UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud

INGENIERÍA DE LA SALUD

TFG del Grado en Ingeniería de la Salud

Análisis bioinformático para la detección de biomarcadores del cáncer empleando Biopython

Presentado por Gabriel Collado Santamaría
en Universidad de Burgos

8 de julio de 2024

Tutores: Rubén Ruiz González – Antonia Maiara Marques do Nascimento

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería de la Salud

D. Rubén Ruiz González, profesor del departamento de Digitalización, área de Ingeniería de Sistemas y Automática.
Dña. Antonia Maiara Marques do Nascimento, profesora del departamento de Biotecnología y Ciencia de los Alimentos, área de Bioquímica y Biología Molecular.

Exponen:

Que el alumno D. Gabriel Collado Santamaría, con DNI 71482493B, ha realizado el Trabajo final de Grado en Ingeniería de la Salud titulado "Análisis bioinformático para la detección de biomarcadores del cáncer empleando Biopython".

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección de quienes suscriben, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 8 de julio de 2024

Vº. Bº. del Tutor:

D. Rubén Ruiz González

Vº. Bº. de la Tutora:

Dña. Antonia Maiara Marques do Nascimento

Resumen

El cáncer de mama es un problema socio-sanitario de primer nivel en todo el del mundo. Se trata del cáncer con mayor incidencia en mujeres, llevándose a su vez la vida de 685.000 personas en el mundo y 6.608 personas tan solo en España para el año 2022.

Ante esta evidente necesidad, y debido al compromiso de todo profesional sanitario, en este proyecto se pretendió diseñar un esquema de procesado de datos genéticos basado en la detección de biomarcadores característicos en el cáncer de mama.

A través de la base de datos de *NCBI* y la aplicación del paquete de *Biopython*, se pudieron realizar sobre un *notebook* de *Jupyter* diferentes alineamientos pareados y múltiples, entre patológicos y controles empleando los algoritmos `Muscle` y `Clustal`.

Gracias a dicho procesado bioinformático, y a pesar de las diversas dificultades experimentadas respecto a las condiciones presentadas por los datos, tuvo lugar al reconocimiento de mutaciones distintivas en aquellos genes vinculados con la patología. Entre ellas se destacan las sustituciones de un simple nucleótido; T>A para *BRCA1*, A>T para *BRCA2*, C>T para el exón 5 de *TP53* y G>A para el exón 8 del mismo.

Los resultados han podido demostrar la gran aplicabilidad del paquete de *Biopython* como herramienta bioinformática para la detección de biomarcadores en el cáncer, una vez que se ha podido detectar mutaciones para los genes (*BRCA1*,*BRCA2*,*PIK3CA* y *TP53*) en células caracterizadas anteriormente como patológicas o normales para el cáncer de mama. Además, este trabajo, sienta las bases para futuros proyectos bioinformáticos de interés, y determina la importancia de la obtención de unos datos de calidad en etapas iniciales.

Entre las posibles mejoras se destacaría la elaboración de un análisis adicional de las regiones intrónicas, dada su fuerte relación con aquellas regiones codificantes, y la complementariedad del proyecto desarrollado con diversas herramientas bioinformáticas alternativas (AlphaFold, DeepVariant y DNABert).

Descriptores

Biopython, alineamiento, `Clustal`, `Muscle`, *NCBI*, secuencias ADN, *TP53*, *BRAC1*, *BRCA2*, *PIK3CA*, biomarcadores, mutaciones, detección.

Abstract

Breast cancer is a major socio-health problem throughout the world. It is the cancer with the highest incidence in women, taking the lives of 685,000 people in the world and 6,608 people in Spain alone by 2022.

Given this evident need, and due to the commitment of every health professional, this project sought to design a genetic data processing scheme based on the detection of characteristic biomarkers in breast cancer.

Through the *NCBI* database and the application of the *Biopython* package, different paired and multiple alignments, between pathological and controls using the `Muscle` and `Clustal` algorithms.

Thanks to said bioinformatics processing, and despite the various difficulties experienced regarding the conditions presented by the data, distinctive mutations were recognized in those genes linked to the pathology. Among them, single nucleotide substitutions stand out; T>A for *BRCA1*, A>T for *BRCA2*, C>T for exon 5 of *TP53* and G>A for exon 8 of the same.

The results have been able to demonstrate the great applicability of the *Biopython* package as a bioinformatics tool for the detection of biomarkers in cancer, once it has been possible to detect mutations for the genes (*BRCA1*,*BRCA2* ,*PIK3CA* and *TP53*) in cells previously characterized as pathological or normal for breast cancer. Furthermore, this work lays the foundations for future bioinformatics projects of interest, and determines the importance of obtaining quality data in the initial stages.

Among the possible improvements, the development of an additional analysis of the intronic regions would be highlighted, given their strong relationship with those coding regions, and the complementarity of the project developed with various alternative bioinformatics tools (AlphaFold, DeepVariant and DNABert).

Keywords

Biopython, alignment, `Clustal`, `Muscle`, *NCBI*, DNA sequences, *TP53*, *BRAC1*, *BRCA2*, *PIK3CA*, biomarkers, mutations, detection.

Índice general

Índice de figuras

Índice de tablas

Introducción

El cuerpo humano está compuesto por billones de células que tienen un tiempo de vida definido. Cuando una célula completa su ciclo vital o presenta una naturaleza anormal, es degradada y es reemplazada por una nueva célula. Este proceso de equilibrio es fundamental para el funcionamiento saludable del organismo siendo realizado por reguladores positivos y negativos como las proteínas. En el caso de las proteínas estimuladoras del ciclo celular se encuentran las proteínas quinasas dependientes de ciclinas, también conocidas como CDK (del inglés cyclin-dependent kinases) y las ciclinas, mientras que la inhibición proteica es llevaba a cabo por las inhibidoras de CDK [López Marure, 2013].

En concreto se dan dos tipos de genes que regulan la conocida regulación celular, los genes supresores de tumores y los protooncogenes. Estos primeros son aquellos que ralentizan la división de la célula o lo detienen para poder corregir el posible error producido en ellas, indicándolas el momento en el que tienen que morir [American Cancer Society, 2022c]. Y, por otro lado, se encuentran los ya nombrados protooncogenes, aquellos encargados de la estimulación o protección de las células frente a la muerte celular, ayudándolas de esta manera a mantenerlas vivas mediante diferentes mecanismos moleculares [Bertram, 2000].

En ocasiones, algunas células pueden experimentar alteraciones, momento en el que actúan los genes reparadores del ADN, encargados de solventar dicha problemática, y en caso de no conseguirlo, desencadenar la muerte celular [American Cancer Society, 2022c]. El problema se encuentra en que si a su vez esas alteraciones se presentan en alguno de los genes nombrados anteriormente, estos pueden ser inactivados perdiendo con ello su capacidad degradativa, o activados en exceso y convirtiéndose ,de esta forma, en oncogenes; dando lugar a una proliferación descontrolada de células

anormales, conocida como cáncer [American Cancer Society, 2022c]. Esta masa celular hiperplásica, con diversas causas y manifestaciones, comienza concentrándose en la región donde ocurre la alteración, y a medida que crece, puede producirse un movimiento migratorio, a través del torrente sanguíneo, del carcinoma a otras regiones del organismo, y producirse la conocida metástasis. Cuando las células cancerosas superan en número a las células normales, pueden interferir con el funcionamiento adecuado del cuerpo [American Cancer Society, 2020].

Trágicamente, el cáncer se ha convertido en la principal causa de muerte a nivel mundial, llegando en 2020, a ser responsable de una de cada seis defunciones registradas [World Health Organization, 2022] y una de cada cuatro por enfermedades no transmisibles [Bray et al., 2024]. Siguiendo con una vista globalizada, en el 2022 hubo casi 20 millones de nuevos diagnósticos, y un total de 9,7 millones de defunciones relacionadas con el cáncer. Pero el pronóstico todavía es peor a futuro, esperándose en 2040 aumentar hasta un 50 % más los nuevos diagnósticos, con 29,9 millones de personas afectadas, y un gran aumento de las muertes alcanzando los 15 millones [NIH, 2024c].

Según la sociedad estadounidense estima que, para este año de 2024, solo en Estados Unidos tengan lugar un total de 2.001.140 nuevos casos, alrededor de un millón para hombres y para mujeres, como podemos ver en la Figura 1.1 mostrada a continuación. Y de entre ellos, llegan a perder la vida más de medio millón de personas [Siegel et al., 2024].

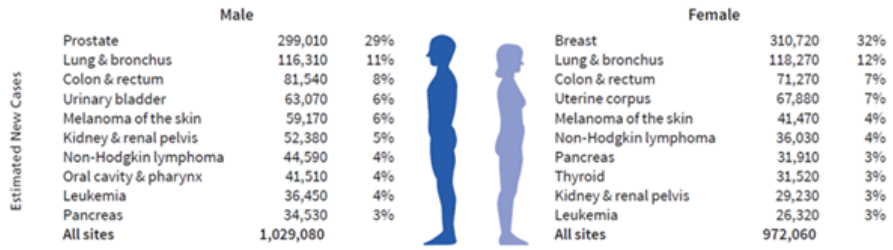

Figura 1.1: Número estimado de nuevos cánceres en 2024 para hombres y mujeres. Fuente: [Siegel et al., 2024].

Entre los numerosos posibles tipos de cáncer, destacamos el cáncer de mama, el cuál es un problema socio-sanitario de primer nivel, tanto en España, como en el resto del mundo (Figura 1.2). Este se caracteriza por su complejidad genética, elevada incidencia sobre la población (tal y como se podía apreciar en la Figura 1.1) e importancia de la detección temprana para la mejora de tasas de supervivencia entre los pacientes [Martín et al., 2015].

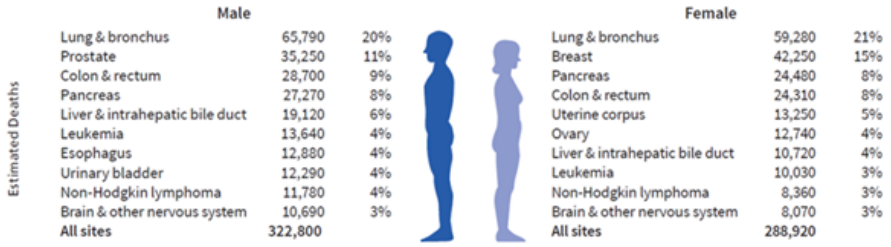

Figura 1.2: Número estimado de muertes por cáncer en 2024 para hombres y mujeres. Fuente: [Siegel et al., 2024].

Cabe destacar que el cáncer de mama llega a presentar tres subtipos principales, según la expresión, es decir, presencia o ausencia de: receptores hormonales como lo son el de estrógeno y progesterona, y el factor de crecimiento epidérmico humano 2 (ERBB2, erb-b2 receptor tyrosine kinase 2), que anteriormente era nombrado como HER2(hormone epidermic repector 2) el cual fomenta el crecimiento rápido de células cancerosas [American Cancer Society, 2022a]. La frecuencia de estos es variable, siendo del 70 % para un diagnóstico de receptor hormonal positivo/ERBB2 negativo (cualquiera de los dos), 15-20 % para un ERBB2 o HER2 positivo, un 15 % para el caso triple negativo, que tiene lugar cuando se carece de los tres marcadores moleculares nombrados, y, por último, aunque sea menos probable, encontramos el caso del triple positivo, la situación diagnostica en la que todos nuestros marcadores moleculares de referencia se encuentran presentes [Waks and Winer, 2019]. Es de gran importancia la clasificación del subtipo de cáncer presente en el paciente, debido a que la elección del tratamiento y su efectividad, cambiará en función de cual se padezca.

Entre casos positivos de receptor hormonal, el receptor hormonal de estrógeno es el principal factor de transcripción que impulsa en gran medida el desarrollo de tumores de mama; y sirve a su vez, de evaluador ante una posible terapia antiestrogénica. Se trata a su vez del subtipo de cáncer de mama con más recaídas tardías, provocando un seguimiento más exhaustivo de los pacientes afectados durante un tiempo más prolongado, lo que ha sido y es un reto clínico significativo [Lim et al., 2012].

Un diagnóstico de ERBB2 positivo se encuentra asociado a un mal pronóstico, además de ser un marcador que permite la magnitud de agresividad de un tumor por su más que justificada relación con la escasez de receptores esteroideos que conduce a peores grados histológicos del tumor, aneuploidía y alta tasa de proliferación entre otros [Révillion et al., 1998].

El cáncer de mama triple negativo es un subtipo de cáncer diagnosticado habitualmente en personas de menor edad y con menor tasa de supervivencia por recaída en comparación con el resto, de subtipos. Biológicamente heterogéneo, implica una gran diversidad de comportamientos clínicos, que, sumado a una falta de tratamiento dirigido, consigue el peor pronóstico existente posible [Zagami and Carey, 2022].

El carcinoma de mama es una patología dependiente del sexo, excepcional para los hombres, suponiendo alrededor de un 1 % de los casos detectados, mientras que habitual para las mujeres, con 35.000 diagnósticos en el 2023; un número de casos que se espera que aumente rondando cifras de 36.300 casos para este nuevo año según la Sociedad Española de Oncología Médica. Es decir, la incidencia ha ido en aumento con un comportamiento ascendente constante a lo largo de los años, tal y como se muestra en la Figura 1.3 y Figura 1.4 ([AECC, 2023]).

