UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Análisis, adaptación y comparación de métodos estadísticos de predicción de hotspots y aplicación en el ámbito de las cardiopatías familiares.

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: García Zarzoso, Alba

Tutor/a: Pastor López, Oscar

Cotutor/a: Tarazona Campos, Sonia

Director/a Experimental: Costa Sánchez, Mireia

CURSO ACADÉMICO: 2023/2024

Análisis, adaptación y comparación de métodos estadísticos de predicción de hotspots y aplicación en el ámbito de las cardiopatías familiares.

Máster en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones.

Alumna: Alba García Zarzoso
Tutor: Óscar Pastor López
Cotutora: Sonia Tarazona Campos
Directora experimental: Mireia Costa Sánchez
Departamento de Estadística e Investigación Operativa Aplicadas y Calidad
Universidad Politécnica de Valencia

14 de septiembre de 2024

Resumen

Resumen

Los hotspots son regiones del ADN susceptibles a presentar variaciones genéticas por distintas razones biológicas e identificarlas es crucial para la medicina de precisión y el diagnóstico precoz de las enfermedades. Sin embargo, estas regiones no se suelen reportar y, por tanto, su predicción sigue siendo un desafío. Por este motivo, el objetivo de este Trabajo Final de Máster es analizar, adaptar y comparar distintos métodos predictivos de hotspots basados en la estadística dentro del ámbito de las cardiopatías familiares. El estudio se divide en 5 etapas: (1) Revisión bibliográfica y selección de métodos, (2) Generación de los datos de entrada, (3) Análisis teórico de los métodos seleccionados, (4) Adaptación de los datos de entrada a cada método e (5) Implementación de métodos. Los resultados se comparan con una base de datos de referencia en este dominio llamada CardioHotspots mediante métricas de evaluación. El método con mejor desempeño es el asociado al algoritmo Smith Waterman y muestra una sensibilidad del 33,33 %, una especificidad del 92,22 % y una exactitud del 68,71 %. Estos resultados revelan que todavía queda mucho margen de mejora dentro de este área de estudio y resalta la necesidad de continuar investigando estos métodos predictivos de hotspots.

Palabras clave: hotspots; predicción; métodos estadísticos; genética.

Resum

Els hotspots són regions de l'ADN susceptibles de presentar variacions genètiques per raons biològiques diferents i identificar-les és crucial per a la medicina de precisió i el diagnòstic precoç de les malalties. Tot i això, aquestes regions no se solen reportar i, per tant, la seva predicció continua sent un desafiament. Per això, l'objectiu d'aquest Treball Final de Màster és analitzar, adaptar i comparar diferents mètodes predictius de hotspots basats en l'estadística dins l'àmbit de les cardiopaties familiars. L'estudi es divideix en 5 etapes: (1) Revisió bibliogràfica i selecció de mètodes, (2) Generació de les dades d'entrada, (3) Anàlisi teòric dels mètodes seleccionats, (4) Adaptació de les dades d'entrada a cada mètode i (5) Implementació de mètodes. Els resultats es comparen amb una base de dades de referència en aquest domini anomenada CardioHotspots mitjançant mètriques d'avaluació. El millor mètode és l'associat a l'algoritme Smith Waterman i mostra una sensibilitat del 33,33 %, una especificitat del 92,22 % i una exactitud del 68,71 %. Aquests resultats revelen que encara queda molt de marge de millora dins d'aquesta àrea d'estudi i ressalta la necessitat de continuar investigant aquests mètodes predictius de hotspots.
Paraules clau: hotspots; predicció; mètodes estadístics; genètica.

Abstract

Hotspots are DNA regions susceptible to genetic variations for different biological reasons and identifying them is crucial for precision medicine and early diagnosis of diseases. However, these regions are not often reported and, therefore, their prediction remains a challenge. For this reason, the aim of this Master's Thesis is to analyze, adapt and compare different statistical-based hotspot prediction methods in the field of familial heart disease. The study is divided into 5 stages: (1) Literature review and method selection, (2) Generation of input data, (3) Theoretical analysis of the selected methods, (4) Adaptation of input data to each method and (5) Implementation of methods. The results are compared with a reference database in this domain called CardioHotspots using evaluation metrics. The best performing method is the one associated with the Smith Waterman algorithm, showing a sensitivity of 33.33 %, a specificity of 92.22 % and an accuracy of 68.71 %. These results reveal that there is still much room for improvement within this area of study and highlight the need to continue researching these hotspot predictive methods.

Keywords: hotspots; prediction; statistical methods ; genetics.

Agradecimientos

En primer lugar, me gustaría dedicarle este apartado de agradecimientos a Mireia Costa Sánchez. Gracias por confiar siempre en mí y en mi potencial de investigadora a lo largo de estos 3 años de trabajo conjunto. Aún recuerdo cuando entré en el grupo por primera vez en tercero de carrera y nos recibiste con los brazos abiertos a Marina y a mí. Todo lo que sé sobre investigación es gracias a los innumerables power points, informes y reuniones que hemos tenido a lo largo de todo este tiempo. Gracias por apostar por mí y siempre darme las facilidades que he necesitado para seguir aprendiendo. Asimismo me gustaría agradecer la contribución de mis tutores Óscar y Sonia. Gracias por estar siempre a mi disposición para resolver dudas y ayudarme en todo lo necesario para que este trabajo saliera adelante.

En segundo lugar, me gustaría agradecer al grupo PROS del instituto VRAIN en la UPV. A lo largo de mis estancias como colaboradora he aprendido de primera mano como funciona el mundo de la investigación y he tenido el placer de compartirlo con profesionales que, a parte de tener una alta calidad investigadora, también destacaba la calidad humana.

En tercer lugar, me gustaría agradecer a la fundación valgrAI. Esta fundación me ha dado la oportunidad de realizar el Máster de Ingeniería en Análisis de Datos, Mejora de Procesos y Toma de Decisiones cuyo final se enmarca en este trabajo. Espero que esta fundación siga dando las oportunidades a los estudiantes que me dieron a mi para poder contribuir al desarrollo del mundo de la inteligencia artificial y el análisis de datos.

