	STK11.S216
STK11.T367	STK11.W239	STK11.W308	TEK.R849	TEK.R915
TEK.R918	TNK2.E346	TNK2.M409	TNK2.R34	TNK2.R99
TP53.R175	TP53.R248	TP53.R273	TTK.D697	

9.2. Anexo 2: Mutaciones *driver* sugeridas

BCOR1.F111	BCOR1.L1132	BCOR1.V1660	BRAF.D594	BRAF.G466
BRAF.V600>	CALR.E364fs	CALR.K368fs	CALR.K385fs	CALR.L367fs
CEBPA.D107fs	CEBPA.E309ins	CEBPA.E316ins	CEBPA.G36fs	CEBPA.H24fs
CEBPA.I62fs	CEBPA.K304ins	CEBPA.K313ins	CEBPA.L315ins	CEBPA.P23fs
CEBPA.Q305ins	CEBPA.Q312ins	CEBPA.Q83fs	CEBPA.S61fs	CEBPA.T310ins
CEBPA.V308ins	CRTC1.S572fs	CRTC1.S588fs	CTNNB1.D32	CTNNB1.G34
CTNNB1.S33	CTNNB1.S37	CTNNB1.S45	CTNNB1.T41	EGFR.A871
EGFR.D770ins	EGFR.G598	EGFR.R108	EGFR.V769ins	FBXW7.R278*
FBXW7.S582	FIP1L1.R481fs	FIP1L1.R487fs	GNA11.Q209	GNAQ.Q209
GNAS1.Q227	GNAS1.R201	H3F3B.G35	H3F3B.K28	HLA-A.D251
HLA-A.E176	HLA-A.E277	HLA-A.L180*	HLA-A.N151	HLA-A.N90
HLA-A.Q86	HLA-A.T187	HLA-A.T345	HLA-A.W191	HRAS2.Q61
IDH1.R132	IDH2.R140	IDH2.R172	JAK2.K539	JAK2.V615
KCNJ5.G151	KCNJ5.L168	KIT.D419del	KIT.N822	KIT.T670
KIT.V654	KIT.Y823	KLF4.K409	KMT2C.C988	KMT2C.G838
KMT2C.I817fs	KMT2C.L291	KMT2C.N729	KMT2C.Q755*	KMT2C.R2481
KMT2C.R909	KMT2C.S772	KMT2C.T316	KMT2C.Y987	MED12.E33ins
MED12.G44	MED12.L36	MED12.Q43	MPL.S505	MPL.W515
NCOR1.E191	NCOR1.G5	NCOR1.K178	NCOR1.R190*	NCOR1.S172
NCOR1.Y20	NOTCH1.D571	NOTCH1.D573	NOTCH1.L1600	NOTCH1.P2514fs
NOTCH1.T194	NOTCH1.T311	NOTCH1.T349	NOTCH2.C19	NOTCH2.E38
NOTCH2.N46	NPM1.W288fs	NPM1.W290fs	PDE4DIP.A1066	PDE4DIP.A167
PDE4DIP.A1742	PDE4DIP.A1757	PDE4DIP.A487fs	PDE4DIP.D1910	PDE4DIP.E410
PDE4DIP.F1013	PDE4DIP.H482	PDE4DIP.I109	PDE4DIP.K1359	PDE4DIP.L1727
PDE4DIP.R1504	PDE4DIP.R171	PDE4DIP.R1867	PDE4DIP.R2291	PDE4DIP.R25
PDE4DIP.R622*	PDE4DIP.R681	PDE4DIP.S275	PDE4DIP.S536	PDE4DIP.T2297
PDE4DIP.V1736	PDE4DIP.W1396	PIK3CA.N345	POLE.P286	POLE.V411
PTEN.C105	PTEN.K267fs	PTEN.N323fs	PTEN.R173	PTPN11.A72
PTPN11.D61	PTPN11.E76	SETBP1.A222	SETBP1.T228fs	SF3B1.K666
SF3B1.K700	TP53.G245	TP53.R282	TRAF7.N520	VHL.P81
VHL.R167				

9.3. Anexo 3: Comparación de los modelos consensuados

A: prueba de Kruskal-Wallis por tipo de cáncer en las métricas

R2/RSE v. interna/externa, donde v. representa validación.