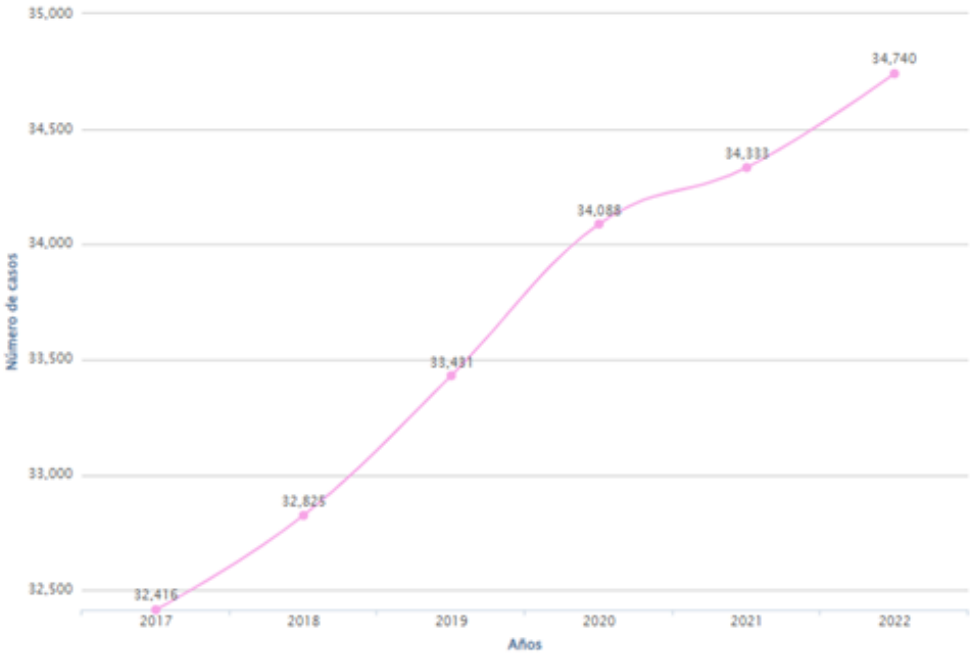

Figura 1.3: Evolución de la incidencia del cáncer de mama en España a lo largo de los años. Fuente: [AECC, 2023]

A nivel mundial, para el 2023, la incidencia del carcinoma de mama junto al de pulmón y cáncer colorrectal conformaban el 52 % de los diagnósticos. La cifra todavía es más sorprendente cuando el cáncer de mama por sí solo suponía el 31 % de los canceres en mujeres [Siegel et al., 2023]. Además, en

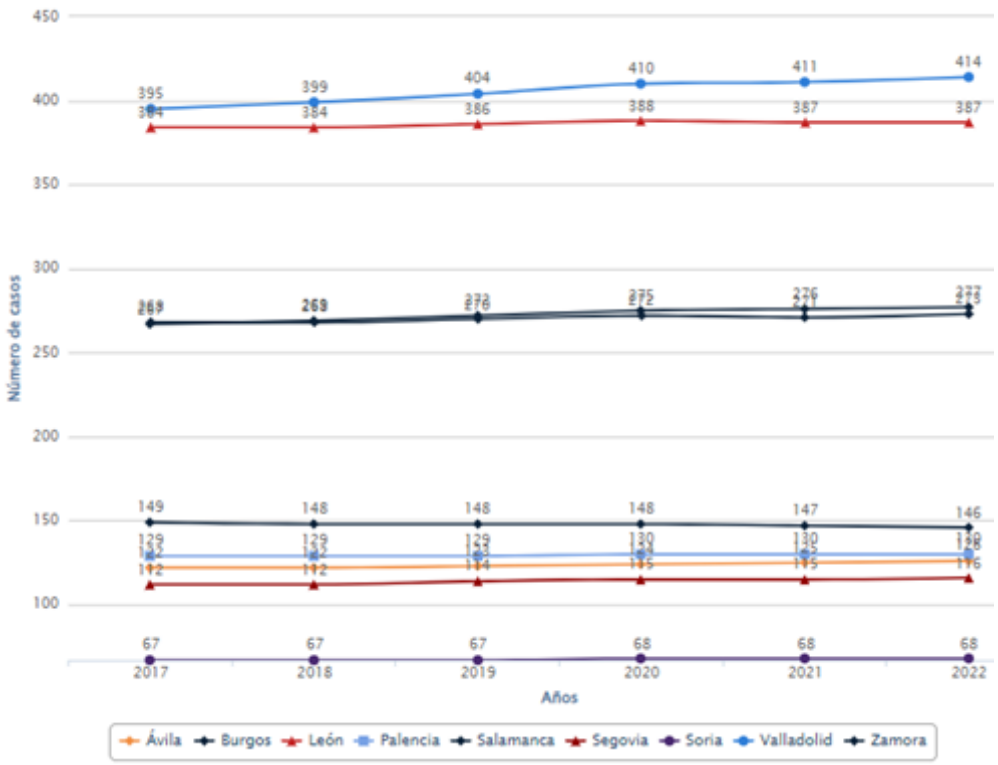

Figura 1.4: Evolución de la incidencia del cáncer de mama en Castilla y León a lo largo de los años. Fuente: [AECC, 2023].

2022 supuso el 11,6 % de todos los casos diagnosticados de cáncer para ese mismo año, encontrándose en la segunda posición de cánceres con mayor incidencia, por detrás del cáncer de pulmón [Bray et al., 2024].

Según la AECC (Sociedad Española Contra el Cáncer) y como se muestra mediante una escala de colores con mayor o menor intensidad en el siguiente mapa gráfico (Figura 1.6), entre las comunidades autónomas dentro de territorio español con más incidencia se encuentran Asturias, Galicia y Castilla y León; mientras que aquellas con menor incidencia encontramos Murcia, Baleares y Andalucía.

La presencia de la enfermedad viene también determinada por la edad. Los estudios realizados han demostrado una mayor predisposición por un curso terapéutico peor en personas jóvenes que en aquellas con cierta edad. Cabe añadir que la incidencia es más frecuente en mujeres mayores de 50 años, siendo solamente entre un 5 % y un 12 % casos en mujeres con menos de 45 años. Esta característica puede orientar a los profesionales sanitarios a

Figura 1.5: Evolución de la incidencia del cáncer de mama en la provincia de Burgos a lo largo de los años. Fuente: [AECC, 2023].

concretar la etiología del tumor, siendo un ejemplo, la aparición temprana de la afección uno de los sellos distintivos de la existencia de un factor genético predisponente en el usuario.

El riego es doble para una persona diagnosticada con cáncer de mama, debido a que dicha condición incentiva la aparición de un segundo carcinoma como el ovárico, cuya probabilidad varía según la edad de esta, siendo del 29 % para personas con menos de 50 años, y del 44 % hasta los 70 [Mehrgou and Akouchekian, 2016].

El cáncer de mama se considera uno de los cánceres con mejor pronóstico, presentando una tasa de supervivencia del 86 %, que, a pesar de ello, y contrario a la idea que puede presentar, provocó la muerte de más de medio millón de personas por todo el globo [Bray et al., 2024], situándola como uno de los carcinomas con mayor mortalidad en los últimos años, y la primera causa de muerte por cáncer entre las mujeres en España con un total de 6.528 fallecidos por año. Las mayores tasas de mortalidad se encuentran distribuidas entre las comunidades autónomas de Asturias, Castilla y León y Aragón, con una tasa y mortalidad media de 18.7 y 290.03 correspondientemente [AECC, 2023].

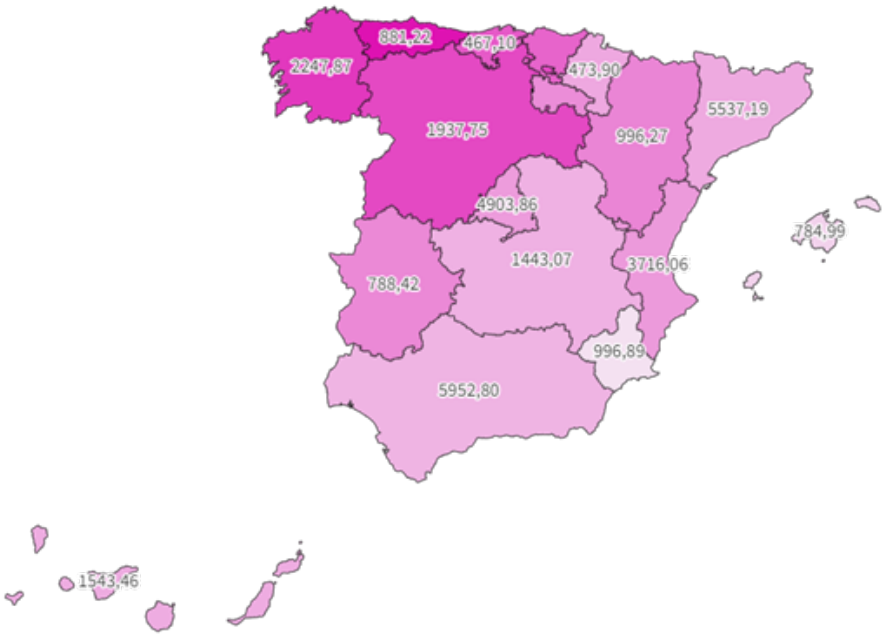

Figura 1.6: Tasa de incidencia por comunidad autónoma en España por cada 100.000 habitantes. Fuente: [AECC, 2023].

Por otro lado, muchos casos podrían llegar a prevenirse teniendo en cuenta factores de riesgo en nuestro estilo de vida que son evitables, como puede ser el tabaquismo, alcoholismo y la obesidad; siendo estos dos últimos especialmente determinantes para nuestro carcinoma a estudio, como podemos ver en las Figuras 1.8 y 1.9. Además de poder llegar a realizar ciertos cursos de prevención y autoexploración, resaltando la importancia del tiempo, debido a lo decisivo que este supone para este tipo de carcinoma [Sociedad Española de Oncología Médica, 2023].

Desde una perspectiva genética, el desarrollo potencial del cáncer de mama puede estar asociado a una variedad de mutaciones en numerosos genes. Entre ellos, es crucial destacar la relevancia de *BRCA1* (breast cancer gene 1) [NIH, 2024a] y *BRCA2* (breast cancer gene 2) [NIH, 2024b], uno de los descubrimientos más importantes del campo de la genética en el cáncer humano, cuya conexión se ha demostrado ampliamente en muchos estudios, sobre todo en cánceres de mama hereditarios, presentando mutaciones para el 30 % del total de los casos [Margarit, 2008]. Ambos genes tienen un papel fundamental en la reparación por recombinación homóloga (HR, homologous recombination), y con su ausencia o si presentan mutaciones (como se ha comentado anteriormente, principalmente en cánceres hereditarios), quedan

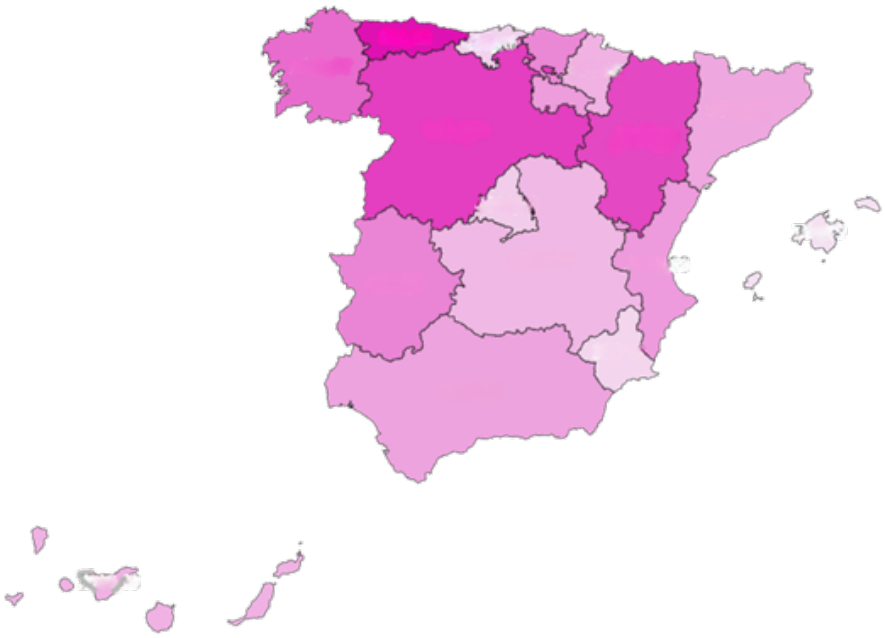

Figura 1.7: Tasa de mortalidad por comunidad autónoma en España por cada 100.000 habitantes. Fuente: [AECC, 2023].

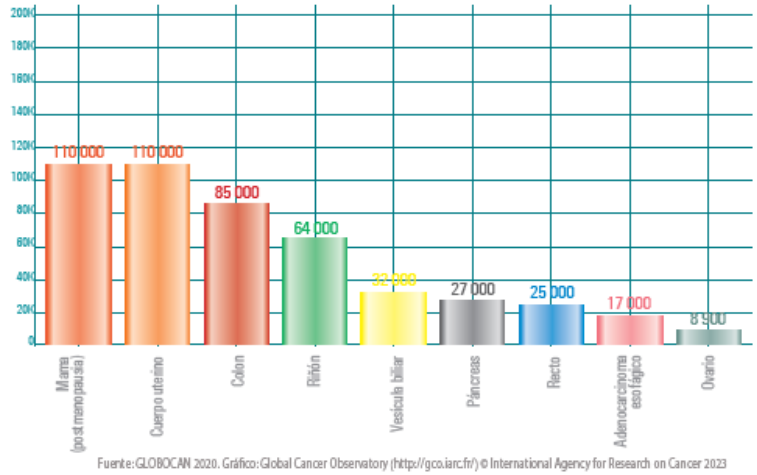

Figura 1.8: Número de casos de cáncer atribuibles al exceso de IMC a nivel global para el año 2012. Fuente: [Sociedad Española de Oncología Médica, 2023].

inutilizables, provocando la inhibición de este mecanismo de reparación, y

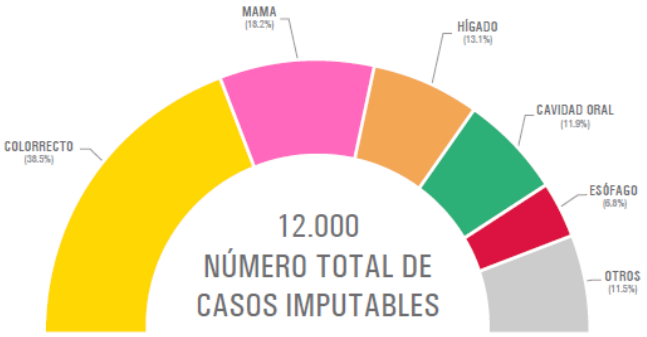

Figura 1.9: Número estimado de nuevos casos de cáncer atribuibles al consumo de alcohol en España en 2020. Fuente: [Sociedad Española de Oncología Médica, 2023].

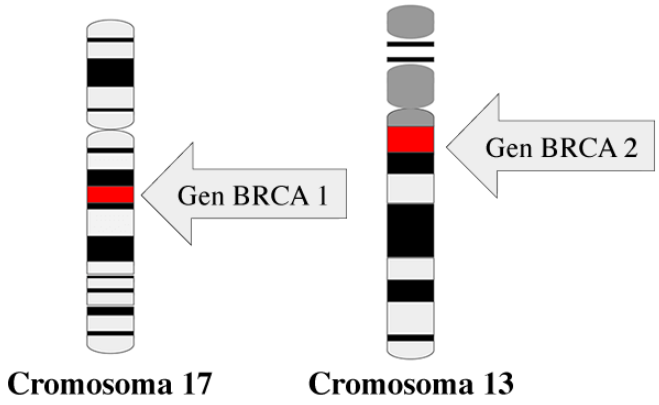

Figura 1.10: Representación visual de los genes *BRCA1* y *BRCA2* en el cromosoma. Fuente: [NCBI, 2024b].

con ello, la reparación del ADN por mecanismos secundarios o alternativos más propensos a errores [Foulkes and Shuen, 2013].

Los cortes de doble cadena son uno de los daños más graves que pueden tener lugar en la cadena de ADN, produciendo la inestabilidad genómica existente por posibles translocaciones y pérdida de material genético. Para realizar su corrección existen dos mecanismos principales a los que se suele

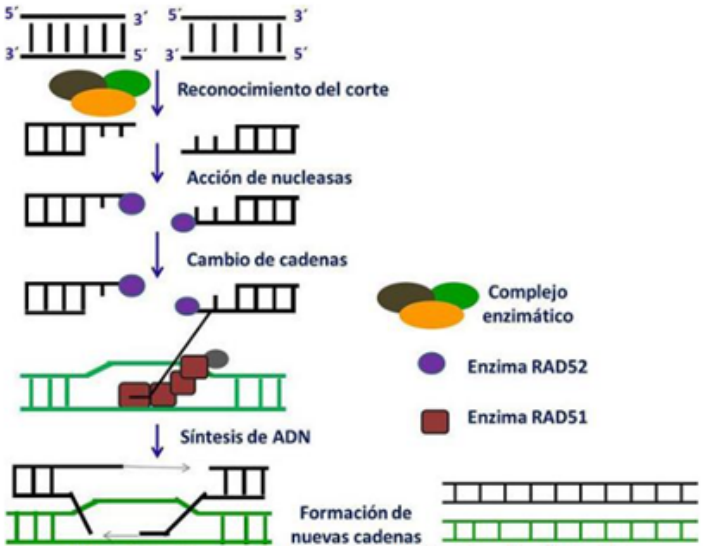

Figura 1.11: Representación gráfica del funcionamiento del sistema de reparación vinculado a *BRCA1* y *BRCA2*, la recombinación homóloga. Fuente: [Yaliana Tafurt Cardona and Maria Aparecida Marin Morales, 2014].

recurrir: la recombinación de extremos no homólogos y la recombinación homóloga (Figura 1.11).