Quería aprovechar este apartado para dedicar este trabajo a toda la gente que ha sido parte de este año. Mamá, papá, tete, gracias por esperarme a cenar todos los días cuando salía a las 9, por apoyarme en mis momentos bajos y enorgulleceros en los momentos altos. Gracias por siempre confiar en mí y dejarme elaborar mi camino académico y formativo, siempre apoyándome desde casa. También me gustaría hacer una mención especial a mis compañeras de clase. Gracias por hacer las tardes de estudio el mejor plan, siempre con las risas por delante y nunca dejando de lado las meriendas en el Trinquet. Finalmente, me gustaría agradecer este trabajo a Aroa. Este año has sido mi pilar fundamental y mi complemento en todas mis versiones. Gracias por escuchar los avances de mi investigación, aunque no entiendas muy bien lo que hago, así como ser el hombro donde apoyarme cuando creía que no iba a conseguirlo. Gracias por siempre confiar más en mí de lo que yo lo hago y por celebrar todos mis logros con incluso más entusiasmo que yo. Este año siempre lo voy a recordar contigo a mi lado.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Antecedentes

El origen de este trabajo se basa en la realización de mi Trabajo Final de Grado titulado "Diseño y desarrollo de una fuente de datos sobre hotspots asociados al criterio PM1 de las guías ACMG – AMP 2015 aplicado a cardiopatías familiares". Este proyecto fue fruto de mi investigación en el grupo PROS del instituto VRAIN, el cual participó activamente en los proyectos OGMIOS y CARDIOVAL. El objetivo principal de estos proyectos fue la clasificación de las variaciones genéticas dentro del área clínica de las cardiopatías.

Ambos proyectos tienen en común su interés en mejorar la clasificación de las variaciones genéticas relacionadas con las cardiopatías familiares siguiendo los criterios de las guías ACMG – AMP del 2015 (American College of Medical Genetics and Genomics - Association for Molecular Pathology). Estas guías son las más usadas para clasificación de variaciones genéticas en 5 tipos según su potencial actuación con respecto a una enfermedad: (1) patogénica, (2) probablemente patogénica, (3) significado incierto, (4) probablemente benigna y (5) benigna [Richards et al., 2015]. Asimismo, también se clasifican en cuatro categorías en función del nivel de evidencia que se proporcione para dicha clasificación: (1) muy fuerte, (2) fuerte, (3) moderado y (4) de soporte [Richards et al., 2015].

Durante la realización de mi Trabajo Final de Grado, me centré en la caracterización del concepto de hotspot - regiones del genoma susceptibles a la mutabilidad - así como en el desarrollo de una fuente de datos de los mismos. Esta investigación surge como consecuencia de la asignación del criterio PM1 - evidencia moderada de patogenicidad - que las guías ACMG - AMP del 2015 les atribuyen a estas regiones. Estas guías son fundamentales para la clasificación de las variaciones y, durante el desarrollo de mi Trabajo Final de Grado, no existían fuentes de datos de hotspots.

Por este motivo, la realización del proyecto anterior se basó en la caracterización del concepto de hotspot mediante el modelado conceptual y el desarrollo de la fuente de datos de hotspots encontrados en la literatura. Además, se desarrolló junto con otros miembros del grupo PROS una interfaz de usuario en la que poder consultar esta fuente. El resultado de este proyecto es CardioHotspots (`https://genomics-hub.pros.dsic.upv.es:3099/`) [S.Garcia et al., 2024], una base de hotspots mutacionales curada a mano y publicada en la revista Database - The Journal of Biological Databases and Curation (Q1 en la categoría *MATHEMATICAL & COMPUTATIONAL BIOLOGY*), que actualmente sigue evolucionando dado que el proyecto sigue su curso.

1.2. Problemática y motivación

Una vez acabado el Trabajo Final de Grado, el grupo de investigación continuó estudiando este tema y se planteó la siguiente pregunta: ¿Existen métodos predictivos de hotspots en algún dominio clínico y pueden adaptarse a las cardiopatías familiares para obtener resultados fidedignos? Para responderla se planteó una investigación complementaria a los intereses del grupo que se desarrolla durante este Trabajo Final de Máster.

Este proyecto tiene como principal objetivo el análisis, adaptación y comparación de métodos estadísticos de predicción de hotspots dentro del caso de uso de las cardiopatías familiares y, por ese motivo, es vital conocer los problemas asociados a este cometido. Los dos principales problemas son: (1) la falta de información de hotspots y (2) la escasez de métodos predictivos de este concepto en esta área clínica.

En primer lugar, solo existe un repositorio que proporcione información precisa sobre hotspots de cardiopatías familiares. Sin embargo, esta fuente contempla únicamente los hotspots encontrados en la literatura y, por tanto, no incluye predicciones derivadas de modelos que identifiquen potenciales nuevos hotspots. Esto implica que la información disponible se limita a regiones previamente analizadas. Por tanto, esta falta de reporte de hotspots provoca una carencia dentro de este área de estudio. Por este motivo, actualmente la comunidad científica se centra en la necesidad de implementar herramientas bioinformáticas que sean capaces de generar nuevas regiones.

La generación e implementación de estas herramientas bioinformáticas para la predicción de conceptos genéticos son esenciales para el avance de la biomedicina moderna. Dentro de la amplia gama de conceptos genéticos a predecir se encuentran los hotspots, unas regiones de alto interés por su elevada mutabilidad y, por tanto, potenciales lugares de hospedaje de futuras enfermedades genéticas. Esta carencia resalta la necesidad de analizar las herramientas creadas para la predicción de hotspots.

Pese a la existencia de herramientas bioinformáticas para la predicción de hotspots, estas son escasas y se desarrollan en marcos de estudio muy específicos. La mayoría de ellas se centran en las enfermedades oncológicas quedando así reflejada la falta de existencia de estos métodos en otras enfermedades. Este hecho deja patente la necesidad de adaptar los métodos predictivos

existentes para el ámbito de las cardiopatías familiares.

Las herramientas de predicción son fundamentales en la medicina personalizada y más concretamente en el análisis de enfermedades genéticas complejas. Este tipo de herramientas predictivas son capaces de analizar grandes cantidades de datos genómicos para la identificación de regiones o variaciones genéticas, así como para la detección de hotspots. Estas herramientas se caracterizan por su gran capacidad de integración y procesado de información para poder aplicar modelos estadísticos que permitan una mejor predicción de los conceptos genéticos de interés. Esto no solo mejora la comprensión de las enfermedades, sino que también facilita la toma de decisiones clínicas, favoreciendo la generación de estrategias preventivas y tratamientos personalizados.

Sin embargo, estos métodos predictivos pueden dar lugar a errores de predicción induciendo a la identificación incorrecta de algunas regiones. Por esta razón se precisa comparar los resultados que da el modelo con los de una referencia y así poder estudiar el rendimiento y la capacidad predictiva de los mismos.