Tumor	Metrica	p.value		Metrica	p.value		Metrica	p.value
Glioma	R2 modelo	4.08E-10		R2 v. interna	1.90E-11		R2 v. externa	1.47E-04
	RSE modelo	1.19E-09		RSE v. interna	3.49E-11		RSE v. externa	3.07E-03
Tumor de piel no-melanoma	R2 modelo	1.28E-07		R2 v. interna	2.66E-09		R2 v. externa	4.70E-09
	RSE modelo	3.05E-06		RSE v. interna	8.37E-09		RSE v. externa	3.20E-07
Tumor colorectal	R2 modelo	1.31E-89		R2 v. interna	5.01E-103		R2 v. externa	2.37E-79
	RSE modelo	3.70E-85		RSE v. interna	3.33E-100		RSE v. externa	7.25E-89
Melanoma	R2 modelo	2.25E-40		R2 v. interna	1.19E-47		R2 v. externa	1.31E-43
	RSE modelo	2.70E-37		RSE v. interna	2.29E-47		RSE v. externa	2.28E-45
Hígado	R2 modelo	3.58E-27		R2 v. interna	3.39E-32		R2 v. externa	4.60E-42
	RSE modelo	9.36E-25		RSE v. interna	2.15E-31		RSE v. externa	4.60E-42
Tumor gástrico	R2 modelo	9.77E-09		R2 v. interna	7.11E-10		R2 v. externa	3.60E-14
	RSE modelo	3.55E-08		RSE v. interna	1.89E-09		RSE v. externa	3.60E-14
Carcinoma seroso de ovario	R2 modelo	4.40E-06		R2 v. interna	4.40E-06		R2 v. externa	4.40E-06
	RSE modelo	4.40E-06		RSE v. interna	4.40E-06		RSE v. externa	5.38E-04

B: prueba post-hoc de Dunn por tipo de cáncer en las métricas. Se utiliza como control la estrategia our-gene, porque tiene mejor mediana en cada tumor en la mayoría de las métricas.

		R^2			RSE		
Tipo de cáncer	**Estrategia**	modelo	v. interna	v. externa	modelo	v. interna	v. externa
Glioma	**CGC-gene**	1.2E-0	3.5E-119	3.1E-03	4.1E-10	1.9E-11	1.5E-04
Tumor colorectal	**our-exon**	9.4E-17	5.5E-21	4.1E-25	1.3E-17	1.5E-20	8.4E-47
	CGC-gene	6.4E-33	3.4E-34	2.0E-16	1.6E-34	1.2E-35	6.3E-02
	CGC-exon	3.0E-46	3.1E-59	6.7E-73	5.6E-49	2.5E-60	5.7E-46
	FO-panel-gene	5.2E-51	3.8E-58	1.6E-33	7.1E-54	5.7E-60	7.3E-19
Hígado	**CGC-gene**	9.4E-25	2.2E-31	4.6E-42	3.6E-27	3.4E-32	4.6E-42
Carcinoma s. ovario	**CGC-gene**	4.4E-06	4.4E-06	5.4E-04	4.4E-06	4.4E-06	4.4E-06
Tumor de piel no-melanoma	**our-exon**	1.1E-02	3.2E-03	2.2E-07	3.2E-03	3.9E-03	8.5E-07
	CGC-gene	1.5E-03	2.6E-05	9.5E-01	4.1E-04	2.1E-05	4.4E-01
	FO-panel-gene	5.7E-06	3.2E-08	6.1E-01	1.5E-07	8.7E-09	1.5E-01
Melanoma	**our-exon**	8.6E-09	2.3E-13	3.9E-33	2.9E-09	1.3E-12	4.1E-30
	CGC-gene	8.4E-08	2.7E-08	9.5E-03	3.6E-09	6.8E-09	7.5E-02
	CGC-exon	2.5E-23	3.4E-34	7.0E-25	5.7E-25	2.6E-29	7.4E-24
	FO-panel-gene	3.4E-28	3.3E-29	1.1E-04	4.8E-31	4.6E-35	3.1E-03
Tumor gástrico	**CGC-gene**	3.6E-08	1.9E-09	3.6E-14	9.8E-09	7.1E-10	3.6E-14