La recombinación homóloga, es un sistema que detecta y repara los daños producidos en ambas cadenas de ADN para aquellas células que se encuentren en la fase S y G2 del ciclo celular, mediante varias enzimas, un complejo enzimático y ambas cadenas desoxirribonucleicas cortadas. En primer lugar el complejo proteico MRN (MRNIP, MRN complex interacting protein) (conformado por RAD50, NBS1 y MRE11) reconoce y degrada las dobles cadenas en el punto donde se produjo el corte, posteriormente, la enzima RAD52 actúa como defensa frente a las posibles exonucleasas inespecíficas que pueda llegar a haber, y la enzima RAD51, se encarga con ayuda de RAD52 y en presencia de ATP, de la síntesis de una cadena nucleoprotéica basándose en la cadena homóloga del cromosoma homólogo que no sufrió el corte, es decir, su cromátida hermana. Ya con las cadenas entrecruzadas tiene lugar la estructura llamada “Cruz de Holiday” con la que se finalizará el proceso de reparación [Tafurt Cardona and Marin Morales, 2014].

Relacionado con la regulación del ciclo celular, estabilidad del genoma y otros procesos fisiológicos de gran importancia, mutaciones en el gen *BRCA1* representan un riesgo importante para el desarrollo oncológico de otros cánceres [Fu et al., 2022].

BRCA1, un gen recesivo incompleto en un autosoma y compuesto por un total de 24 exones y un total de 2843 aminoácidos, tiene como producto génico la Ubiquitina ligasa E3, una proteína encargada entre múltiples funciones de mediar específicamente la formación de cadenas de poliubiquitina unidas a Lisina-6 y desempeñar un papel principal en la reparación por recombinación homóloga del ADN, facilitando de esta forma, las respuestas celulares al daño del ADN [UCSC, 2024a].

En 1995 se identificó un gen aún más grande que *BRCA1*(con 3418 aminoácidos), *BRCA2*. *BRCA2* se une a RAD51 y potencia la reparación recombinacional del ADN, al igual que *BRCA1*, al promover el ensamblaje de RAD51 en ADN monocatenario (ssADN) [UCSC, 2024b]. También este realiza un importante papel en la recombinación meiótica con las interacciones que llega a establecer con la recombinasa DMC1 (ADN Meiotic Recombinase 1) [HGNC, 2007].

Por consiguiente, la identificación de familias con alto riesgo para el cáncer de mama es esencial para facilitar una derivación oportuna hacia asesoramiento genético, basado en una evaluación del riesgo, lo que a su vez favorece la detección temprana.

Sin embargo, es importante tener en cuenta que el cáncer de mama puede derivarse de mutaciones en otros genes además de *BRCA1* y *BRCA2*, como son el caso de *PIK3CA*(Tumor Protein p53) y *PIK3CA* (phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha), como podemos observar en la Figura 1.12, que son comunes y pueden ocurrir de manera habitual en ausencia de antecedentes familiares identificables, como puede ser la acumulación de mutaciones adquiridas a lo largo de la vida.

PIK3CA es el gen que codifica una de las proteínas supresoras de tumores por excelencia, encargándose en gran medida del control de la división y destrucción de las células del organismo. Un gen relacionado con múltiples funciones moleculares como puede ser la unión de la cromatina y regulación de la transcripción actuando como un factor determinante; y por otro lado, numerosos procesos a nivel biológico, como la reparación de bases y renaturalización de la cadena de ADN, influyendo además en el comportamiento circadiano habitual [NIH, 2024d].

Suele ser diana de muchas mutaciones (la mayoría adquiridas, desarrollando el cambio en algún momento dado después del nacimiento), conteniéndolas aproximadamente en el 50-60 % de cánceres humanos, en el que, en 90 de cada 100 ocasiones, se codifican proteínas mutantes sin sentido; siendo estas las más comunes en los cánceres. Todo ello produce la incapacidad o

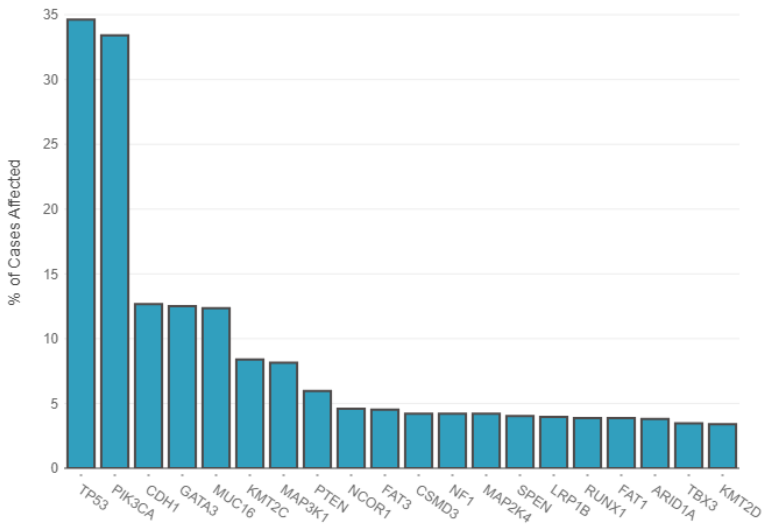

Figura 1.12: Gráfica de *TCGA* que presenta los genes ordenados por frecuencia de mutación. Fuente: [NCI and National Human Genome Research Institute, 2006].

reducción de la capacidad natural de unión a aquella secuencia de ADN relacionada con la regulación de este conocido gen.

Se ha desarrollado incluso, una técnica diagnóstica destinada concretamente a su evaluación dada la importancia que tiene, y la motivación por la obtención de una terapia más personalizada a los pacientes, la prueba genética *PIK3CA*. Una prueba destinada a pacientes diagnosticados de un carcinoma o a quienes guarden relación con un pariente que lo haya padecido. Se basa en un análisis de sangre, y la recogida de tejido tumoral mediante una biopsia sólida o una muestra de la parte posterior del hueso de la cadera, para una recogida de la médula ósea [MedlinePlus, 2024].

Las mutaciones de *PIK3CA* también son comunes en el cáncer de mama, estando presentes entre el 20 % al 40 % de los casos de este tipo de carcinoma, dependiendo además su presencia del tamaño que llegue a conseguir el tumor y cómo de desarrollado se encuentre, es decir, su estadio [Børresen-Dale, 2003].

Por otro lado, se encuentra *PIK3CA*, gen que codifica la subunidad catalítica p110 alpha de la fosfatidilinositol 3-quinasa, mejor conocida como "PI3K", que, gracias a su papel realizado en la vía PI3K/Akt, realiza

funciones proliferativas, metabólicas, angiogénica y apoptóticas entre otras [UCSC, 2024c].

Numerosos estudios afirman la presencia de un *PIK3CA* mutado en diversos canceres, con especial frecuencia en el cáncer de mama y endometrio, tratándose por ello un oncogén a considerar en numerosas ocasiones (Figura 1.13). Sus mutaciones suelen concentrarse en ciertos dominios específicos como los son el helicoidal, ubicado en el exón 9, y quinasa, situado en el exón número 20. El receptor inhibidor (*NVP-BYL719*) de este, por tanto, también es de interés, desarrollándose mediante él como principal diana, terapias anticancerígenas prometedoras. Terapias incluso, más allá de su inhibición, utilizando la propia enzima para la cual sintetiza su subunidad, aprovechando la capacidad regulatoria de autofagia que presenta y sus propiedades pro y anticancerígenas. De forma que "PI3K" se ha convertido en un área de investigación de especial intensidad, llegando a superar las 36.000 publicaciones en PubMed y guardar relación con las publicaciones que se cuelgan a diario [Arafeh and Samuels, 2019].

Atendiendo a los datos contenidos en las diferentes bases de datos, estos dos genes (*PIK3CA* y *PIK3CA*) presentan una importancia crucial en la investigación y diagnóstico del cáncer de mama, debido a que son aquellos genes con mayor prevalencia en esta patología [Xiao-Yi Lin et al., 2023].

Numerosos estudios han comprobado la relación que guardan ambos genes y el cáncer de mama, concretamente entre las mutaciones que pueden llegar a presentarse en ambos y una supervivencia libre de enfermedad más corta. Uno de ellos, realizado por el departamento de cirugía mamaria del Hospital Popular Provincial de Guangdong, afirma una clínica desfavorable y peor pronóstico en la situación de una co-mutación de los genes. La co-mutación disminuye la sensibilidad de ciertas terapias neoadyuvantes, especialmente la conocida quimioterapia con taxanos, en comparación con sus mutaciones individuales. Además de presentar una mayor asociación con tumores de detección tardía, HR-negativos y de alto grado. Por todo ello, determinaron y solicitaron la urgencia de desarrollar nuevas estrategias terapéuticas efectivas para este tipo de co-mutaciones.

Existen diversos métodos que sirven como prueba diagnóstica para hallar un posible carcinoma, desde la exploración física del profesional sanitario, pasando por las conocidas mamografías, IRMs (Espectrometría de masas de relaciones isotópicas de gases) o una técnica tan habitual como lo es una biopsia o ganglios centinelas, con su posterior análisis en el laboratorio. Decenas de pruebas pueden ser de gran utilidad para la detección de un carcinoma, aunque dependiendo de la especulación del tipo de cáncer que

Tipo de cáncer	Frecuencia de mutación (%)
endometrio	24-46%
Mama	20-32%
Vejiga	20-27%
Cervical	14-23%
colorrectal	13-28%
Cabeza y cuello	12-15%
ampular	9-15%
Esófago/estómago	2-18%
pulmón escamoso	5-10%
Glioblastoma	2-8%
Melanoma	2-5%
Pancreático	2-3%
Riñón	1-5%
Tejido blando/Sarcoma	1-3%
Hígado	1-3%
testicular	2%
Hueso	1,87%

Figura 1.13: Mutaciones de *PIK3CA* en cáncer. Fuente: [Arafeh and Samuels, 2019].

se tenga, se exponen ciertas técnicas con una eficacia y efectividad más favorables que con el resto de ellas, lo que provoca una mayor predilección en estas por parte de los profesionales sanitarios ocasionando su uso frecuente.

Un claro ejemplo sería la mamografía, la prueba diagnóstica común por excelencia del cáncer de mama. La mamografía presenta una imagen del interior de la mama del paciente mediante radiografía, permitiendo encontrar tumores de pequeño tamaño al tacto y más definido como el carcinoma ductal in situ. En caso de detección exitosa de un carcinoma, el tiempo, la dosis de radiación y las tomas en diferentes ángulos aumentarán, para aclarar el diagnóstico preciso del paciente [NIH, 2023b].

A su vez, para una exploración y análisis más exhaustivo, aunque se presente una mayor probabilidad de falso positivo en este tipo de prueba, para casos con especial riesgo se podría recurrir a la resonancia magnética con la adición del químico gadolinio en el torrente sanguíneo que ayudará a mostrar cualquier tejido anormal de la mama [American Cancer Society, 2022].

También dentro del cáncer de mama, una de las pruebas habituales en su diagnóstico es el análisis genómico mediante una muestra de sangre de los genes de *BRCA1* y *BRCA2* (pudiendo incluirse también los genes de *PIK3CA* y *PIK3CA*), aquellos los cuales, como se ha nombrado con anterioridad presentan un mayor riesgo para un desarrollo de cáncer hereditario que TP35 y *PIK3CA*.Prueba la cual es recomendada realizar en caso de que un paciente presente relación con un familiar cercano, ya sea masculino o femenino, con cáncer de mama y ovario; e incluso en caso de que el paciente sea de ascendencia judía procedente de Europa Oriental [MedlinePlus, 2023].

Pruebas diagnósticas específicas para genes también tienen cabida, como es el ejemplo de *PIK3CA* y la prueba diagnóstica desarrollada por la empresa farmacéutica multinacional Roche. La prueba de mutación cobas® *PIK3CA* (Roche), es una técnica de diagnóstico de alta sensibilidad basada en una reacción en cadena de la polimerasa (PCR) en tiempo real destinada al hallazgo e identificación de mutaciones en los exones 2,5,8,10 y 21 del gen *PIK3CA* en pacientes con un cáncer de mama avanzado o metastásico [Roche, 2024].

Aquellos pacientes que fueran diagnosticados con cánceres invasivo de mama también deberían serles analizados el estado de la proteína HER2 con la extracción de muestra mediante un biopsia o extracción del carcinoma a través de cirugía y evaluada por técnicas inmunohistoquímicas o de hibridación fluorescente in situ (FISH); proteína la cual ya se indicó de ser una de las causantes de la existencia de subtipos dentro de este tipo de cáncer [American Cancer Society, 2022b].

En primera instancia tiene lugar la prueba inmunohistoquímica, dado el costo y el tiempo que supone realizar la técnica diagnostica de FISH, siendo esta última utilizada como prueba complementaria ante un posible resultado de ambigüedad dentro de la primera. Para una prueba de IHC (Immunohistochemistry, IHC) los resultados obtenidos pueden ser de 0/HER2-negativo, +1/HER2-negativo, +2/HER2-ambiguo y +3/HER2-positivo; siendo todos ellos de gran importancia para un posterior tratamiento exitoso [American Cancer Society, 2022b].

Aunque evidentemente no todos los diagnósticos se asocian solo a un análisis de HER, sino que la presencia de receptores de estrógeno y progesterona a su vez presentan un papel importante [American Cancer Society, 2021].

Añadiendo la posibilidad de detectar hasta dos diagnósticos más el cáncer de mama triple negativo y triple positivo [American Cancer Society, 2022b].

Cabe recalcar que las pruebas diagnósticas no solo se quedan en las ya nombradas, sino que ya con una visión más a futuro, existen gran cantidad de ellas que se encuentran en desarrollo en ciertas fases avanzadas de ensayos clínicos, y destinados al diagnóstico de cáncer y su aplicación en mama como lo son el examen de mama, la termografía o muestreo de tejido entre muchas más [NIH, 2023a].

Como se mencionó anteriormente, el tratamiento al cual se someta un paciente depende en gran medida del estadio de la enfermedad en el que el paciente se encuentre y la clasificación del subtipo del cáncer. El objetivo de todos es paliar los síntomas y prolongar la vida de los pacientes todo lo posible, evitando las secuelas que podrían ocurrir en ellos. De forma general los tratamientos más conocidos relativos al cáncer se pueden resumir en tres, una cirugía destinada a la extirpación de la masa tumoral del cuerpo, el bombardeo de altas dosis de radiación para la destrucción de tejidos cancerosos con la radioterapia, y la conocida quimioterapia, con el tratamiento mediante diversos fármacos para la eliminación de las células de carácter oncológico o incluso el alivio de los síntomas que se pueden desarrollar debido a un cuadro cancerígeno [NIH, 2024].

Aplicables estos a cualquier tipo de cáncer, incluido el de mama. Aunque este, a su vez, presenta tratamientos dirigidos a paliarlo. El tratamiento para aquel carcinoma de mama no metastásico, tal y como se ha repetido a lo largo del documento, depende mucho del subtipo que tenga lugar; para un receptor hormonal positivo se recurre a un tratamiento endocrino y en casos más concretos una quimioterapia complementaria. Los casos de ERB2 o HER2 positivos reciben una terapia con anticuerpos dirigidos,

o, como alternativa inhibidores de moléculas pequeñas acompañadas de quimioterapia [Waks and Winer, 2019].