1.3. Alineación con los contenidos del máster

Este trabajo no podría haber sido realizado sin los conocimientos adquiridos durante el Máster de Ingeniería de Análisis de Datos, Toma de Decisiones y Mejora de Procesos, el cual me ha dotado de los conocimientos suficientes para la interpretación de los modelos, la adaptación de estos y su posterior análisis. Todo este trabajo se ha realizado empleando R como lenguaje de programación dado que es uno de los lenguajes referentes dentro del mundo de la estadística y la bioinformática. Además, durante el desarrollo del máster, se dedica una asignatura entera a comprender los fundamentos de este lenguaje y a ser capaces de desarrollar código propio en este lenguaje.

Centrándonos más en los contenidos teóricos del máster, todo lo aprendido en las asignaturas de Modelos de Regresión Lineal y ANOVA, Diseño de Experimentos I y Análisis Multivariante, me han capacitado para comprender el fundamento teórico de los modelos que se implementan en el mismo. Conocer las bases y el fundamento de los modelos es crucial para poder tener una visión holística y no conformarse con interpretar los resultados. Más concretamente, estas asignaturas han sido de especial utilidad a la hora de comprender las técnicas de inferencia estadística empleadas en los modelos, así como los diferentes enfoques que se aplican.

Asimismo, para el desarrollo de este proyecto se han empleado todas las técnicas de preprocesado de datos adquiridas de forma autónoma a través de los trabajos académicos realizados en las asignaturas de Minería de Datos, Análisis Multivariante y Programación en R para el Análisis de Datos. A lo largo de todo el máster nos hemos enfrentado a bases de datos reales y lo que supone prepararlas para poder analizarlas y gracias a trabajos que comenzaban desde cero, he adquirido las habilidades y las técnicas oportunas para poder preparar una base de datos completa para poder emplearla como entrada para los métodos predictivos.

1.4. Estructura del trabajo

Este Trabajo de Final de Máster se organiza en 8 capítulos:

- **Capítulo 1. Introducción:** En él se introduce el trabajo y los antecedentes del mismo, la motivación y problemática del trabajo, la alineación con los contenidos del máster y la estructura del trabajo.
- **Capítulo 2. Objetivos:** En él se especifican los objetivos y las preguntas de investigación asociadas a cada objetivo.
- **Capítulo 3. Base metodológica:** En él se detalla la metodología Design Science empleada para desarrollar este trabajo.
- **Capítulo 4. Análisis del problema de investigación:** En él se presenta el marco contextual del trabajo, definiendo conceptos clave para el entendimiento del mismo, así como la presentación del problema de investigación.
- **Capítulo 5. Diseño de investigación e inferencia:** En este se explica el diseño de investigación planteado para poder llevar a cabo este proyecto.
- **Capítulo 6. Ejecución de la investigación:** En él se exponen los problemas encontrados a la hora de la implementación de los métodos, las soluciones y los resultados de los mismos.
- **Capítulo 7. Análisis de datos:** En él se expone el análisis de los resultados de la implementación de los métodos predictivos en comparación con la fuente de datos de CardioHotspots.
- **Capítulo 8. Conclusiones:** En él se resuelven las preguntas de investigación para lograr los objetivos redactados en el capítulo 2.

Capítulo 2

Objetivos

El objetivo del Trabajo Final de Máster es analizar, adaptar y comparar distintos métodos predictivos de hotspots en el área clínica de las cardiopatías familiares. Para poder lograr este objetivo global se plantean a continuación una serie de objetivos específicos que se resolverán mediante preguntas de investigación.

1. Análisis de investigación del problema
    1.1. ¿Qué es la genética?
    1.2. ¿Qué es un hotspot?
    1.3. ¿Qué fuentes de datos existen sobre hotspots relacionados con cardiopatías familiares?
2. Diseño de investigación e inferencia
    2.1. ¿Cómo se identifican los métodos predictivos existentes?
    2.2. ¿Cómo se seleccionan los métodos más importantes?
    2.3. ¿Qué fundamentos teóricos sustentan los métodos?
    2.4. ¿Qué datos existen sobre variaciones genéticas de cardiopatías familiares?
    2.5. ¿Cómo se generaliza la entrada de datos a cada modelo?
3. Ejecución de la investigación
    3.1. ¿Cuáles son los resultados?
4. Análisis de datos
    4.1. ¿Qué enfoque se emplea para la comparativa?
    4.2. ¿Qué parámetros son clave para la comparativa?
    4.3. ¿Qué método ofrece mejores parámetros de evaluación con respecto a la referencia?

Capítulo 3

Base metodológica

3.1. Design science

Este trabajo se basa en una metodología conocida como Design Science (DS) propuesta por Wieringa en 2014 y esta se define como el diseño e investigación de artefactos en un contexto específico [Wieringa, 2014]. Los artefactos que se pretenden diseñar vienen acompañados de un entorno que los contextualiza ya que estos por sí mismos no son capaces de resolver un problema [Wieringa, 2014]. Para resolver un problema de DS se debe acudir a la interacción entre el artefacto y el contexto del problema. El artefacto de este Trabajo de Final de Máster consiste en analizar, adaptar y comparar los métodos estadísticos de predicción de hotspots con la fuente de datos de referencia llamada CardioHotspots [S.Garcia et al., 2024].

En la metodología del Design Science se clasifican los problemas de dos formas: problemas prácticos y preguntas de conocimiento. En el primer tipo de problema, se busca aplicar el DS para producir un cambio en el mundo real, de modo que el diseño de la solución propuesta tenga presente las necesidades de todos los usuarios o *stakeholders*. En otras palabras, este tipo de problemas pretenden desarrollar nuevas soluciones a problemas reales de forma empírica. Cabe añadir que la metodología empleada para resolver este tipo de problemas es el conocido Ciclo de Diseño.

En el segundo tipo de problema, las preguntas de conocimiento, el Design Science se centra en un marco contextual, donde el principal objetivo es la obtención de más conocimiento acerca del mundo, sin la necesidad de proponer una solución al problema. Es decir, estos pretenden resolver una pregunta de investigación de un dominio concreto con el conocimiento existente del mismo. Cabe añadir que la metodología empleada para resolver este tipo de problemas es el conocido Ciclo Empírico.

En la Tabla 3.1 se observa la heurística empleada para distinguir los problemas de investigación de las preguntas de conocimiento.