Carcinoma s. ovario: carcinoma seroso de ovario

9.4. Anexo 4: Comparación de métodos de alineamiento

A: prueba Kruskal-Wallis

Tipos alineamientos	P.value Kruskal Wallis	Score de NorMd	Post hoc Dunn Test
Completo	0.00	Figura 23A	Anexo 4B
Regiones desordenadas	9.14e-181	Figura 23B	Anexo 4C
Regiones con términos de ontología	6.11e-184	Figura 23C	Anexo 4D

En los anexos 4B-D, los 3 métodos de alineamiento se representan con C-O, M-D y M-L son Crustal-Omega, MAFFT-Default y MAFFT-linsi respectivamente. Los valores 60, 60_50 y 60_75 representan los clúster al 60% de identidad con cubrimiento: sin especificar, al 50% y al 75%, similar ocurre con 80, 80_50 y 80_75 pero al 80% de identidad. Las comparaciones, en rojo sin diferencias significativas (p.value > 0.05); azul, en el mismo método con diferencias significativas; naranja, de diferentes % de identidad y métodos con diferencias significativas.

B: p.value de prueba post-hoc Dunn en alineamiento completo

	C-O_60	C-O_60_50	C-O_60_75	C-O_80	C-O_80_50	C-O_80_75	M-D_60	M-D_60_50	M-D_60_75	M-D_80	M-D_80_50	M-D_80_75	M-L_60	M-L_60_50	M-L_60_75	M-L_80	M-L_80_50
C-O_60_50	0.53																
C-O_60_75	0.83	0.38															
C-O_80	0.00	0.00	0.00														
C-O_80_50	0.00	0.00	0.00	**0.81**													
C-O_80_75	0.00	0.00	0.00	**0.86**	**0.69**												
M-D_60	0.00	0.03	0.00	0.00	0.00	0.00											
M-D_60_50	0.03	0.16	0.02	0.00	0.00	0.00	0.47										
M-D_60_75	0.81	0.36	0.98	0.00	0.00	0.00	0.00	0.02									
M-D_80	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00								
M-D_80_50	0.00	0.00	0.00	0.05	**0.09**	0.03	0.00	0.00	0.00	**0.38**							
M-D_80_75	0.00	0.00	0.00	**0.81**	**0.64**	**0.96**	0.00	0.00	0.00	0.00	0.02						
M-L_60	0.02	0.12	0.01	0.00	0.00	0.00	0.58	0.89	0.01	0.00	0.00	0.00					
M-L_60_50	0.10	0.36	0.05	0.00	0.00	0.00	0.23	0.70	0.05	0.00	0.00	0.00	0.60				
M-L_60_75	0.61	0.22	0.79	0.00	0.00	0.00	0.00	0.01	0.80	0.00	0.00	0.00	0.00	0.02			
M-L_80	0.00	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	**0.78**	**0.61**	0.00	0.00	0.00	0.00		
M-L_80_50	0.00	0.00	0.00	**0.10**	**0.17**	**0.06**	0.00	0.00	0.00	**0.22**	**0.79**	**0.05**	0.00	0.00	0.00	**0.39**	
M-L_80_75	0.00	0.00	0.00	**0.81**	**0.64**	**0.96**	0.00	0.00	0.00	0.00	0.02	**1.00**	0.00	0.00	0.00	0.00	**0.05**