En el caso de triple positivo pueden darse tratamientos con medicamentos hormonales o que tienen como diana la proteína HER2 o ERB2. Un caso de triple negativo es más particular al no presentar estado positivo alguno en los biomarcadores de diagnóstico de referencia, resultado en una ineficacia más que demostrada en la mayoría de los tratamientos dirigidos a cáncer de mama y se recurriría directamente a quimioterapia [American Cancer Society, 2022a].

En último lugar se podría recurrir a la cirugía o lumpectomía, pero previamente debería de haberse intentado el tratamiento mediante alguna estrategia sistémica; algo que se pretende seguir priorizando dada la extrema solución que supone [Waks and Winer, 2019].

En la actualidad, contamos con mejores herramientas y enfoques para abordar el cáncer de mama, lo que nos permite obtener resultados más favorables en comparación con años anteriores, aunque teniendo en cuenta que la incidencia en menores de 45 años ha ido en aumento y se da una posibilidad significativa de recaída aun cuando su descubrimiento es precoz; hace replantear al mundo de la medicina, que es indispensable promover la investigación para el diseño de nuevas terapias y técnicas de detección del cáncer de mama. Algo crucial para una reducción del impacto de dicha enfermedad tanto a nivel nacional como global. Además de lograr una mejora en la calidad de vida, salud y bienestar de las mujeres en todo el mundo [Fundación Grupo Español de Investigación en Cáncer de Mama, 2024].

1.1. Estructura de la memoria

La memoria se divide en las siguientes secciones:

- **Introducción**: en este se describe el contenido y la estructura del proyecto realizado, además de presentar un contexto actual del cáncer de mama.
- **Objetivos**: apartado que presenta las metas analíticas, técnicas y personales que se pretenden cumplir con el proyecto.
- **Conceptos teóricos**: conceptos teóricos básicos para una total comprensión del proyecto propuesto.

- **Metodología**: inicialmente con una descripción de los datos utilizados, presenta las técnicas y metodologías aplicadas para lograr las metas planteadas.
- **Resultados**: resultado final obtenido del procesamiento y análisis de los datos iniciales, mediante el conjunto de técnicas empleadas.
- **Conclusiones**: resumen y comparación de los resultados personales con la literatura contenida en la nube.
- **Líneas futuras**: contiene las ideas y herramientas adicionales, mejoras, y reflexión crítica del proyecto desarrollado para continuar en próximas líneas de desarrollo.

Objetivos

1. Objetivos del estudio.

    a) Presentar los datos derivados de la confrontación de alineamientos.

    b) Realizar un análisis de biomarcadores basado en el alineamiento de secuencias.

    c) Elaborar representaciones visuales para los diferentes alineamientos múltiples.

    d) Efectuar comparaciones utilizando diversos algoritmos (`Clustal` y `Muscle`) y metodologías (perspectiva global y local).

    e) Identificar motivos conservados a través de la obtención de una secuencia consenso.

2. Objetivos de calidad.

    a) Conseguir unos resultados de calidad razonables.

    b) Realizar alineaminetos múltiples con un tiempo de ejecución, y, por tanto, velocidad de ejecución aceptables, sin recurrir a clústeres externos.

3. Objetivos personales.

    a) Aprender a utilizar LaTeX, formatear tablas e imágenes.

    b) Familiarizarse con el entorno de *Github* y realizar con éxito un workflow propuesto.

    c) Conseguir experiencia en el uso de herramientas bioinformáticas.

d) Realizar un estudio que sirva como referencia o plantilla inicial para la realización de posteriores trabajos.

e) Cumplimiento de los ODS planteados por la ONU relacionados con el desarrollo del proyecto.

Conceptos teóricos

3.1. Conceptos Biológicos

El ADN o ácido desoxiriribonucleico, descrito por Watson y Crick en 1953, es una de las moléculas fundamentales de todo organismo, el cual se encarga de contener la información genética necesaria para determinar el desarrollo, funcionamiento y la reproducción de estos [NIH, 2024]. De forma coloquial se podría entender que son como las instrucciones que nos definen a cada uno de nosotros, y provoca que funcionemos de la forma en que lo hacemos. Además de ser reconocido como un pilar y unidad fundamental en los campos de la biología y la genética [Watson and Crick, 1953].

Su estructura presenta una forma de doble hélice, similar a una escalera de caracol. Cada una de las hélices está formada por una serie de bases nitrogenadas dispuestas en un orden y sentido específicos (ya sea 5'→ 3' para el caso de la cadena codificante, o 3 '→ 5' para la cadena molde), recibiendo el nombre de secuencia o cadena de ADN. Ambas cadenas están unidas por puentes de hidrógeno formados entre cada par de bases nitrogenadas. Esos puentes de hidrógeno son uniones débiles, las cuales permiten la separación de estas para que tengan lugar diversos procesos biológicos. Las bases nitrogenadas que conforman el alfabeto del ADN, son un total de cuatro: Adenina, Guanina, Citosina y Timina; las cuales, como pares, se emparejan dos a dos, una pirimidina con una purina y viceversa. Uniéndose en la mayoría de las ocasiones por convenido una Adenina (purina) con una Timina (pirimidina), y una Citosina (pirimidina) con una Guanina(purina) [Jorde et al., 2020].

Dentro de una secuencia existen dos tipos principales de regiones, aquellas regiones formadas por nucleótidos que pueden ser transcritas a ARN mensa-

jero (exones) y regiones que inicialmente no pueden ser transcritas (intrones). Los exones, mediante la agrupación en tripletes de nucleótidos, darán lugar a la serie de aminoácidos concatenados correspondientes que, a su vez, conformarán un complejo funcional conocido como proteína (regiones codificantes). Por otro lado, las regiones que no pueden ser transcritas también tienen un papel importante en ciertos procesos regulatorios [Benítez et al., 2024].

En caso de que se produjera otro tipo de unión, se daría a entender que se ha producido un cambio. Las secuencias, vistas como la unidad principal de estudio en bioinformática, están expuestas siempre a una serie de procesos y factores, tanto internos como externos, que pueden llegar producir a cambios sobre la secuencia de nucleótidos, conocidos como mutaciones, dando lugar a variantes de un mismo gen con secuencias distintas [Benítez et al., 2024]. Estas mutaciones pueden ser muy variables pudiendo afectar distintos niveles, desde genéticos, cromosómicos y hasta genómicos; teniendo un papel importante en la determinación del diagnóstico de un paciente, en función de si se causa un efecto benigno o maligno sobre el individuo. A nivel genético se pueden resumir en [UCM, 2024]:

- Sustituciones: una de las mutaciones o cambios más comunes que puede sufrir la secuencia de nucleótidos. Consiste en el cambio de una base nitrogenada por otra (SNV), pudiendo producir cambios de alta o baja importancia, benignos o malignos, dada la degeneración del código genético presente en los tripletes. Pueden ser transiciones (cambios entre bases del mismo tipo) o transversiones (cambio de purina por pirimidina y viceversa).
- Inserciones: introducción atípica de material genético adicional, pueden ser desde una a varias bases nitrogenadas, que puede provocar graves consecuencias en la secuencia, como lo es el desplazamiento del marco de secuencia.
- Deleciones: se trata de la eliminación de material genético, lo cual puede provocar una situación muy similar a la de una inserción, y causar importantes consecuencias.
- Transposiciones: cuando una secuencia de nucleótidos cambia de posición dentro de su propio gen o en otro lugar del genoma.

Para detectar este tipo de cambios, se pueden recurrir a diversas estrategias. Dentro del campo de bioinformática, la forma más habitual de hacerlo es mediante la confrontación de cadenas. Esta comparación puede

representarse de forma visual mediante un alineamiento, el cual puede ser de dos tipos [Blanca and Cañizares, 2010]:

- Alineamiento global. En el que se comparan las secuencias enteras, pudiendo introducirse huecos que igualen longitudes de secuencias.
  - Lo cual supone un coste computacional elevado.
  - Utilizado para secuencias con alta similitud y de longitud parecida.
- Alineamiento local. Donde se alinean las partes más parecidas, permitiendo identificar secuencias con cierto grado de similitud.
  - Con bajo coste computacional, ideal para la comparación de secuencias en bases de datos de gran tamaño. (En realidad es un alineamiento global de secuencias cortas)
  - Utilizado para secuencias con baja similitud y con diferencias en la longitud.

Otra ventaja que puede presentar cualquier tipo de alineamiento es la posibilidad de generar una matriz de frecuencias que pueda mostrar las coincidencias y no coincidencias presentes en el alineamiento obtenido de una forma más sintética.

Cuando un alineamiento pasa de dos a tres o más secuencias de forma parcial o completa, recibe el nombre de alineamiento múltiple. Este tipo de alineamiento puede ser útil para la generación de árboles filogenéticos con el propósito de establecer ciertas relaciones evolutivas, clasificación de las diferentes secuencias en una amplia diversidad de familias, o incluso informa sobre la función, estructura y evolución de una secuencia (dentro de lo cual se incluye la identificación de patrones comunes de un conjunto de secuencias) [Jones and Pevzner, 2004].

Existen dos formas principales de obtener un alineamiento múltiple, ya sea mediante un algoritmo de programación dinámica o un algoritmo heurístico. El primero de ellos, a pesar de su alto coste computacional, da con la mejor solución, aunque para un alineamiento múltiple en pocas ocasiones resulta ser viable, dado que el número de alineamientos a comprobar para dar con el mejor de ellos, podría presentar un tiempo de ejecución y coste computacional exageradamente altos [Jones and Pevzner, 2004].

Por ello, en la gran mayoría de las veces se recurre a los algoritmos heurísticos, algoritmos, que a pesar no devolver la mejor solución, consiguen dar con una óptima solución con un coste computacional suficientemente razonable [Jones and Pevzner, 2004][Gagniuc, 2021].

Estos se pueden clasificar en:

- Progresivos: aquellos en los que se comienza con el alineamiento de dos secuencias y de forma iterativa se van añadiendo el resto de secuencias al alineamiento. En ellos la elección del primer alineamiento es de gran importancia. Un ejemplo de este tipo de algoritmos sería `ClustalW`
- Iterativos: En los que se realiza un alineamiento progresivo y se intenta mejorar el alineamiento moviendo, añadiendo o eliminando `gaps`. Un ejemplo de este tipo de algoritmos sería `Muscle`
- Híbridos: Combinan diferentes estrategias empleando información complementaria (información estructural de las proteínas o bases de datos con información de buenos alineamientos locales).

Gracias a la realización de este tipo de alineamientos, se puede llegar a obtener lo entendido como una secuencia consenso. Aquella secuencia que está compuesta por los nucleótidos más representados o frecuentes, para las distintas posiciones de las secuencias confrontadas. Esto puede ser de especial utilidad a la hora de determinar cambios conservativos ocultos en secuencias patológicas que no se incluyan en las secuencias controles [Jones and Pevzner, 2004][Gagniuc, 2021].

3.2. Conceptos Informáticos

Python es un lenguaje de programación de alto nivel y multiparadigma, desarrollado por Guido van Rossum en 1991. Se trata de un lenguaje con una sintaxis clara y sencilla, de modo que facilita su aprendizaje y mantenimiento [Python Software Foundation, 2024c].

Es conocido por ser potente, flexible y versátil, pudiendo apoyarse en una extensa biblioteca e innumerables paquetes, lo cual le permite utilizarse como herramienta y medio para una amplia variedad de aplicaciones.

Por otro lado, *Jupyter notebook* es una aplicación web de código abierto que permite la creación y divulgación de documentos que mezclan bloques de código y texto narrativo, como si se tratan de un tipo de informe. Donde

el código utilizado puede pertenecer a diversos lenguajes de programación, entre los cuales destacan R y `Python` [Jupyter Team, 2015].

El lenguaje `Python` presentan una gran variedad de estructuras de datos, las cuales permiten la organización, manipulación y almacenamiento de los datos de una manera eficiente. La elección de una estructura adecuada de datos en función de una situación específica puede mejorar la eficiencia y legibilidad del código empleado.

Estas estructuras de datos, en el lenguaje `Python`, pueden dividirse en dos [Python Software Foundation, 2024a]:

- Datos básicos, entre los cuales se incluyen los tipos de datos que se puede asignar a los datos.
- Datos compuestos, aquellas estructuras que permiten agrupar múltiples elementos en una sola unidad.Estos son:
  - Listas: son colecciones ordenadas y mutables de datos con una gran tolerancia al multitipado. En ellas se permite añadir, eliminar y modificar los diferentes elementos.
  - Tuplas: aquellas colecciones ordenadas e inmutables de datos que también pueden presentan multitipado. Una vez creadas no se pueden modificar.
  - Conjuntos: se trata de una colección de datos únicos desordenados. Se emplean para la eliminación de duplicados
  - Diccionarios: son colecciones de datos desordenadas de pares clave-valor. Cada una de las claves presenten en ella es única, y se utiliza para el acceso al diccionario de su contenido asociado. Útiles en la búsqueda y manipulación de datos.

En todo lenguaje de programación, a su vez existe lo conocido como funciones. Bloques de código diseñado para realizar una acción específica. Se emplean como medida de depuración, reutilización y brevedad del código implementado. Estas están formadas por diferentes partes, las cuales pueden resumirse en [Python Software Foundation, 2024d]:

1. Nombre: se trata del identificar único que se utilizar para dar uso a y referirse a la función creada.

2. Parámetros: aquellas variables que la función recibe cuando es llamada, permitiendo la utilización de datos externos para que puedan ser utilizados. Estos son opcionales, pudiendo tener valores predefinidos.
3. Cuerpo: Es la parte principal de la función, siendo el bloque de código que contiene las instrucciones que realizará la función llamada y las cuales la definen.
4. Retorno: parámetro opcional que permite a la función devolver un valor al exterior tras ejecutarse las instrucciones necesarias pertenecientes al cuerpo. Vienen definidos por el comando return.

Otro elemento común en programación son las bibliotecas y paquetes que presente el lenguaje utilizado. Las bibliotecas contienen unas colecciones de funciones, clases y métodos predefinidos que extienden la funcionalidad básica del lenguaje empleado; soliendo organizarse en módulos o archivos de código.

Mientras que un paquete se trata de una carpeta o estructura más grande organizada que puede contener diversos módulos, bibliotecas o recursos relacionados entre sí [Python Software Foundation, 2024b].

Un ejemplo sería la biblioteca de *Biopython* [Chang et al., 2024] empleada en el proyecto. Una biblioteca diseñada para la manipulación y análisis de datos biológicos, mediante la disponibilidad de diversas herramientas y módulos incluidos en ella. La biblioteca contiene una serie de objetos que ayudan a categorizar y manipular los diferentes datos biológicos, entre los cuales se destacan:

- `Seq`: objeto correspondiente a una secuencia biológica de ADN, ARN o aminoácidos, con sus diferentes métodos específicos asociados.
- `SeqRecord`: aquel objeto que se encuentra compuesto por una secuencia biológica y unos metadatos asociados, como lo son el identificador, una descripción y ciertas anotaciones adicionales.
- `MultipleSeqAlignment`: representa un alineamiento múltiple de secuencias, en el cual se muestran las diferencias y similitudes encontradas en las secuencias comparadas.

Los datos obtenidos se han podido conseguir gracias al empleo de bases de datos biológicas, para este caso `NCBI`. Una base de datos biológica es un conjunto de datos, biológicos organizado de tal modo que permita obtener con rapidez diversos tipos de información [RAE, 2024].