Problemas de diseño	Preguntas de conocimiento
Llamada al cambio del mundo	Pregunta para el conocimiento del mundo
El diseño es la solución	La respuesta es una proposición
Muchas soluciones	Una solución
Evaluado por utilidad	Evaluado por la verdad
Utilidad depende de los objetivos	Verdad no dependiente de los objetivos

Tabla 3.1: *Heurística para distinguir los problemas de diseño con los problemas de conocimiento [Wieringa, 2014].*

Tras conocer los tipos de problemas que abarca el DS y los objetivos de este proyecto, la metodología que se va a seguir es la relacionada con las preguntas de conocimiento. Esto es debido a que, siguiendo la Tabla 3.1 , se defina únicamente una pregunta de investigación y cuya solución es única. Asimismo, la verdad no es dependiente de los objetivos ya que es posible que la respuesta a la pregunta sea una negativa, negando la posibilidad de realizar esta adaptación.

3.2. Adaptación del Ciclo Empírico

El ciclo empírico es un proceso racional que responde a los problemas de conocimiento que consta de cinco fases. En este proyecto se va a realizar una adaptación del ciclo empírico inspirado por Wieringa planteando este a través de cuatro fases. En la Figura 3.1 se puede observar la adaptación del ciclo empírico.

Figura 3.1: *Adaptación del cíclo empírico [Elaboración propia].*

La definición de cada una de estas cuatro fases y sus tareas principales se exponen a continuación:

1. **Análisis del problema de investigación:** Contextualización del problema de investigación para conseguir una visión global y un entendimiento total del mismo. Para lograrlo se presenta la definición de conceptos biológicos relevantes y la presentación de fuentes de datos existentes de hotspots dentro del ámbito de las cardiopatías familiares.
2. **Diseño de la investigación y de inferencia:** Especificación de todos los requerimientos de diseño para la resolución de las preguntas de investigación y planteamiento de una estrategia de resolución del proyecto.
3. **Ejecución de la investigación:** Realización de ese diseño de la investigación a través de la implementación de los métodos, así como la presentación de los problemas surgidos durante el desarrollo y la solución empleada. La estructura interna del mismo se corresponde de forma fidedigna al diseño de investigación propuesto en la fase anterior.
4. **Análisis de los resultados:** Análisis de los resultados a través de la comparación de los resultados obtenidos con los de referencia presentados en el análisis del problema de investigación.

Capítulo 4

Análisis del problema de investigación

Conocer el problema a la hora de empezar una investigación es fundamental para establecer unas bases conceptuales sólidas. Por este motivo, este capítulo se divide en dos grandes secciones: marco contextual y CardioHotspots. En la primera sección se definen algunos conceptos genéticos relevantes para la comprensión del problema y en la segunda sección se describe CardioHotspots, una base de datos de hotspots desarrollada en colaboración con el grupo PROS del instituto VRAIN de la Universidad Politécnica de Valencia.

4.1. Marco contextual

Esta sección se centra en la definición de los conceptos de medicina de precisión, genética, variación genética y hotspots. Estas definiciones resultan cruciales para entender el problema de investigación y también para tener una visión un poco más completa del área de estudio en la que se ve envuelto este trabajo.

4.1.1. Medicina de precisión y genética

La **medicina de precisión** es un nuevo enfoque terapéutico que pretende encontrar la personalización del tratamiento y la prevención de las enfermedades estudiando las características de cada paciente [Ackerman et al., 2013]. Este enfoque se basa en la idea de que dos personas con la misma enfermedad pueden responder de forma distinta a un tratamiento. Esto se debe a sus diferencias genéticas y para conocerlas es necesario realizar pruebas que revelen la secuencia de cada paciente. Gracias al estudio de esas secuencias se puede estudiar las variaciones genéticas que presenta cada individuo y determinar si son patogénicas o no.

Otra de las acepciones más empleadas para definir medicina de precisión es que esta medicina se basa en adaptar el tratamiento y la prevención de las enfermedades considerando las diferencias

en factores genéticos, ambientales o incluso de estilo de vida, específico de grupos de personas [Hurtado, 2022].

Debido a la propia definición de medicina de precisión, la genética juega un papel fundamental en esta disciplina, ya que cualquier variación genética puede suponer, entre otras cosas, la predisposición de padecer ciertas enfermedades. Según el Instituto Nacional de Investigación del Genoma Humano, se define como la rama de la biología ocupada de la herencia, incluyendo la interacción entre genes, las variaciones del ADN[1] y sus interacciones con otros factores ambientales [Genética, NHGRI, 2024].

Por todos estos motivos, la genética es uno de los principales pilares del desarrollo de tratamientos médicos avanzados. El motivo de su uso reside en que algunas variaciones genéticas influyen en la predisposición de padecer ciertas enfermedades y, por tanto, conocerlas es vital para detectarlas, estudiarlas y poder tratar las enfermedades derivadas de esas variaciones. Esta disciplina consituye un paso más para la transformación de la medicina convencional a la medicina de precisión.

4.1.2. Variación genética

A raíz de lo comentado anteriormente se presenta la definición de variación genética. Una variación genética se define como el cambio de secuencia del ADN de un individuo con respecto a una secuencia de referencia [EBI, 2024]. Estas variaciones nos permiten diferenciarnos entre los individuos de la misma especie y algunas de ellas son responsables de algunas enfermedades complejas como el cáncer o trastornos cardiovasculares. Es por ello por lo que identificarlas y comprenderlas resulta fundamental, no solo para el diagnótico preciso sino que también para su prevención. En este contexto, la medicina de precisión aparece adaptando los tratamientos y medidas preventivas en función del perfil genético de cada paciente. Al ser capaces de saber la predisposición genética a padecer ciertas enfermedades, se pueden implementar intervenciones preventivas para así mejorar la calidad de vida y reducir el riesgo de futuras complicaciones.

Por este motivo, el estudio de estas variaciones genéticas no solo es clave para el diagnóstico de enfermedades, sino que también es un pilar fundamental para la evolución de la medicina personalizada. Este estudio de variaciones genéticas supone la transformación del enfoque tradicional, el cual se basa en la implementación del mismo tratamiento para todos los pacientes, hacia un enfoque dirigido a satisfacer las necesidades específicas de cada paciente.