C: p.value de prueba post-hoc Dunn en alineamiento de regiones desordenadas

	C-O_60	C-O_60_50	C-O_60_75	C-O_80	C-O_80_50	C-O_80_75	M-D_60	M-D_60_50	M-D_60_75	M-D_80	M-D_80_50	M-D_80_75	M-L_60	M-L_60_50	M-L_60_75	M-L_80	M-L_80_50
C-O_60_50	0.034																
C-O_60_75	0	0.035															
C-O_80	0	0	0														
C-O_80_50	0	0	0	0.027													
C-O_80_75	0	0	0	0	0.049												
M-D_60	0.02	0.821	0.057	0	0	0											
M-D_60_50	0	0.02	0.818	0	0	0	0.034										
M-D_60_75	0	0	0.005	0.009	0	0	0	0.01									
M-D_80	0	0	0	**0.304**	**0.247**	0.002	0	0	0								
M-D_80_50	0	0	0	0.001	**0.352**	**0.319**	0	0	0	0.034							
M-D_80_75	0	0	0	0	0.003	**0.321**	0	0	0	0	0.042						
M-L_60	0.056	0.821	0.02	0	0	0	0.68	0.01	0	0	0	0					
M-L_60_50	0	0.064	0.799	0	0	0	0.102	0.635	0.002	0	0	0	0.038				
M-L_60_75	0	0	0.031	0.001	0	0	0	0.052	0.532	0	0	0	0	0.015			
M-L_80	0	0	0	**0.468**	**0.142**	0.001	0	0	0.001	**0.782**	0.016	0	0	0	0		
M-L_80_50	0	0	0	0.004	**0.533**	**0.19**	0	0	0	**0.067**	**0.781**	0.02	0	0	0	0.034	
M-L_80_75	0	0	0	0	0.007	**0.469**	0	0	0	0	**0.076**	**0.799**	0	0	0	0	0.038

D: p.value de prueba post-hoc Dunn en alineamiento de regiones con términos de ontología

	C-O_60	C-O_60_50	C-O_60_75	C-O_80	C-O_80_50	C-O_80_75	M-D_60	M-D_60_50	M-D_60_75	M-D_80	M-D_80_50	M-D_80_75	M-L_60	M-L_60_50	M-L_60_75	M-L_80	M-L_80_50
C-O_60_50	0.027																
C-O_60_75	0	0.008															
C-O_80	0	0	0														
C-O_80_50	0	0	0	0.018													
C-O_80_75	0	0	0	0	0.014												
M-D_60	0.005	0.578	0.036	0	0	0											
M-D_60_50	0	0.005	0.918	0	0	0	0.027										
M-D_60_75	0	0	0.002	0.001	0	0	0	0.002									
M-D_80	0	0	0	**0.226**	**0.26**	0	0	0	0								
M-D_80_50	0	0	0	0	**0.26**	**0.198**	0	0	0	0.022							
M-D_80_75	0	0	0	0	0	**0.245**	0	0	0	0	0.013						
M-L_60	0.018	0.891	0.013	0	0	0	0.69	0.009	0	0	0	0					
M-L_60_50	0	0.02	0.754	0	0	0	0.08	0.668	0	0	0	0	0.029				
M-L_60_75	0	0	0.011	0	0	0	0	0.015	0.568	0	0	0	0	0.004			
M-L_80	0	0	0	**0.254**	**0.234**	0	0	0	0	**0.943**	0.018	0	0	0	0		
M-L_80_50	0	0	0	0.001	**0.299**	**0.167**	0	0	0	0.028	**0.935**	0.01	0	0	0	0.023	
M-L_80_75	0	0	0	0	0	**0.267**	0	0	0	0	0.015	**0.947**	0	0	0	0	0.012