3.3. Estado del arte y trabajos relacionados.

Como uno de los primeros pasos, se realizó una búsqueda bibliográfica para tomar alguna referencia de los proyectos o trabajos realizados hasta la fecha, que tuvieran relación con los objetivos planteados. El proceso de selección fue largo y tedioso, dado que este tipo de herramientas bioinformáticas a pesar de estar en auge, los estudios son muy limitados y no se encuentran tan asentados como en otros campos científicos. Por lo tanto, se optó por la selección de una combinación de estudios de diferentes campos relacionados, pudiendo estos dividirse, en aquellos con una carga más biológica y biológica y otros, al contrario, con una perspectiva más informática.

Documentación clínica y biológica

En esta sección se encontró gran cantidad de artículos relacionados con el cáncer de mama. Alguno de ellos fueron:

1. Análisis de las firmas mutacionales que registran la actividad mutagénica en el desarrollo del carcinoma de mama, que puede suponer una herramienta adicional para el diagnóstico de precisión [Nik-Zainal and Morganella, 2017].

2. Análisis pronóstico y predictivo de las mutaciones concomitantes que tienen lugar en los genes *PIK3CA* y *TP53* para cáncer de mama [Lin et al., 2023].

Documentación bioinformática

Los artículos que aplican herramientas bioinformáticas han tenido cierta influencia en la elaboración del proyecto. Entre ellos se destacan:

1. Identificación y predicción de biomarcadores mediante el uso de microarrays de ARNm para el diagnóstico y tratamiento de cáncer de mama [Zeng et al., 2021].

2. Estudio del papel de `PTPRZ1` en diferentes tumores, destacando el cáncer cerebral, mediante tecnología NGS y el paquete de *Biopython* [Kumari and Gupta, 2023].

3. Detección de genes potenciales de cáncer gástrico mediante el empleo de múltiples herramientas bioinformáticas y métodos de aprendizaje automático [Chen et al., 2022].

Metodología

4.1. Descripción de los datos.

Gran parte de los datos fueron recogidos del Centro Nacional para la Información Biotecnológica de Estados Unidos, mejor conocido como `NCBI`, y principalmente de la base de datos de `Nucleotide`, mediante la funcionalidad de `clipboard` que presenta la biblioteca.

Los cuales pueden dividirse en dos conjuntos de datos principales: uno asociado a pacientes enfermos y los otros a pacientes sanos que actúan como controles en este proyecto. Cuatro archivos de formato `fasta` por parte de los pacientes enfermos, y cinco archivos más del mismo formato por los pacientes sanos. Cada uno de los archivos, correspondientes a los genes y regiones codificantes a estudio causantes de cáncer de mama: *BRCA1*, *BRCA2*, *TP53* y *PIK3CA*.

Cada archivo `fasta` de cada gen, contiene un número variable de secuencias de ADN, todas ellas en sentido 5'→ 3'(cadenas codificantes) de diferentes tamaños y contenidos, pero con una estructura común.

El resto, fue extraído de la base de datos de *gnomAD*. Consisten en un par de archivos de formato `csv`, que representan la información de diferentes variantes genéticas documentadas de forma tabular.

Todo ello esta detallado en el apéndice *D: Descripción de adquisición y tratamiento de los datos.*

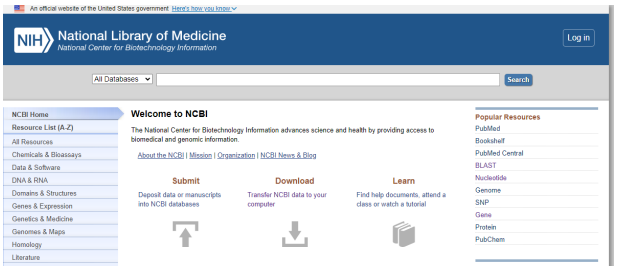

Figura 4.1: Portal de inicio de `NCBI`. Fuente: [NCBI, 2024b].

```
>OM524595.1 Homo sapiens isolate Sh520B1 breast cancer type 1 susceptibility protein (BRCA1) gene,
exon 20 and partial cds
AAGATCTTCTGATCCAGTAGTGTTCTGGACATTGGACTGCTTGTCCCTGGGAAGTAGCAGCAGAAATCAT
CAGGTGGTGAACAGAAGAAAAAGAAAAGCTCTTCCTTTTTGAAAGTCTGTTTTTTGAATAAAAGCCAATA
TTCTTTTATAACTAGATTTTCCTTCTCTCCATTCCCCTGTCCCTCTCTCTTCCTCTCTTCTTCCAGATCT
TCAGGGGGCTAGAAATCTGTTGCTATGGGCCCTTCACCAACATGCCCACAGGTAAGAGCCTGGGAGAACC
CCAGAGTTCCAGCACCAGCCTTTGTCTTACATAGTGGAGTATTATAAGCAAGATCCCACGATGGGGGTTC
CTCAGATTGCTGAAATGTTCTAGAGGCTATTCTATTTCTCTACCACTCTCCAAACAAAACAGCACCTAAA
TGTTATCCTATGGCAAAAAAAAACTATACCTTGTCCCCCTTCTCAAGAGCATGAAGGTGGTTAATAGTTA
GGATTCAGTATGTTATGTGTTCAGAGGCGTTG

>OM524594.1 Homo sapiens isolate Sh420B1 breast cancer type 1 susceptibility protein (BRCA1) gene,
exon 20 and partial cds
AAGATCTTCTGAATCCATGTAGTGTTCTGGACATTGGACTGCTTGTCCCTGGGAAGTAGCAGCAGAAATC
ATCAGGTGGTGAACAGAAGAAAAAGAAAAGCTCTTCCTTTTTGAAAGTCTGTTTTTTGAATAAAAGCCAA
TATTCTTTTATAACTAGATTTTCCTTCTCTCCATTCCCCTGTCCCTCTCTCTTCCTCTCTTCTTCCAGAT
CTTCAGGGGGCTAGAAATCTGTTGCTATGGGCCCTTCACCAACATGCCCACAGGTAAGAGCCTGGGAGAA
CCCCAGAGTTCCAGCACCAGCCTTTGTCTTACATAGTGGAGTATTATAAGCAAGATCCCACGATGGGGGT
TCCTCAGATTGCTGAAATGTTCTAGAGGCTATTCTATTTCTCTACCACTCTCCAAACAAAACAGCACCTA
AATGTTATCCTATGGCAAAAAAAAACTATACCTTGTCCCCCTTCTCAAGAGCATGAAGGTGGTTAATAGT
TAGGATTCAGTATGTTATGTGTTCAGTGGC
```

Figura 4.2: Ejemplo secuencias `fasta` patológicas *BRCA1*. Fuente: elaboración propia.

4.2. Técnicas y herramientas.

Herramientas

En el desarrollo del proyecto se han utilizado una gran variedad de herramientas de alta importancia para cubrir los objetivos y metas planteados. Cada una de ellas, fueron seleccionadas por diversos motivos para tratar de la mejor manera los diferentes retos y puntos de interés que fueron surgiendo en el transcurso de las etapas que conforman el proyecto. En esta sección se abordarán las principales herramientas y bibliotecas empleadas, así como las alternativas consideradas, destacando su funcionalidad, ventajas y como han contribuido al éxito del proyecto.

En primer lugar, al tratarse de un proyecto centrado en el software, se debía elegir el lenguaje con el que elaborar las diferentes instrucciones y código necesario para lograr producir el cambio deseado sobre los datos obtenidos. El escogido fue el lenguaje `Python` [Python Software Foundation, 2024c].

Un lenguaje destinado a las aplicaciones web, el desarrollo de software, la ciencia de datos y el `Machine learning`. Un lenguaje con una sintaxis clara y legible, y de gran versatilidad y portabilidad, cuya elección en gran medida fue motivada por ser el lenguaje de mayor uso en el transcurso del grado. A su vez, presentaba la posibilidad de emplear los cuadernos de *Jupyter notebook* [Jupyter Team, 2015], siendo esta una alternativa de especial interés al poder combinar áreas de documentación y bloques de código integrado en un único archivo, además de presentar una gran comodidad para el estudiante.

También se llegó a plantear el uso de algún lenguaje adicional o alternativo a este como es el caso de `R`. Un lenguaje a su vez muy extendido en el campo de ciencias de datos y el sanitario. Presentaba una sintaxis clara, gran aplicabilidad para información tabulada y cierta experiencia obtenida por parte del estudiante, que, a pesar de ello, no fueron unos argumentos lo suficientemente significativos como para que finalmente se llevara a la práctica [R Foundation, 2024].

Dado los intereses y los objetivos que se habían planteado, se asimiló que el lenguaje `Python` fuera a cubrirlos de una forma óptima.

Dentro de la gran variabilidad de paquetes que presenta este lenguaje, y dado que el proyecto presentaba una temática biológica y sanitaria, se exploró la posibilidad de mezclar ambos campos.

De forma que se terminó encontrando la herramienta clave y fundamental de este proyecto, el bien conocido paquete biológico por excelencia del lenguaje `Python`, *Biopython* [Chang et al., 2024]. Un paquete gratuito que incluye una inmensa cantidad de herramientas aplicables para casi cualquier proceso biológico a analizar. La utilidad dentro del proyecto no pudo ser mayor, ni mejor; aprovechando su uso para cada proceso incluido en él. Desde la realización de alineamientos en conjunto con softwares que posteriormente serán comentados, la facilidad en la lectura y guardado de ficheros con información biológica, generación de secuencias consenso y hasta la posibilidad de la obtención de matrices de frecuencias, que dieron lugar a una posible alternativa de detección de biomarcadores.

En resumen, una herramienta que ha facilitado en gran medida el trabajo desarrollado en el proyecto.

Como se ha mencionado en el párrafo anterior, se requirió de la utilización de dos softwares adicionales para lograr numerosos alineamientos múltiples de las secuencias control y patológicas, mediante los cuales se determinaría una secuencia consenso definida para cada uno de ellos; siendo estas dos funcionalidades el tema central del proyecto y a través de los cuales se

consigue un intento de detección de biomarcadores. Los softwares referidos son `Clustal Omega` [Dineen, 2016] y `Muscle` [Edgar, 2021].

Ambos se tratan de unos programas, que una vez incorporados al dispositivo, permiten su integración en `Python` mediante el paquete *Biopython*, y consiguen realizar de una forma cómoda y sencilla el alineamiento múltiple de las secuencias contenidas en archivos de formato `fasta`. La elección de recurrir a ambos, en lugar de únicamente a uno de ellos, esta basada en gran parte debido a lo comentado en el apartado de conceptos teóricos y es meramente con una finalidad comparativa. Tanto un programa como otro consiguen el mismo objetivo, pero cada uno de ellos mediante diferentes procesos y mecanismos, que los hacen más adecuados de ejecutar en determinadas ocasiones.

Para la visualización de los alineamientos múltiples también se barajaron diferentes alternativas. Herramientas en auge como lo es `MSABrowser` [Bilgin et al., 2024], una herramienta web gratuita y de código abierto, que permite visualizar de forma rápida, explorar y navegar de forma interactiva, además de integrar de forma sorprendente anotaciones específicas de variantes. O incluso aquellas más establecidas y reconocidas a nivel global, como lo es el `Multiple Sequence Alignment Viewer` [NCBI, 2023] de `NCBI`. Un programa web capaz de visualizar el alineamiento de nucleótidos y proteínas, incluir archivos de secuencias alineadas en diferentes formatos e incluir una secuencia consenso generada automáticamente con sus diferentes anotaciones genéticas.

Pero, debido por un lado a que para `MSAbrowser` se requería de cierto conocimiento del formato `Json` [JSON, 2024], la secuencia consenso que generaba era poco manipulable y el estado avanzado del proyecto; junto con que se habían encontrado otras alternativas que cubrían de igual manera y estaban más adaptadas a nuestra situación, que el visualizador de `NCBI`, se renunció a ellas.

La función de obtención de consenso del módulo de `AlignInfo` perteneciente a *Biopython* y el paquete de `Bokeh` [de Ven, 2024]. utilizado en una función referenciada del *notebook* de *Jupyter*, fueron las alternativas escogidas.

Bibliotecas

El `NCBI` o Centro Nacional para la Información Biotecnológica se trata de una biblioteca en línea que permite el acceso gratuito a los textos completos de numerosas bases de datos que incluye, con el propósito de lograr convertir-

se en una fuente de información integral de biología molecular [NCBI, 2024b]. Asimismo, goza de múltiples herramientas bioinformáticas que hoy en día son utilizadas por múltiples profesionales del ámbito bioinformático a nivel global [NCBI, 2024a].

Por todo ello se trata de una de las mayores fuentes de datos de temática biológica y gran reconocimiento a nivel internacional, pudiendo encontrar en ella según los términos de búsqueda utilizados, gran cantidad de información de interés de una forma rápida y accesible. Como es el caso de las secuencias nucleotídicas, las cuales se encuentra rigurosamente documentada, permiten su descarga en el dispositivo; una gran ventaja frente al resto de bases de datos consultadas, donde, en la gran mayoría, esta opción se encuentra limitada o directamente no se presenta.

Durante la realización se llegaron a considerar otras fuentes de información o bibliotecas biológica de gran utilidad; como es el caso de *gnomAD* [gnomAD Steering Committee, 2024a] y *TCGA*. *GnomAD* recoge de forma óptima grandes cantidades de variantes sobre un gen, y las muestra tanto de forma visual como tabulada, garantizando una compresión más sencilla. *TCGA* [NCI and National Human Genome Research Institute, 2006], por otra parte, incluye gran variedad de información debido a la integración de varios proyectos y numerosas herramientas de visualización, permitiendo realizar múltiples análisis de los datos, desde diferentes perspectivas en un propio apartado denominado centro de análisis.

Esta primera se pudo dar uso durante la fase de análisis del proyecto, para la obtención de ciertos archivos `csv` con información de las diferentes variantes tanto inciertas, como patológicas como benignas, que pueden presentar las muestras descargadas para los genes y exones de interés. *TCGA* por el contrario. y a pesar de ser una base con muchas ventajas y herramientas de análisis, permitió un uso limitado de su información dada su restricción a investigadores disponedores de cierta licencia y ley de privacidad de datos.

Ya en último lugar, se encuentra el navegador genómico de la `UC Santa Cruz` [UCSC, 2024d]. Un navegador lineal interactivo que tuvo sus inicios en el Proyecto Genoma Humano y es utilizado hasta hoy en día. La libertad y rapidez con la que se puede navegar entre las diferentes versiones genómicas, para múltiples organismos, además de la increíble cantidad de información que ofrece, la convierte en una biblioteca complementaria básica para cualquier proyecto o trabajo del campo de la bioinformática. Para este proyecto, se llegó a recurrir en contadas ocasiones a esta, dada su utilidad y experiencia del estudiante en ella.

Resultados

5.1. Resumen de resultados.

Como resultados finales del proyecto se han obtenido posibles biomarcadores benignos y malignos de aquellos exones pertenecientes a genes susceptibles a cáncer de mama. Los resultados obtenidos a partir del proceso de análisis realizado pueden dividirse en:

Resultados de Secuencias individuales

Dentro de dicho estudio se analizó el contenido de GC, longitud y contenido correspondientes de cada una de las secuencias consenso, tanto patológicas como controles.

Secuencias Patológicas

En el estudio independiente de cada una de las secuencias consenso patológicas se encuentra recogido en las tablas: Tabla 5.1 para los resultados conseguidos mediante el algoritmo de *MSA*(Multiple Sequence Alignment) `Clustal`, y Tabla 5.2 para los obtenidos mediante `Muscle`.