Para poder identificar esas alteraciones, es de vital importancia conocer las secuencisa de referncia en las cuales se respaldan. Estas secuencias atienden a individuos sanos y, por tanto, al compararlas con las del paciente, se pueden detectar las diferencias que se observan. Es importante recalcar que no todas las variaciones son patogénicas ni todas tienen por qué estar ligadas

[1] **ADN o ácido desoxirribonucléico:** Molécula que transporta información genética para el desarrollo y funcionamiento de un organismo. Esta molécula se compone de dos cadenas complementarias enrolladas entre sí en forma de doble hélice y cada cadena está formada por un azucar, un fosfato y una base (A,T,C o G) unida al azúcar [ADN, NHGRI, 2024].

a un fenotipo [2]. Esto ocurre debido a que no toda la secuencia se encuentra en zonas codificantes y, por tanto, no todas las alteraciones producen cambios reales en la persona que las alberga.

En el ADN existen cuatro bases nitrogenadas y son la adenina (A), la timina (T), la guanina (G) y la citosina (C). Estas bases se emparejan dos a dos de la siguiente manera: AT- y G - C. En el caso del ARN [3] se sustituye la base timina por el uracilo (U).

Por lo tanto, cuando se da una variación genética, lo que ocurre es que las parejas de bases nitrogenadas no se respetan según el estándar. Existen diferentes tipos de variaciones - inserciones, deleciones, duplicaciones etc. - pero las más comunes son las SNP (*Single Nucleotide Polimorphism*) que son las que hacen referencia a la alteración de una base, produciendose así un desajuste con la secuencia de referencia de una sola base nitrogenada. En la Figura 4.1 se observa de forma gráfica una variación genética.

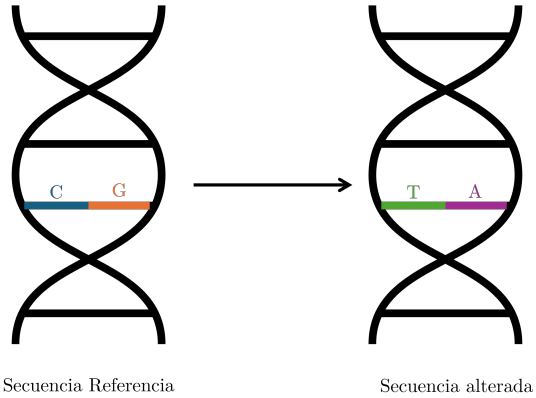

Figura 4.1: *Representación gráfica de una variación genética [Elaboración propia].*

Otro aspecto importante a recalcar a la hora de estudiar las variaciones genéticas es conocer en que assembly han sido secuenciados. Un assembly puede definirse como una reconstrucción de secuencia del genoma completo mediante el alineamiento y combinación de lecturas generadas por lecturas de secuencias [ScienceDirect, sf]. Actualmente existen dos versiones distintas de secuenciación del genoma humano de referencia, la versión GRCh37 (o hg19) y la versión GRCh38 (o hg38) y a cada una de ellas se les llama assembly. La principal diferencia entre ambas es que la versión del 37 es más antigua y que la 38 tiene mayor cobertura que la anterior. Es de vital importancia no mezlcar ambos *assembly* ya que puede haber diferencias en coordenadas genómicas produciéndose incompatibilidades entre datos de referencia.

[2] **Fenotipo:** Rasgos y características físicas de un individuo que están influenciados por su genotipo y el entorno [Fenotipo, NHGRI, 2024].

[3] **ARN o ácido ribonucleico:** Molécula similar al ADN con una sola cadena encargada de dar instrucciones desde el ADN hasta la maquinaria celular encargada de la síntesis de proteínas [ARN, 2024].

4.1.3. Hotspots

El concepto de hotspots tiene una amplia gama de acepciones dependientes de la disciplina científica en la que se emplee, presentando significados distintos dentro de la geofísica, las ciencias ambientales y la genética. Comúnmente se denomina hotspot a un área de gran actividad o importancia [Rodrigues, 2013]. En el caso de estudio de este Trabajo Final de Máster se emplea el concepto de hotspot ligado a la genética, el cual tiene implicaciones relevantes para el diagnóstico y prevención de enfermedades genéticas.

Sin embargo, es crucial comprender que existen diferentes tipos de hotspots según la función que tengan dentro del genoma humano. Actualmente se conocen dos tipos de hotspots diferentes: los hotspots de recombinación y los hotspots mutacionales.

Por un lado, los hotspots de recombinación son regiones del genoma caracterizadas por distribuirse de forma heterogénea donde la frecuencia de recombinación oscila entre una y dos kilobases [Paul, 2016] y pueden causar distorsiones del mapa genético[4]. También se pueden definir los hotspots de recombinación como regiones locales dentro de cromosomas en las que la recombinación se concentra entre regiones de poca frecuencia combinatoria o coldspots. [Choi and Henderson, 2015].

Por el otro lado, los hotspots mutacionales tienen una definición menos precisa y por tanto, más compleja. Sin embargo, estos son los que pueden albergar regiones con variaciones potencialmente patogénicas y, por tanto, son el objeto de estudio de este Trabajo Final de Máster debido a su alto interés clínico.

La Fundación Instituto Roche define hotspot mutacional como "secuencias de DNA muy susceptibles de ser mutadas debido a una inestabilidad inherente, tendencia al entrecruzamiento desigual o predisposición química a sustituciones de nucleótidos simples; región en la que observan mutaciones con más frecuencia de lo habitual"[Roche, sf].

La guía de aplicación clínica de la secuenciación masiva en síndromes mielodisplásicos y leucemia mielomonocítica crónica define un hotspot como una "zona dentro del genoma propensa a ser alterada y en la cual se detectan variantes más frecuentes. Dicha región puede comprender un solo nucleótido, un codón o un exón" [GCECGH, sf].

Del mismo modo, este término se encuentra definido en la revisión de Rogozín como "posiciones de nucleótidos con una frecuencia de mutación excepcionalmente alta" [Rogozin and Pavlov, 2003]. En este artículo también se describen las características de estas regiones, hablando de ellas como zonas que muestran el nivel de interacción entre mutágenos o como zonas con mecanismos específicos de mutación.

[4]**Mapa genético o de ligamiento:** Mapa que muestra la ubicación relativa de marcadores genéticos (que reflejan sitios de variantes genómicas) en un cromosoma [Mapa Genético, NHGRI, 2024].

Todas las definiciones anteriores se basan en la idea de ser regiones con alta presencia mutacional y estas mutaciones se miden a raíz de las variaciones genéticas. Tal y como queda expuesto en el subapartado anterior, una variación genética es cualquier alteración de la secuencia de referencia y, por tanto, una mutación es una variación genética. Por este motivo, es crucial tanto conocer cuáles son las variaciones genéticas existentes en las secuencias de los pacientes como donde se encuentran a lo largo de la secuencia.