10. Índices

10.1. Índice de Figuras

10.2. Índice de Tablas

10.3. Índice de Ecuaciones

10.4. Abreviaturas

Abreviatura	Descripción
CB	Beneficio clínico
CI	Intervalo de confianza
CGC	Censo de genes del cáncer
COSMIC	Catálogo de mutaciones somáticas en cáncer
FO-panel	Panel de genes de Foundation One
ICGC-DCC	Consorcio internacional del genoma del cáncer
MSA	alineamientos múltiples de secuencia
OS	Supervivencia total
PFS	Supervivencia libre de progresión
RSE	Raíz cuadrada de la varianza estimada del error aleatorio
R^2	Coeficiente de determinación es la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes
TCGA	Atlas del genoma del cáncer
TMB	Carga mutacional total. Por sus siglas del inglés Total Mutation Burden.
WES	Secuenciación del exoma completo
WGS	Secuenciación del genoma completo

10.5. Glosario de proteínas

El nombre del gen se toma de la notación HUGO. Las proteínas obtenidas de un gen toman obtienen el nombre del gen.

Nombre gene	Descripción de proteína
ABL1	ABL proto-oncogene 1, non-receptor tyrosine kinase
ACE2	angiotensin I converting enzyme 2
ALK	ALK receptor tyrosine kinase
APC	APC regulator of WNT signaling pathway
ATM	Serine-protein kinase ATM
AXIN1	axin 1
BCORL1	BCL6 corepressor like 1
BMPR1A	Bone morphogenetic protein receptor type-1A
BRAF	B-Raf proto-oncogene, serine/threonine kinase
BRCA1	BRCA1 DNA repair associated
BRCA2	BRCA2 DNA repair associated
CACNG3	calcium voltage-gated channel auxiliary subunit gamma 3
CALR	calreticulin
CDK4	Cyclin-dependent kinase 4
CDKN2A	cyclin dependent kinase inhibitor 2A
CTLA4	cytotoxic T-lymphocyte associated protein 4
CTNNB1	catenin beta 1
DCC	DCC netrin 1 receptor
DNMT3A	DNA methyltransferase 3 alpha
DST	dystonin
EGFR	epidermal growth factor receptor
ERBB2	erb-b2 receptor tyrosine kinase 2
ESR1	estrogen receptor 1
FAM135B	family with sequence similarity 135 member B
FBXW7	F-box and WD repeat domain containing 7
FLG	filaggrin
FLT3	fms related tyrosine kinase 3
HLA-A	major histocompatibility complex, class I, A
IDH1	isocitrate dehydrogenase (NADP(+)) 1
JAK2	Janus kinase 2
KCNJ5	potassium inwardly rectifying channel subfamily J member 5
KEAP1	kelch like ECH associated protein 1
KIT	KIT proto-oncogene, receptor tyrosine kinase

KMT2C	lysine methyltransferase 2C
KRAS	KRAS proto-oncogene, GTPase
MACF1	microtubule actin crosslinking factor 1
MAP2K1	mitogen-activated protein kinase kinase 1
MAP2K2	mitogen-activated protein kinase kinase 2
MDM2	MDM2 proto-oncogene
MDM4	MDM4 regulator of p53
MET	MET proto-oncogene, receptor tyrosine kinase
MGA	MAX gene-associated protein
MLH1	DNA mismatch repair protein Mlh1
MSH2	DNA mismatch repair protein Msh2
MUC16	mucin 16, cell surface associated
MYC	Myc proto-oncogene protein
NCOR1	nuclear receptor corepressor 1
NEB	nebulin
NOTCH1	notch receptor 1
NOTCH2	notch receptor 2
NPM1	nucleophosmin 1
NRAS	NRAS proto-oncogene, GTPase
OBSCN	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF
PD-1 / PDCD1	programmed cell death 1
PD-L1 / CD274	CD274 molecule
PDE4DIP	phosphodiesterase 4D interacting protein
PDGFRA	platelet derived growth factor receptor alpha
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PTEN	phosphatase and tensin homolog
PTPRT	protein tyrosine phosphatase receptor type T
RB1	Retinoblastoma-associated protein
SETBP1	SET binding protein 1
SMO	smoothened, frizzled class receptor
STK11	Serine/threonine-protein kinase STK11
SYNE1	spectrin repeat containing nuclear envelope protein 1
TP53	tumor protein p53
TTN	titin
VHL	von Hippel-Lindau disease tumor suppressor