Secuencias Controles

Por otro lado, los resultados logrados para las secuencias consenso de origen sano, también se dividen y recogen en dos tablas correspondientes. Tabla 5.3 para `Clustal` y Tabla 5.4 para `Muscle`.

Resultados Secuencias Consenso Patológicas con `Clustal`			
Genes/Exón	Contenido GC	Longitud	Nucleótidos
BRCA1/5	29 %	368 pb	'C': 54, 'T': 125, 'G': 55, 'A': 133, 'X': 1
BRCA1/19	36 %	255 pb	'G': 39, 'X': 49, 'A': 66, 'C': 36, 'T': 65
BRCA2/2	34 %	253 pb	'T': 76, 'C': 43, 'A': 86, 'G': 42, 'X': 6
BRCA2/11	24 %	710 pb	'G': 52, 'T': 109, 'A': 174, 'X': 333, 'C': 42
PIK3CA/9	34 %	196 pb	'C': 29, 'A': 80, 'G': 38, 'T': 46, 'X': 3
TP53/5	52.3 %	172 pb	'T': 49, 'C': 34, 'A': 33, 'G': 56
TP53/8	46.7 %	197 pb	'G': 34, 'A': 54, 'C': 58, 'T': 51

Tabla 5.1: Tabla de resultados de las secuencias consenso patológicas para `Clustal`

Resultados de Alineamientos pareados

Los alineamientos pareados fueron analizados entre secuencias consenso patológicas y controles, y, secuencias patológicas aleatorias y secuencias consenso controles. El primer estudio, para la determinación de las mutaciones comunes, y el segundo para el reconocimiento de aquellas específicas. Estos datos de valor, son recogidos en la Tabla 5.5.

Resultados de Alineamientos múltiples

En último lugar se obtuvieron de forma tabulada los cambios más habituales dentro de las secuencias controles y patológicas, para alineamientos múltiples con `Clustal` en Tabla 5.6 y con `Muscle` en la Tabla 5.7.

Resultados Secuencias Consenso Patológicas con `Muscle`			
Genes/Exón	Contenido GC	Longitud	Nucleótidos
BRCA1/5	29 %	421 pb	'X': 3, 'T': 139, 'G': 70, 'A': 145, 'C': 64
BRCA1/19	39 %	276 pb	'X': 25, 'G': 50, 'T': 77, 'A': 76, 'C': 48
BRCA2/2	34 %	253 pb	'T': 77, 'C': 43, 'A': 85, 'G': 42, 'X': 6
BRCA2/11	31 %	904 pb	'C': 101, 'X': 258, 'A': 260, 'T': 184, 'G': 101
PIK3CA/9	34 %	196 pb	'C': 29, 'A': 80, 'G': 38, 'T': 46, 'X': 3
TP53/5	52.6 %	173 pb	'T': 49, 'C': 34, 'A': 33, 'G': 57
TP53/8	46.7 %	197 pb	'G': 34, 'A': 54, 'C': 58, 'T': 51

Tabla 5.2: Tabla de resultados de las secuencias consenso patológicas para `Muscle`

5.2. Discusión.

En los resultados correspondientes al estudio básico de secuencias consenso, se puede observar que se analizaron las secuencias en función a tres aspectos característicos: contenido de GC, longitud que presenta y el contenido de nucleótidos que la conforman. El estudio de contenido en GC como parámetro presenta gran importancia en bioinformática, debido a que puede representar la estabilidad del ADN y ARN, la clasificación de algoritmos y comparación de genomas. También el contenido dentro de la secuencia, en especial para el caso que acontece, pudiendo servir como una medida de la heterogeneidad contenida en el alineamiento múltiple. Tras el estudio, se puede determinar:

- Para el caso de las secuencias patológicas consenso, los análisis muestran diferencias en los resultados según el algoritmo utilizado. Para *BRCA1/5*, *BRCA2/2*, *PIK3CA/9*, *TP53/5* y *TP53/8*, los resultados son similares en longitud, contenido y porcentaje de bases desconocidas. Sin embargo, en *BRCA1/19* y *BRCA2/11*, las diferencias son notables en longitud, contenido y cantidad de bases ambiguas, indicando varia-

Resultados Secuencias Consenso Controles con `Clustal`			
Genes/Exón	Contenido GC	Longitud	Nucleótidos
BRCA1/5	34.3 %	372 pb	'C': 63, 'T': 140, 'A': 103, 'G': 64, 'X': 2
BRCA2/11	23.4 %	287 pb	'G': 16, 'X': 146, 'C': 17, 'T': 47, 'A': 61
TP53/5	64.2 %	148 pb	'G': 38, 'A': 28, 'T': 25, 'C': 57
TP53/8	63.1 %	84 pb	'G': 29, 'A': 26, 'C': 24, 'T': 5

Tabla 5.3: Tabla de resultados de las secuencias consenso controles para `Clustal`

Resultados Secuencias Consenso Controles con `Muscle`			
Genes/Exón	Contenido GC	Longitud	Nucleótidos
BRCA1/5	34.3 %	372 pb	'C': 63, 'T': 141, 'A': 102, 'G': 64, 'X': 2
BRCA2/11	31.9 %	295 pb	'X': 151, 'T': 36, 'A': 62, 'C': 20, 'G': 26
TP53/5	64.2 %	148 pb	G': 38, 'A': 28, 'T': 25, 'C': 57
TP53/8	63.1 %	84 pb	'G': 29, 'A': 26, 'C': 24, 'T': 5

Tabla 5.4: Tabla de resultados de las secuencias consenso controles para `Muscle`

bilidad en los alineamientos y afectando los resultados. Esto resalta la necesidad de elegir el algoritmo adecuado según la similitud de las secuencias estudiadas.

- La variabilidad entre diferentes algoritmos es prácticamente inexistente en el caso de las secuencias controles. Tanto `Muscle` como `Clustal` producen los mismos resultados para estas secuencias, con longitudes, contenidos de GC y cantidades de bases ambiguas similares. La variabilidad observada en las secuencias consenso patológicas no se presenta en las secuencias controle, lo cual justifica estos resultados. *BRCA2/11* es la excepción, mostrando que las secuencias controles a

Genes	Secuencia aleatoria - Consenso		Entre consensos
	Primera secuencia - Consenso	Segunda secuencia - Consenso	
BRCA1	T>A	T>A	T>A
BRCA2	T>A	A>T	A>T
TP53/5	C>T	C>T	C>T
TP53/8	G>A	G>A	G>A

Tabla 5.5: Mutaciones más frecuentes por cada gen, entre secuencias consenso patológica - control, y secuencias patológicas aleatorias - secuencia consenso control

	Alineamiento `Clustal`	
Gen/Exón	Controles	Patológicos
BRCA1/5	R>G/G>R	A>T/T>A
BRCA1/19	No hay controles	A>G/G>A
BRCA2/2	No hay controles	A>G/G>A
BRCA2/11	A>T/T>A	A>T/T>A
PIK3CA/9	No hay controles	A>G/G>A
TP53/5	A>C/C>A	A>G/G>A
TP53/8	No hay cambios en secuencias controles	G>T/T>G

Tabla 5.6: Sustituciones de mayor frecuencia en los alineamientos de múltiples secuencias de controles y patológicos mediante el algoritmo de `Clustal`

	Alineamiento `Muscle`	
Gen/Exón	Controles	Patológicos
BRCA1/5	C>A/A>C	A>T/T>A
BRCA1/19	No hay controles	A>G/G>A
BRCA2/2	No hay controles	A>G/G>A
BRCA2/11	A>T/T>A	A>T/T>A
PIK3CA/9	No hay controles	A>G/G>A
TP53/5	A>C/C>A	A>G/G>A
TP53/8	No hay cambios en secuencias controles	G>A/A>G

Tabla 5.7: Sustituciones de mayor frecuencia en los alineamientos de múltiples secuencias de controles y patológicos mediante el algoritmo de `Muscle`

partir de las cuales se genera la secuencia consenso son heterogéneas, aunque sean pocas.

- Por último, al comparar entre secuencias sanas y enfermas, encontramos:

  - *BRCA1/5*: La secuencia consenso tiene una longitud variable de unas 50 pb (pares de bases) aproximadamente. A su vez, la secuencia control muestra un mayor contenido de GC, lo que sugiere mayor estabilidad, y una ligera disminución en el número de bases ambiguas, indicando bases más conservadas.
  - *BRCA2/11*: Existe una diferencia extrema en la longitud de las secuencias consenso, con más de 600 bases de diferencia. Por otro lado, el contenido de GC es similar en ambas, con diferencias mínimas en porcentaje. Y cabe recalcar, que la secuencia control tiene aproximadamente la mitad de bases ambiguas que la secuencia patológica, lo que es algo inesperado.
  - *TP53/5*: La longitud en ambos casos no es la misma, pero llegan a ser ciertamente similares. El contenido de GC en la secuencia es mayor para el caso control, lo cual era de esperar.Ninguno de ellos presenta base ambigua alguna. Y para finalizar, la cantidad y tipos de aminoácidos, son muy distintos, debido en parte a la diferencia del número de bases existente.
  - *TP53/8*: Para el exón 8 del mismo gen, se encuentran longitudes muy distintas, logrando incluso la secuencia patológica doblar la longitud de la secuencia control. El contenido de GC es ampliamente superior en la secuencia control analizada. Para ambas, el contenido de bases ambigua es nulo, y el gráfico de aminoácidos, al igual que se mencionó en el anterior exón, es distinto en número y tipos.

Como ya se mencionó, el contenido de GC es un marcador biológico de estabilidad genómica .Tal y como se ha demostrado [Oliver and Marín, 1996], las regiones codificantes o exones, presentan un mayor contenido de GC que las regiones intrónicas. Esto se debe, a una cuestión de importancia, las regiones codificantes presentan una información más valiosa que las no codificantes, y por ello muestran una gran estabilidad respecto a las regiones intrónicas, aquellas donde los cambios o mutaciones están a la orden del día. Las secuencias patológicas, secuencias teóricamente similares a las controles, pero con alguna mutación contenida, presentan menores porcentajes, dando

a entender que han sufrido alguna clase de modificación alternativa, y de origen posiblemente patológico.

Los cambios de longitud y contenido no son más que una prueba adicional y fiel reflejo de algo ya conocido desde el principio, la alta heterogeneidad lejos de la habitual entre secuencias de un mismo gen para un exón específico.

Contrario a lo que en un principio cabría esperar, los alineamientos globales obtenidos muestran menores similitudes que los alineamientos locales destinados al análisis de secuencias divergentes. Esto revela que las secuencias consenso patológicas y controles son más diferentes que similares. En los gráficos `Dot plot`, se esperaba encontrar una diagonal descendente continua que representara las coincidencias, pero esto no se observa para ninguno de los casos, indicando diferencias significativas entre las secuencias. El algoritmo utilizado para calcular cada uno de los alineamientos influye en los resultados. Se ha demostrado que `Muscle` generalmente ofrece mejores resultados que `Clustal` para los alineamientos globales realizados.

Estas diferencias y a pesar de realizar el análisis con una herramienta tan útil como los alineamientos de secuencias, los resultados no son lo suficientemente buenos ni representativos para lo que se pretendía comprobar.

Los alineamientos largos como los obtenidos, pueden llegar a dificultar la detección visual de los diferentes biomarcadores presentes en las secuencias, por ello se elaboró adicionalmente un análisis basado en las matrices de frecuencias obtenidas. Estos no mostraron mutaciones de tipo deleción en las secuencias patológicas, sin embargo las inserciones eran numerosas. Las grandes diferencias de longitud entre las secuencias control y patológicas, justifican esta mayor frecuencia. Por otro lado, las sustituciones detectadas variaban en función del exón estudiado, y fueron recogidas en 5.5:

- En el caso del exón 5 de *BRCA1*, se obtiene como mutación más frecuente la sustitución T>A. Recurriendo a la bibliografía existente sobre las firmas mutacionales más frecuentes en cáncer de mama, para el caso de *BRCA1* suelen darse 3 mutaciones: C>A(8), C>G(3) y T>A(8). La segunda presenta una gran relación con una deficiencia en la reparación por recombinación homóloga y presentando una prevalencia sobre el 20 % de los cánceres de mama. Por otro lado, la primera y la tercera, presentes en más del 60 % de los cánceres de mama, se encuentran vinculadas con un aumento en la deficiencia de recombinación homóloga y en etapas tardías de la evolución del cáncer, pero también está presente en niveles más bajos en muchos otros tumores [Nik-Zainal and Morganella, 2017].

De forma que se puede afirmar que la mutación detectada para *BRCA1/5*, T>A, muestra una naturaleza maligna

- Para *BRCA2/11* la mutación más frecuente es una variante maligna del exón, provocando una mutación sin sentido que resultaba en un aminoácido diferente. También se identificaron otras sustituciones malignas y se determinó la ausencia de sustituciones benignas en los datos ([gnomAD Steering Committee, 2024b].

- Dentro de las posibles mutaciones de *TP53* para cáncer de mama, las más frecuentes son: G>A (28.88 %), C>T(23.99 %) y en menor medida G>T(10.77 %) y A>G(10,47 %) [Cosmic, 2024d]. Si se comparan las mutaciones reconocidas mediante el *notebook* desarrollado y *Biopython*, se destaca la mutación C>T por parte del exón 5, y G>A para el exón 8 como biomarcadores. De forma que coincide con la bibliografía consultada, y se pueden determinar cómo biomarcadores malignos para dicho gen.

En el alineamiento en el que se confrontaban secuencias patológicas aleatorias con la secuencia consenso control para dicho exón, tenía por objetivo la identificación de sustituciones singulares. Para ciertos exones, como son los correspondientes a *BRCA1*, exón 5 y exón 8 de *TP53*, en la comparación de sus diferentes secuencias con la secuencia consenso, se obtienen las mismas mutaciones habituales. Y de manera opuesta, para la primera secuencia del exón 11 de *BRCA2*, se obtiene una diferente mutación de mayor frecuencia (T>A), con respecto a la segunda y la consenso (A>T); ambas malignas [gnomAD Steering Committee, 2024b]. Aunque la similitud de nucleótidos presente da a creer de la existencia de algún tipo de relación entre ambas. Esta singularidad mostrada por las secuencias de *BRCA2*, estaría relacionada por la alta variabilidad mencionada en reiteradas ocasiones entre sus secuencias.

Algoritmos de Alineamientos múltiples

Tal y como se muestran en la Tabla 5.6 y Tabla 5.7, existen diferencias entre aquellos cambios presentados en controles y patológicos, salvo para el exón 11 del gen *BRCA2*. Algunas de las mutaciones se mantienen, como es el caso de *BRCA1*, *BRCA2* y el exón 8 de *TP53* para `Muscle`. La conservación de estos cambios, afianzan algunas de las posibles sustituciones patológicas detectadas y la diversidad del resto, da a entender la gran posibilidad de mutaciones que pueden llegar a tener lugar para ambos tipos de secuencias.

Las mutaciones detectadas varían según el algoritmo utilizado. En particular, esto se observa en el exón 8 del gen *TP53* y el exón 5 de *BRCA1*. El algoritmo de `Muscle` ofrece mejores resultados en comparación con `Clustal`, gracias a su estrategia de funcionamiento.