Si bien se aclara que los hotspots son regiones con alta presencia de mutaciones, no quedan tan claros otros aspectos de la definición tales como se caracteriza la región y la parametrización del adjetivo alto. Esta imprecisión en la definición implica que cada investigador interprete de una forma este concepto. En el contexto del trabajo supone que cada método tenga una serie de asunciones en las que se respalda. Este aspecto es importante ya que, a través del estudio de las asunciones de cada método junto con el análisis posterior a realizar en el Capítulo 7, se puede concretar qué asunciones provocan una mejor detección de esas regiones.

4.2. CardioHotspots

CardioHotspots es una base de datos curada a mano con información de regiones consideradas como hotspots mutacionales y de regiones descartadas de serlo elaborada durante mi Trabajo Final de Grado y mejorada tras la defensa del mismo conjuntamente con investigadores del grupo PROS [S.Garcia et al., 2024]. Esta fuente de datos se estructura en base al modelado conceptual de un hotspot, definido durante mi Trabajo Final de Grado. En la Figura 4.2 se observa el modelado conceptual[5] empleado para definir este concepto genético y, a través del cual, se definen las variables necesarias para identificarlo y poder caracterizarlo correctamente.

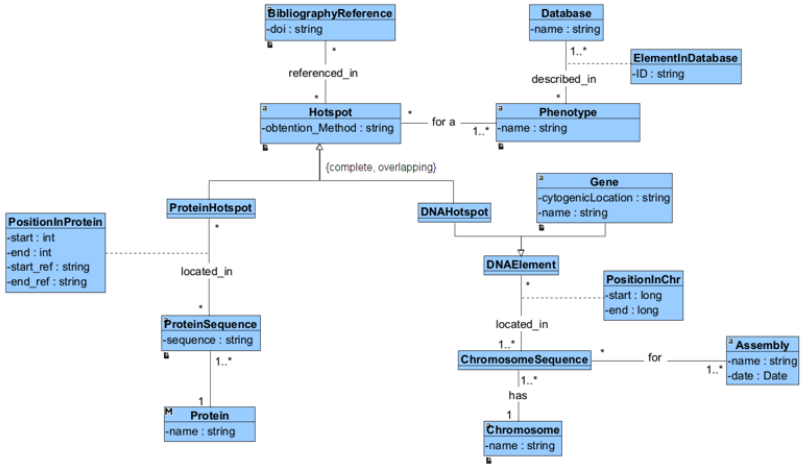

Figura 4.2: *Modelado conceptual del concepto de hotspot[García Zarzoso, 2023].*

[5]**Modelado conceptual:** Descripción del conocimiento del dominio en el cual se desarrollará un sistema de información [Olivé, 2007].

Conocer la descripción de las características que definen estas regiones son de vital importancia para entender de una forma global, no solo la definición de este concepto, sino también que se necesita para poder identificarlos. Si se conocen estas características se puede estudiar si los métodos predictivos que se van a implementar contemplan todas las características necesarias para la definición completa o, sin embargo, solo emplean algunas de ellas. Asimismo, también es interesante conocer sus descripciones para así, tras detectar los hotspots, poder determinar cuáles son las características que emplean para identificarlos. Del modelado conceptual de la 4.2 se extraen 14 características distintas que definen un hotspot y la descripción de las mismas se observa en la Tabla 4.1.

Variable	**Tipo**	**Descripción**
CROMOSOMA	categórica	Cromosoma al que pertenece, siendo el 23 y el 24 referentes al cromosoma X y al cromosoma Y.
LOCALIZACIÓN CROMOSÓMICA	categórica	Localización específica de un gen en un cromosoma.
GEN	categórica nominal	Símbolo del gen al que pertenece.
UNIPROTID	identificador	Referencia a la proteína en UniprotKB .
PROTCHANGE	categórica nominal	Cambio proteico definido con los aminoácidos de referencia y alternativos en su forma corta y la posición que ocupa. En caso de ser una región se separa el inicio del fin por un guión.
INICIO GRCH37	numérica discreta	Posición genómica de inicio del assembly GRCh37.
FIN GRCH37	numérica discreta	Posición genómica de final del assembly GRCh37.
FENOTIPO	categórica nominal	Fenotipo asociado a cada región o variación estandarizado con las ontologías HP [Human Phenotype Ontology (HP),], OMIM [OMIM, sf] y MONDO [MONDO Consortium, sf].
DOI	texto	Digital Object Identifier o identificador del artículo.
REFERENCIAS ADICIONALES	texto	Identificadores adicionales con los que se identifica la región.
FUENTE AUXILIAR	categórica nominal	Variable categórica con 3 niveles, en función del tipo de fuente auxiliar empleada para obtener toda la información. Los tres niveles son NINGUNA, WINTERVAR o VARSOME.
REFERENCIAS ADICIONALES	texto	Identificadores adicionales con los que se identifica la región.
COMENTARIOS	texto	Comentarios acerca de la región o variación.
HOTSPOT	categórica dicotómica	1 si es un hotspot y 0 en caso contrario.

Tabla 4.1: *Descripción de las variables de CardioHotspots [Elaboración propia].*

Capítulo 5

Diseño de investigación e inferencia

El diseño de la investigación es crucial para definir la estrategia de realización del proyecto y viabilidad del mismo y garantizar la reproducibilidad del experimento. En este proyecto la estrategia se divide en 5 etapas: (1) Revisión bibliográfica y selección de métodos, (2) Generación de los datos de entrada, (3) Análisis teórico de los métodos seleccionados, (4) Adaptación de los datos de entrada a cada método e (5) Implementación de los métodos. En la Figura 5.1 se observa un esquema de esta estrategia en la que se esquematizan las principales tareas a realizar en cada etapa.