Comparando con la bibliografía consultada:

- Dentro de los controles, se hallarán de forma teórica mutaciones que raramente sean patológicas:
  - *BRCA1/5*: las mutaciones de C>A y A>C, son consideradas como poco frecuentes dentro del cáncer de mama, tiendo lugar aproximadamente en el 4.1 % y 3.59 % de los diagnósticos de cáncer de mama [Cosmic, 2024a].
  - *BRCA2/11*: al igual que en el anterior gen, los cambios A>T(6.45 %) y T>A((3.55 %), para *BRCA2* son pocos frecuentes en cáncer de mama malignos reconocidos [Cosmic, 2024b].
  - *TP53/5*: en el gen *TP53* las mutaciones de A>C y C>A presentan una probabilidad de sucedan del 1.7 % y 2.9 % [Cosmic, 2024d].
- Al contrario sucedería con las mutaciones en las secuencias patológicas, cuyas mutaciones presentan las siguientes frecuencias [Cosmic, 2024d, Cosmic, 2024a, Cosmic, 2024b, Cosmic, 2024c]:
  - *BRCA1/5*: A>T $\longrightarrow$ 5.13 % T>A $\longrightarrow$ 4.1 %
  - *BRCA1/19*: A>G $\longrightarrow$ 17.44 % G>A $\longrightarrow$ 22.05 %
  - *BRCA2/2*: A>G $\longrightarrow$ 10.97 % G>A $\longrightarrow$ 14.84 %
  - *BRCA2/11*: A>T $\longrightarrow$ 6.45 % T>A $\longrightarrow$ 3.55 %
  - *PIK3CA/9*: A>G $\longrightarrow$ 47.3 % G>A $\longrightarrow$ 36.28 %
  - *TP53/5*: A>G $\longrightarrow$ 10.47 % G>A $\longrightarrow$ 28.88 %
  - *TP53/8* `Clustal`: G>T $\longrightarrow$ 10.77 % T>G $\longrightarrow$ 4.03
  - *TP53/8* `Muscle`: A>G $\longrightarrow$ 10,47 % G>A $\longrightarrow$ 28.88 %

Conclusiones

Tras la evaluación de los resultados en la discusión, y habiendo sido el proyecto finalmente completado, se puede determinar:

- Ha sido posible realizar un reconocimiento de biomarcadores mediante el paquete de *Biopython* y ciertos softwares complementarios.
- El *pipeline* de procesamiento desarrollado ha sido correcto, a pesar de no obtener de forma general resultados válidos o aplicables, aunque sí razonables dado el contexto experimentado.
- Se han experimentado las grandes dificultades que supone realizar un estudio en el campo de la bioinformática.
- Se han sentado las bases para futuros proyectos, pudiendo servir como referencia.
- Se ha comprobado la importancia de que los datos iniciales, sean homogéneos y de cantidad.
- Aplicación de diversas técnicas y estructuras sobre el proyecto de `Overleaf`.

6.1. Aspectos relevantes

Los aspectos más interesantes e importantes a abordar de cada etapa son:

Obtención de datos

Una de las etapas cruciales en el desarrollo del proyecto, condicionando en gran medida al resto de ellas. Se barajaron diferentes bases de datos posibles para conseguir el estudio y objetivos planteados, pero solamente una de ellas proporcionaba en gran medida lo que se buscaba, `NCBI` [NCBI, 2024b]. Para el resto de bases de datos, o no incluían la opción de descarga de secuencias nucleotídicas, o esta se encontraba restringida a ciertas personas las cuales presentaran una serie de requisitos [NCI and National Human Genome Research Institute, 2006]. En el campo de la bioinformática, se reconoce la dificultad de obtener datos de estudios homogéneos y en grandes cantidades. Existen múltiples variables a considerar, y la elección de sus valores pueden determinar la validez del estudio.De forma, que se podría considerar como el talón de Aquiles del proyecto y cualquier estudio bioinformático Todo ello, justificó la cantidad de tiempo y recursos invertidos en esta etapa, que hasta relativamente poco no se disponía de la totalidad de los datos. Aun así la descarga de datos patológicos y controles no se pudo obtener ni en la misma cantidad, ni de la misma información a estudiar. No obstante, y pese a los esfuerzos tomados, la cantidad y de homogeneidad pretendidas en los datos no se lograron, condicionando desde un inicio la validez del trabajo elaborado.

Tratamiento

El tratamiento sobre los datos conseguidos fue el mismo para los datos patológicos y controles. Este estuvo determinado por el formato e información contenida de los datos descargados, teniendo que realizar un estudio inicial del cual dependería la estrategia a seguir. Sobre los exones de interés se aplicó un filtrado basado en la mediana de las longitudes presentadas por las secuencias asociadas, eliminando de gran parte de los datos conseguidos y provocando la eliminación de algunos de los exones previstos a análisis. De forma que ya se estaban experimentando las consecuencias resultantes de la etapa anterior.

Por otro lado, el tiempo requerido no fue tanto como en un principio se esperaba, ya que se realizó de una forma rápida y eficaz al conocer de antemano las herramientas necesarias y la estructura de los archivos tratados.

Análisis

Desde un inicia se tuvo claro que el análisis a realizar para el reconocimiento de los biomarcadores genéticos debía basarse en la confrontación o

alineamiento de secuencias. El alineamiento de secuencias es la herramienta fundamental de la bioinformática a partir de la cual obtener una información añadida. Existen múltiples formas de realizarlo y representarlo visualmente, transmitiendo al investigador la información deseada. Adicionalmente se creó un análisis automatizado de las mutaciones como biomarcadores, basado en las coincidencias encontradas en los alineamientos, que proporcionaron un valor añadido al proyecto. Sobre este último se basaron la gran mayoría de los resultados. Al realizar la confrontación, las bases teóricas inicialmente establecidas no se cumplían, atribuyéndolo a los datos tan heterogéneos conseguidos. A pesar de que el proceso realizado era el correcto, los análisis carecían de lógica alguna.

Resultados

En ellos se pudieron determinar las mutaciones o biomarcadores más y menos habituales dentro de los datos controles y patológicos disponibles. La comparación en la discusión de estos, mediante una u otra fuente de información, estuvo marcada por el sentido que presentaban la secuencias obtenidas y la disponibilidad de datos controles para los diferentes exones. De nuevo, consecuencias adicionales de las condiciones que presentaban los datos obtenidos en la etapa inicial. De igual forma, y en la medida de lo posible, se pudieron comprobar ciertos biomarcadores mediante la bibliografía y la bases de datos consultadas, estableciendo su posible naturaleza y frecuencia en cáncer de mama.

Dificultades

El proyecto enfrentó grandes dificultades debido a la heterogeneidad de los datos, tanto en forma como en número, lo cual complicó el análisis y redujo la veracidad y eficacia de los resultados. Además, las secuencias de datos presentan diferentes orientaciones (5'→ 3' o 3'→ 5'), lo que puede ser problemático, especialmente cuando se comparan variantes genéticas, como con *gnomAD*. A pesar de los avances, la recolección de datos biológicos sigue siendo compleja y, en muchos casos, privatizada, indicando que aún queda un largo camino por recorrer.

También en el desarrollo del script de `R`, surgieron diferentes complicaciones en la importación y descarga de los paquetes necesarios. Por ello y al creer que no podría aportar información de valor complementaria al proyecto, no se llegó a incluir.

Lineas de trabajo futuras

Las mejoras propuestas, o aspectos a implementar en futuras líneas de trabajo, son las siguientes:

- La primera faceta a mejorar sería el proceso de obtención de los datos. Establecido como uno de los puntos débiles del trabajo, los datos, han supuesto una bajada más que justificada de la fiabilidad de los resultados conseguidos, además de limitar un análisis estandarizado para todos los genes y regiones codificantes relacionados con el cáncer de mama. Una obtención adecuada en cantidad y calidad de los datos iniciales, podrían suponer un lavado de cara notorio en el proyecto.

- El empleo de otras estructuras para su análisis a diferentes niveles, como es el caso de ARN y aminoácidos, que pueden ser proporcionadas a partir de los mismo datos, aportándonos una mayor información y siendo viable, destacándose como una de las mejores mejoras que podrían tener lugar. El estudio del material genético únicamente desde una perspectiva y en una etapa tan inicial como es el ADN, dada la complejidad del tema a tratar, hace que el proyecto pierda cierta fiabilidad y significado válido. Por tanto, la adición de diversos estudios a otros niveles estructurales, serviría como una buena alternativa de continuidad para el proyecto.

- Como alternativa complementaria de la anterior propuesta serviría la aplicación de nuevas herramientas bioinformáticas en auge como `MSAbrowser` [Bilgin et al., 2024], `DNABert` [Ji et al., 2023] y `DeepVariant` [Chang et al., 2023, Poplin et al., 2018]; o incluso ya establecidas como lo es el conocido `AlphaFold` [AlphaFold, 2024]. Este último sería

de gran interés para comprobar la estructura presenta tras las diferentes mutaciones oportunas que tuvieran lugar en las secuencias, ya que mantiene una gran relación con la funcionalidad. También una evaluación del rendimiento de la aplicación de varios de ellos que consigan un mismo objetivo sería interesante (y valorar de paso el rendimiento de cada uno).

- Este proyecto se ha centrado en el estudio de las regiones codificantes, pero el resto de regiones que representan la gran mayoría de la de información genómica, tienen la misma importancia para el control y regulación génica. Las regiones intrónicas o no codificantes presentan una mayor tolerancia a las mutaciones, sin embargo, sufren un número mayor de ellas. Cabe destacar que el carácter funcional de una secuencia comienza en estas extensiones. La información y veracidad que pueden aportar al proyecto, la convierten en una mejora casi obligatoria a futuro.

- Como alternativa a la estrategia actualmente propuesta, se podría explorar el estudio de otros genes comúnmente asociados al cáncer de mama o relacionados con la regulación génica. Esta aproximación podría enriquecer el proyecto con nuevos y distintos biomarcadores identificados en las secuencias analizadas.

- Otra opción, sería la realización de un estudio de minería de datos, junto con la aplicación de diferentes estrategias de aprendizaje automático, para la determinación mediante clasificación de un diagnóstico preciso de cáncer de mama. Resultaría ser una opción muy viable y más eficiente incluso que la actualmente elaborada, planteando desde un enfoque distinto el problema a abordar. Además de ser factible de implementar gracias a las herramientas y conocimientos disponibles.

- Por último, se propone la implementación de una App web o Api, que a partir de los avances logrados hasta el momento y a través de una plataforma de desarrollo de web o framework, se logre identificar el sentido clínico de unas muestras pasadas. Pudiendo vincular otras plataformas web que contengan sus respectivos métodos para ello, con la propia desarrollada. Un framework de ejemplo podría ser `StreamLit`, con el que se han llegado a diseñar modelos parecidos.

Bibliografía

[AECC, 2023] AECC (2023). Home | AECC Observatorio — observatorio.contraelcancer.es. https://observatorio.contraelcancer.es/. [Accessed 06-07-2024].

[AlphaFold, 2024] AlphaFold (2024). AlphaFold Protein Structure Database — alphafold.ebi.ac.uk. https://alphafold.ebi.ac.uk/. [Accessed 06-07-2024].

[American Cancer Society, 2020] American Cancer Society (2020). ¿qué es el cáncer? [Internet; recuperado 01-abril-2024].

[American Cancer Society, 2021] American Cancer Society (2021). Estado del receptor hormonal del cáncer de seno receptor de estrógeno positivo | Receptor de estrógeno — cancer.org. https://www.cancer.org/es/cancer/tipos/cancer-de-seno/comprension-de-un-diagnostico-de-cancer-de-seno/estado-del-receptor-hormonal-del-cancer-de-seno.html. [Accessed 06-07-2024].

[American Cancer Society, 2022a] American Cancer Society (2022a). Estatus HER2 del cáncer de seno | ¿Qué es el estatus HER2? — cancer.org. https://www.cancer.org/es/cancer/tipos/cancer-de-seno/comprension-de-un-diagnostico-de-cancer-de-seno/estado-de-her2-del-cancer-de-seno.html. [Accessed 06-07-2024].

[American Cancer Society, 2022b] American Cancer Society (2022b). Estatus HER2 del cáncer de seno | ¿Qué es el estatus HER2? — cancer.org. https://www.cancer.org/es/cancer/tipos/cancer-de-seno/comprension-de-un-diagnostico-de-cancer-de-

seno/estado-de-her2-del-cancer-de-seno.html. [Accessed 06-07-2024].

[American Cancer Society, 2022c] American Cancer Society (2022c). Oncogenes, genes supresores de tumores y genes reparadores del adn. [Internet; recuperado 06-abril-2024].

[American Cancer Society, 2022] American Cancer Society (2022). ¿Qué es una resonancia magnética del seno? | Pruebas de detección para el cáncer de seno — cancer.org. https://www.cancer.org/es/cancer/tipos/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/imagenes-por-resonancia-magnetica-de-los-senos.html. [Accessed 06-07-2024].

[Arafeh and Samuels, 2019] Arafeh, R. and Samuels, Y. (2019). Pik3ca in cancer: The past 30 years. *Seminars in Cancer Biology*, 59:36–49. PI3K/AKT signaling in human cancer and New insights in melanoma biology: running fast towards precision medicine.

[Benítez et al., 2024] Benítez, J., González Neira, A., Malats, N., Osorio, A., Robledo, M., Rodríguez, S., and Urioste, M. (2024). Fundación Instituto Roche - Glosario de genética — institutoroche.es. https://www.institutoroche.es/recursos/glosario. [Accessed 06-07-2024].

[Bertram, 2000] Bertram, J. S. (2000). The molecular biology of cancer. *Molecular Aspects of Medicine*, 21(6):167–223.

[Bilgin et al., 2024] Bilgin, H. I., Torun, F. M., and Kaplan, O. I. (2024). MSABrowser — thekaplanlab.github.io. https://thekaplanlab.github.io/. [Accessed 06-07-2024].

[Blanca and Cañizares, 2010] Blanca, J. and Cañizares, J. (2010). Alineamiento de secuencias 2014; Bioinformatics at COMAV 0.1 documentation — bioinf.comav.upv.es. https://bioinf.comav.upv.es/courses/intro_bioinf/alineamientos.html. [Accessed 06-07-2024].

[Børresen-Dale, 2003] Børresen-Dale, A.-L. (2003). Tp53 and breast cancer. *Human mutation*, 21(3):292–300.

[Bray et al., 2024] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263.

[Chang et al., 2024] Chang, J., Chapman, B., Friedberg, I., Hamelryck, T., de Hoon, M., Cock, P., Antao, T., Talevich, E., and Wilczyński, B. (2024). Biopython Tutorial and Cookbook — biopython.org. https://biopython.org/DIST/docs/tutorial/Tutorial. [Accessed 06-07-2024].

[Chang et al., 2023] Chang, P.-C., akolesnikov, Nattestad, M., McLean, C., sgoe1y, and Cook, D. E. (2023). GitHub - google/deepvariant: DeepVariant is an analysis pipeline that uses a deep neural network to call genetic variants from next-generation DNA sequencing data. — github.com. https://github.com/google/deepvariant. [Accessed 06-07-2024].

[Chen et al., 2022] Chen, Q., Wang, Y., Liu, Y., and Xi, B. (2022). Esrrg, atp4a, and atp4b as diagnostic biomarkers for gastric cancer: A bioinformatic analysis based on machine learning. *Frontiers in Physiology*, 13:905523.

[Cosmic, 2024a] Cosmic (2024a). BRCA1_ENST00000357654 Gene - COSMIC — cancer.sanger.ac.uk. https://cancer.sanger.ac.uk/cosmic/gene/analysis?all_data=&coords=AA%3AAA&dr=&end=1864&gd=&id=279822&ln=BRCA1_ENST00000357654&seqlen=1864&sn=breast&start=1#ts. [Accessed 06-07-2024].

[Cosmic, 2024b] Cosmic (2024b). BRCA2_ENST00000544455 Gene - COSMIC — cancer.sanger.ac.uk. https://cancer.sanger.ac.uk/cosmic/gene/analysis?all_data=&coords=AA%3AAA&dr=&end=3419&gd=&id=330603&ln=BRCA2_ENST00000544455&seqlen=3419&sn=breast&start=1#ts. [Accessed 06-07-2024].

[Cosmic, 2024c] Cosmic (2024c). PIK3CA Gene - COSMIC — cancer.sanger.ac.uk. https://cancer.sanger.ac.uk/cosmic/gene/analysis?all_data=&coords=AA%3AAA&dr=&end=1069&gd=&id=276592&ln=PIK3CA&seqlen=1069&sn=breast&start=1#ts. [Accessed 06-07-2024].

[Cosmic, 2024d] Cosmic (2024d). TP53 Gene - COSMIC — cancer.sanger.ac.uk. https://cancer.sanger.ac.uk/cosmic/gene/analysis?all_data=&coords=AA%3AAA&dr=&end=394&gd=&id=348585&ln=TP53&seqlen=394&sn=breast&start=1#ts. [Accessed 06-07-2024].

[de Ven, 2024] de Ven, B. V. (2024). Bokeh — bokeh.org. https://bokeh.org/. [Accessed 06-07-2024].

[Dineen, 2016] Dineen, D. (2016). Clustal Omega - fast, accurate, scalable multiple sequence alignment for proteins — clustal.org. http://www.clustal.org/omega/. [Accessed 06-07-2024].

[Edgar, 2021] Edgar, R. C. (2021). Muscle v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. *bioRxiv*.

[Foulkes and Shuen, 2013] Foulkes, W. D. and Shuen, A. Y. (2013). In brief: Brca1 and brca2. *The Journal of pathology*, 230(4):347–349.

[Fu et al., 2022] Fu, X., Tan, W., Son, Q., Pei, H., and Li, J. (2022). Frontiers | BRCA1 and Breast Cancer: Molecular Mechanisms and Therapeutic Strategies — doi.org. https://doi.org/10.3389/fcell.2022.813457. [Accessed 06-07-2024].

[Fundación Grupo Español de Investigación en Cáncer de Mama, 2024] Fundación Grupo Español de Investigación en Cáncer de Mama (2024). El cáncer de mama en españa: situación actual. [Internet; recuperado 01-abril-2024].

[Gagniuc, 2021] Gagniuc, P. A. (2021). *Algorithms in bioinformatics: Theory and implementation*. John Wiley and Sons.

[gnomAD Steering Committee, 2024a] gnomAD Steering Committee (2024a). gnomAD — gnomad.broadinstitute.org. https://gnomad.broadinstitute.org/. [Accessed 06-07-2024].

[gnomAD Steering Committee, 2024b] gnomAD Steering Committee (2024b). gnomAD — gnomad.broadinstitute.org. https://gnomad.broadinstitute.org/gene/ENSG00000139618?dataset=gnomad_r4. [Accessed 06-07-2024].

[HGNC, 2007] HGNC (2007). Gene symbol report | HUGO Gene Nomenclature Committee — genenames.org. https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:2927. [Accessed 06-07-2024].

[Ji et al., 2023] Ji, Y., Zhou, Z., Lau, T., bioRoastery, and Kalia, R. (2023). GitHub - jerryji1993/DNABERT: DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome — github.com. https://github.com/jerryji1993/DNABERT. [Accessed 06-07-2024].

[Jones and Pevzner, 2004] Jones, N. C. and Pevzner, P. A. (2004). *An introduction to bioinformatics algorithms*. MIT press.

[Jorde et al., 2020] Jorde, L. B., Carey, J. C., and Bamshad, M. J. (2020). Genética Médica — books.google.es. https://books.google.es/books?hl=es&lr=&id=nh__DwAAQBAJ&oi=fnd&pg=PP1&dq=genetica&ots=v0WNO6eqFk&sig=xJDosDZD4nI7-GLTRCT5qzun8c8#v=onepage&q&f=false. [Accessed 06-07-2024].

[JSON, 2024] JSON (2024). JSON — json.org. https://www.json.org/json-en.html. [Accessed 06-07-2024].

[Jupyter Team, 2015] Jupyter Team (2015). Project Jupyter Documentation 2014; Jupyter Documentation 4.1.1 alpha documentation — docs.jupyter.org. https://docs.jupyter.org/en/latest/. [Accessed 06-07-2024].

[Kumari and Gupta, 2023] Kumari, U. and Gupta, S. (2023). Ngs and sequence analysis with biopython for prospective brain cancer therapeutic studies.

[Lim et al., 2012] Lim, E., Metzger-Filho, O., and Winer, E. P. (2012). The Natural History of Hormone Receptor-Positive Breast Cancer - ProQuest — proquest.com. https://www.proquest.com/openview/cac749309f31d7909027ef9a8e20968f/1?pq-origsite=gscholar&cbl=38461. [Accessed 06-07-2024].

[Lin et al., 2023] Lin, X.-Y., Guo, L., Lin, X., Wang, Y., and Zhang, G. (2023). Concomitant pik3ca and tp53 mutations in breast cancer: An analysis of clinicopathologic and mutational features, neoadjuvant therapeutic response, and prognosis. *Journal of Breast Cancer*, 26(4):363.

[López Marure, 2013] López Marure, R. (2013). LA REGULACIÓN DEL CICLO CELULAR Y EL CÁNCER. *Vertientes. Revista Especializada en Ciencias de la Salud*, 6(1).

[Margarit, 2008] Margarit, S. (2008). Cáncer hereditario de mama. *Revista chilena de radiología*, 14(3):135–141.

[Martín et al., 2015] Martín, M., Herrero, A., and Echavarría, I. (2015). El cáncer de mama. *Arbor*, 191(773):a234–a234.

[MedlinePlus, 2023] MedlinePlus (2023). Pruebas genéticas para BRCA1 y BRCA2: MedlinePlus enciclopedia médica — medlineplus.gov. https://medlineplus.gov/spanish/ency/patientinstructions/000690.htm. [Accessed 06-07-2024].

[MedlinePlus, 2024] MedlinePlus (2024). Prueba genética TP53 (proteína tumoral 53): Prueba de laboratorio de MedlinePlus — medlineplus.gov. https://medlineplus.gov/spanish/pruebas-de-laboratorio/prueba-genetica-tp53-proteina-tumoral-53/. [Accessed 06-07-2024].

[Mehrgou and Akouchekian, 2016] Mehrgou, A. and Akouchekian, M. (2016). The importance of BRCA1 and BRCA2 genes mutations in breast cancer development — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4972064/. [Accessed 06-07-2024].

[NCBI, 2023] NCBI (2023). NCBI Multiple Sequence Alignment Viewer 1.25.0 — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/projects/msaviewer/?appname=ncbi_msav&openuploaddialog. [Accessed 06-07-2024].

[NCBI, 2024a] NCBI (2024a). Analyze - NCBI — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/home/analyze/. [Accessed 06-07-2024].

[NCBI, 2024b] NCBI (2024b). National Center for Biotechnology Information — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/. [Accessed 06-07-2024].

[NCI and National Human Genome Research Institute, 2006] NCI and National Human Genome Research Institute (2006). The Cancer Genome Atlas Program (TCGA) — cancer.gov. https://www.cancer.gov/ccg/research/genome-sequencing/tcga. [Accessed 06-07-2024].

[NIH, 2023a] NIH (2023a). Exámenes de detección del cáncer de seno (mama) — cancer.gov. https://www.cancer.gov/espanol/tipos/seno/paciente/deteccion-seno-pdq. [Accessed 06-07-2024].

[NIH, 2023b] NIH (2023b). Mamografías — cancer.gov. https://www.cancer.gov/espanol/tipos/seno/hoja-informativa-mamografias#que-diferencia-hay-entre-las-mamografias-de-deteccion-y-las-de-diagnostico. [Accessed 06-07-2024].

[NIH, 2024a] NIH (2024a). BRCA1 BRCA1 DNA repair associated [Homo sapiens (human)] - Gene - NCBI — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/gene/672. [Accessed 06-07-2024].

[NIH, 2024b] NIH (2024b). BRCA2 BRCA2 DNA repair associated [Homo sapiens (human)] - Gene - NCBI — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/gene/675. [Accessed 06-07-2024].

[NIH, 2024c] NIH (2024c). Cancer Statistics — cancer.gov. https://www.cancer.gov/about-cancer/understanding/statistics#:~:text=Cancer%20is%20among%20the%20leading,related%20deaths%20to%2015.3%20million. [Accessed 06-07-2024].

[NIH, 2024d] NIH (2024d). Diccionario de cáncer del NCI — cancer.gov. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/gen-tp53. [Accessed 06-07-2024].

[NIH, 2024] NIH (2024). Diccionario de genética del NCI — cancer.gov. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-genetica/def/adn. [Accessed 06-07-2024].

[NIH, 2024] NIH (2024). Tipos de tratamiento — cancer.gov. https://www.cancer.gov/espanol/cancer/tratamiento/tipos. [Accessed 06-07-2024].

[Nik-Zainal and Morganella, 2017] Nik-Zainal, S. and Morganella, S. (2017). Mutational signatures in breast cancer: the problem at the DNA level. *Clinical Cancer Research*, 23(11):2617–2629.

[Oliver and Marín, 1996] Oliver, J. L. and Marín, A. (1996). A relationship between GC content and coding-sequence length. *Journal of molecular evolution*, 43(3):216–223.

[Poplin et al., 2018] Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., and DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987.

[Python Software Foundation, 2024a] Python Software Foundation (2024a). 5. Data Structures — docs.python.org. https://docs.python.org/3/tutorial/datastructures.html#data-structures. [Accessed 06-07-2024].

[Python Software Foundation, 2024b] Python Software Foundation (2024b). 5. El sistema de importación — docs.python.org. https://docs.python.org/es/3/reference/import.html#packages. [Accessed 06-07-2024].

[Python Software Foundation, 2024c] Python Software Foundation (2024c). The Python Tutorial — docs.python.org. https://docs.python.org/3/tutorial/index.html. [Accessed 06-07-2024].

[Python Software Foundation, 2024d] Python Software Foundation (2024d). Tipos integrados — docs.python.org. https://docs.python.org/es/3/library/stdtypes.html#functions. [Accessed 06-07-2024].

[R Foundation, 2024] R Foundation (2024). R: The R Project for Statistical Computing — r-project.org. https://www.r-project.org/. [Accessed 06-07-2024].

[RAE, 2024] RAE (2024). base | Diccionario de la lengua española — dle.rae.es. https://dle.rae.es/base#Ciiosq0. [Accessed 06-07-2024].

[Roche, 2024] Roche (2024). Prueba de mutación cobas® PIK3CA — diagnostics.roche.com. https://diagnostics.roche.com/es/es/products/params/cobas-pik3ca-mutation-test.html. [Accessed 06-07-2024].

[Révillion et al., 1998] Révillion, F., Bonneterre, J., and Peyrat, J. (1998). Erbb2 oncogene in human breast cancer and its clinical significance. *European Journal of Cancer*, 34(6):791–808.

[Siegel et al., 2024] Siegel, R. L., Giaquinto, A. N., and Jemal, A. (2024). Cancer statistics, 2024 - PubMed — pubmed.ncbi.nlm.nih.gov. https://pubmed.ncbi.nlm.nih.gov/38230766/. [Accessed 06-07-2024].

[Siegel et al., 2023] Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1).

[Sociedad Española de Oncología Médica, 2023] Sociedad Española de Oncología Médica (2023). Las cifras del cáncer en España 2023. [Internet; recuperado 01-abril-2024].

[Tafurt Cardona and Marin Morales, 2014] Tafurt Cardona, Y. and Marin Morales, M. A. (2014). PRINCIPALES MECANISMOS DE REPARACIÓN DE DAÑOS EN LA MOLÉCULA DE ADN — scielo.org.co. http://www.scielo.org.co/scielo.php?pid=S1657-95502014000200008&script=sci_arttext. [Accessed 06-07-2024].

[UCM, 2024] UCM (2024). LA MUTACIÓN. https://www.ucm.es/data/cont/media/www/pag-56185/11-La%20mutaci%C3%B3n.pdf. [Accessed 06-07-2024].

[UCSC, 2024a] UCSC, U. S. C. (2024a). Human Gene BRCA1 (ENST00000357654.9) from GENCODE V46 — genome.ucsc.edu. https://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=ENST00000357654.

9&hgg_chrom=chr17&hgg_start=43044294&hgg_end=43125364&hgg_type=knownGene&db=hg38. [Accessed 05-07-2024].

[UCSC, 2024b] UCSC, U. S. C. (2024b). Human Gene BRCA2 (ENST00000380152.8) from GENCODE V46 — genome.ucsc.edu. https://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=ENST00000380152.8&hgg_chrom=chr13&hgg_start=32315507&hgg_end=32400268&hgg_type=knownGene&db=hg38. [Accessed 06-07-2024].

[UCSC, 2024c] UCSC, U. S. C. (2024c). Human Gene PIK3CA (ENST00000263967.4) from GENCODE V46 — genome.ucsc.edu. https://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=ENST00000263967.4&hgg_chrom=chr3&hgg_start=179148356&hgg_end=179240093&hgg_type=knownGene&db=hg38. [Accessed 06-07-2024].

[UCSC, 2024d] UCSC, U. S. C. (2024d). UCSC Genome Browser Home — genome.ucsc.edu. https://genome.ucsc.edu. [Accessed 06-07-2024].

[Waks and Winer, 2019] Waks, A. G. and Winer, E. P. (2019). Breast Cancer Treatment in 2019 — jamanetwork.com. https://jamanetwork.com/journals/jama/article-abstract/2721183. [Accessed 06-07-2024].

[Watson and Crick, 1953] Watson, J. and Crick, F. (1953). Molecular Structure of Nucleic Acids | Learn Science at Scitable — nature.com. https://www.nature.com/scitable/content/Molecular-Structure-of-Nucleic-Acids-16331/. [Accessed 06-07-2024].

[World Health Organization, 2022] World Health Organization (2022). Cancer/datos y cifras. [Internet; recuperado 01-abril-2024].

[Xiao-Yi Lin et al., 2023] Xiao-Yi Lin, Lijuan Guo, Xin Lin, Yulei Wang, and Guochun Zhang (2023). Concomitant PIK3CA and TP53 Mutations in Breast Cancer: An Analysis of Clinicopathologic and Mutational Features, Neoadjuvant Therapeutic Response, and Prognosis — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10475711/. [Accessed 06-07-2024].

[Yaliana Tafurt Cardona and Maria Aparecida Marin Morales, 2014] Yaliana Tafurt Cardona and Maria Aparecida Marin Morales (2014). PRINCIPALES MECANISMOS DE REPARACIÓN DE DAÑOS EN LA MOLÉCULA DE ADN. http://vip.ucaldas.edu.co/biosalud/downloads/Biosalud13%282%29_8. [Accessed 08-07-2024].

[Zagami and Carey, 2022] Zagami, P. and Carey, L. A. (2022). Triple negative breast cancer: Pitfalls and progress - npj Breast Cancer — nature.com. https://www.nature.com/articles/s41523-022-00468-0. [Accessed 06-07-2024].

[Zeng et al., 2021] Zeng, X., Shi, G., He, Q., and Zhu, P. (2021). Screening and predicted value of potential biomarkers for breast cancer using bioinformatics analysis. *Scientific reports*, 11(1):20799.