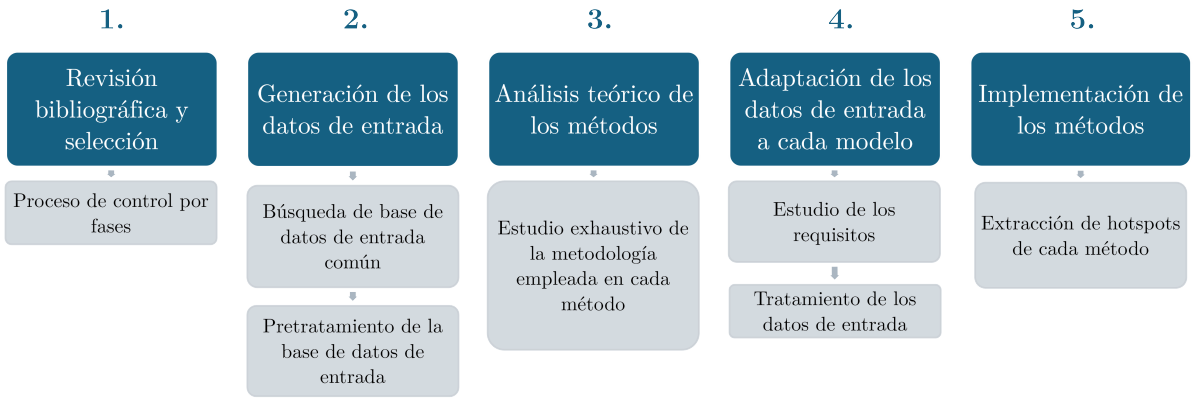

Figura 5.1: *Diseño de la estrategia de investigación [Elaboración propia]*

5.1. Etapa 1: Revisión bibliográfica.

La revisión bibliográfica es una etapa imprescindible en cualquier proyecto de investigación, ya que brinda el fundamento teórico necesario para contextualizar el proyecto dentro de un campo. Gracias a ello, se pueden identificar los problemas existentes y evitar técnicas ya probadas con resultados no satisfactorios. Además, analizar de forma crítica los estudios previos y estudiar las técnicas ya empleadas permite evitar errores y asegura reproducir de forma fidedigna las técnicas que han resultado exitosas. Este proceso no solo fortalece la credibilidad de este proyecto, sino

que también facilita la construcción de un marco teórico robusto.

En este Trabajo Final de Máster se opta por una revisión bibliográfica secuencial y de cribado que consta de 4 fases. En cada una se establecen criterios de aceptación o rechazo de los métodos para así, poder establecer unas directrices estandarizadas a la hora de seleccionarlos. En esta primera etapa del ciclo, se sigue el siguiente proceso:

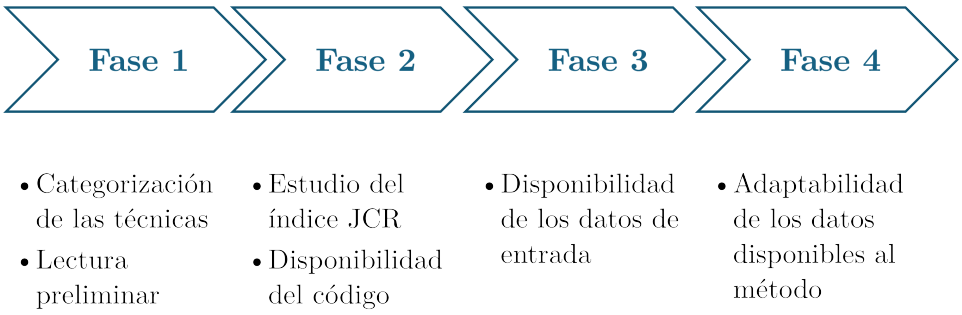

Figura 5.2: *Esquema de las tareas principales de la revisión bibliográfica [Elaboración propia]*

En la **fase 1** se parte de una lista de 48 artículos a revisar extraídos de un estudio del arte anterior centrada en la detección de hotspots relacionados con el cáncer [Almeida et al., 2020] y del resultado del buscador **predictive models AND hotspots** en PubMed [PubMed, NCBI, sf]. El primer paso a realizar para aceptar o descartar artículos es la clasificación de la técnica de extracción de los hotspots se acoge realmente a la pregunta de investigación. Por este motivo, en este trabajo se clasifican 5 técnicas distintas: (1) técnicas basadas en 3D, (2) técnicas basadas en clústeres, (3) técnicas basadas en posición, (4) técnicas basadas en regresión y (5) miscelánea.

Las técnicas basadas en clústeres son las más comunes para este tipo de algoritmos. Estas se basan en encontrar hotspots mediante la agrupación de mutaciones de una cierta región considerando tanto la frecuencia de mutaciones como otras características biológicas relevantes.

Las técnicas basadas en posición son las que se basan en la identificación de hotspots por proxumidad o dominio de los clústeres de mutaciones. El criterio de dominio se basa en si las mutaciones pertenecen a la misma proteina o por el mismo motivo de ADN no codificante. El criterio de proximidad se define como la distancia medida en codones o bases y si este se aplica a regiones codificantes o no. Definida así la medica de proximidad, los clústeres de mutaciones se identifican mediante el uso de algoritmos de ventana de todo tipo.

Las técnicas basadas en 3D son aquellas que detectan los hotspots teniendo en cuenta el doblamiento de las proteínas en su estructura tridimensional. Para conseguir este cometido se debe encontrar información tridimensional de la estructura de las proteínas para así ser capaz de, manualmente, localizar las mutaciones en la estructura para poder aplicar métodos de detección.

Las técnicas basadas en regresión identifican los hotspots mediante el ajuste de los datos de referencia a cualquier tipo de regresión para, posteriormente, ver si los datos que se pretende analizar realmente se ajustan a esa regresión o, por el contrario, muestran diferencias significativas.

Las técnicas clasificadas como miscelánea son aquellas que proceden de diferentes orígenes y que no son lo suficientemente comunes como para crear una categoria propia. Dentro de esta categoría se acogen los métodos basados en estructura secundaria de proteinas, los métodos basados en el impacto funcional o los basados en el dominio de la cromatina. Además también se acogen técnicas basadas técnicas químicas entre otras.

Tras la clasificación, se opta por descartar las técnicas basadas en 3D y, dentro de miscelánea. El motivo de descarte de las técnicas 3D es que la curación a mano de las mutaciones puede inducir a errores de posición. El motivo de descarte de las técnicas clasificadas como miscelánea es que estas proceden de experimentos de laboratorio o que forman parte de un estudio de pacientes. Evocando al objetivo del trabajo es encontrar métodos computacionales con un enfoque estadístico y, por tanto, este tipo de técnicas no cumplen ese requisito.

Después se pasa al segundo cribado de esta fase; una lectura preliminar. En este apartado se lee la definición de hotspot que emplea cada artículo para así asegurarse de que se escogen los métodos que detectan hotspots mutacionales y no hotspots de recombinación . Tal y como queda expuesto en el Capítulo 4 , hay diferentes tipos de hotspots y diferentes acepciones de este y por tanto, es importante descartar los que no son de interés.

En la **fase 2** se criba por la reproducibilidad del método y el prestigio de la revista medido por el índice JCR (Journal Citation Reports)[Clarivate, 2024]. Este índice sirve para evaluar las revistas científicas haciendo hincapié en el impacto y relevancia que estas tienen dentro de la comunidad científica. Esta métrica mide cuantas veces se citan los artículos de una revista durante dos años en promedio, para así determinar que revistas son más influyentes en cada campo. En este cribado se escogieron tan solos los artículos con un índice JCR de Q1 y Q2, valorando así el proceso de revisión por el que han pasado los artículos. Gracias a este filtrado se asegura una alta calidad en los métodos que se van a probar.

El otro ítem contemplado en esta fase es la reproducibilidad del método, el cual se centra en la disponibilidad del código. Este ítem es crucial para garantizar un entendimiento completo del método a través del estudio del código que lo estructura.

La **fase 3** del cribado se centra en la disponibilidad de los datos de entrada de los métodos predictivos. En este tipo de estudios es importante entender a la perfección la estructura de datos para así poder estudiar la reproducibilidad de los mismos. Por este motivo, en esta fase se comprueba que se tenga la información suficiente para poder conocer la estructura de los datos que precisa cada método además de disponer de datos de ejemplo para observar un caso concreto y así asegurar una reproducción más precisa.

Finalmente, en la **fase 4**, el filtrado se basa en si es posible adaptarlo a un marco común o no. Esta última fase se centra en la adaptablidad de los datos que se disponen a las necesidades de cada modelo. Esta fase también hace hincapié en el enfoque estadístico que fundamenta el método, dejando de lado aquellos que carecen de sentido teórico.

5.2. Etapa 2: Generación de datos de entrada.

Esta etapa consiste en la búsqueda y adaptación de una base de datos común de la cual partir a la hora de implementar los métodos. Ser capaces de generar una base de datos de entrada común permite partir del mismo punto a todos los métodos para realizar comparaciones de los mismos.

Actualmente existen muchos repositorios públicos donde se almacena información de variaciones genéticas y de los cuales podemos extraer información que puede ser útil para este trabajo.

Clinvar es una base de datos pública mantenida por el NCBI (*National Center for Biotechnology Information*)[Sayers et al., 2022], la cual almacena información sobre la relación que hay entre variaciones genéticas y fenotipos [Landrum et al., 2014]. El objetivo principal de esta base de datos es centralizar y estandarizar datos genéticos clínicos y así facilitar la interpretación de variaciones en un contexto médico.

GWAS Catalog (*Genome - Wide Association Studies*) es una base de datos pública basada en estudios de asociación a nivel genómicos que almacena información sobre variaciones genéticas (SNP o polimorfismos) asociadas a enfermedades o fenotipos [Sollis et al., 2022].

NCBI Gene es una base de datos con información de genes de diferentes organismos. Dentro de la información que tiene se destaca la función y expresión de esos genes, las vías biológicas en las que participan y la interacción con otros genes y proteínas [NCBI, 2005].

RefSeq o *Reference Sequence Database* es una base de datos sobre secuencias de referencia mantenida por el NCBI para proporcionar un conjunto completo de secuencias genómicas, transcriptómicas y proteómicas representativas de varios organismos [NLM(US), NCBI, 2002]

LOVD o *Leiden Open Variation Database* es una base de datos de variaciones genómicas gestionada por el Centro Médico Universitario Leiden en los Países Bajos. En esta base de datos se almacena información de variaciones genéticas relacionadas con enfermedades hereditarias [Fokkema et al., 2011].

Para este proyecto, se escoge Clinvar como repositorio de donde extraer la información de partida debido a diversos motivos. En primer lugar, la definición de hotspot implica variaciones genéticas y este repositorio almacena la información por variaciones. En segundo lugar, el NCBI es uno de los institutos referentes dentro del campo de la genética y, por tanto, es el que más información contiene. En tercer lugar, Clinvar tiene una FTP a través de la cual se puede descargar la información en distintos formatos (.txt, .vcf etc.) de una forma sencilla y accesible. En tercer lugar, este repositorio contiene tanto la información de la secuenciación del genoma humano de la versión GRCh37 (o hg19) y la de la versión GRCh38 (o hg38). Esto es importante ya que, los métodos pueden estar empleandose en cualquiera de los dos *assembly* y tener la oportunidad de descargarse ambos permite más flexibilidad.

5.3. Etapa 3: Análisis teórico de cada método.

Tras generar unos datos de entrada comunes, se pasa al análisis teórico de los métodos. En esta etapa se hace un estudio de los fundamentos teóricos subyacentes de cada método para estudiar no solo el código para implementarlo, también la estadística detras del mismo. En este apartado se estudian tanto las asunciones de los métodos como la metodología empleada para extraer los hotspots.

Por ejemplo, asumir que los datos siguen una distribución normal y realizar una regresión lineal a partir de esa asunción cuando realmente los datos no se ajustan a esta, supone un error crucial a la hora de implementar cualquier método predictivo que se sustenta en esa asunción. Del mismo modo, asumir la independencia entre muestras que realmente son dependientes puede suponer un error de interpretación de los resultados así como la equivocación en la elección de los tests y las métricas de evaluación.

Asimismo, en esta etapa se detallan de forma teórica los parámetros de extracción y evaluación de los métodos para así poder tener una visión holística del mismo. La estadística es una disciplina científica que se ocupa de obtener y analizar datos para obtener explicaciones y predicciones de esos datos. Por tanto, conocer la teoría detrás de cada método ayuda a comprenderlos.

5.4. Etapa 4: Adaptación de los datos de entrada a cada modelo.

En este apartado del diseño se adapta la base de datos de Clinvar en función de las especificaciones de cada método. Este paso es crucial porque cada método requiere de una estructura de datos distinta. Por este motivo, este apartado se compone de dos grandes tareas; el estudio de los requisitos y el tratamiento de los datos de entrada.

La primera tarea consiste en un estudio exhaustivo de los archivos y ficheros necesarios para implementar cada método. Esta tarea es fundamental para conocer los requisitos de entrada. En función del método que se estudie, los datos de entrada pueden venir comprimidos en un solo archivo o, por el contrario, precisar información adicional para realizar la predicción. Asimismo, la gran heterogeneidad a la hora de almacenar los datos y sus diferentes características revelan la necesidad de esta etapa dentro del diseño de la investigación.

La segunda etapa consiste en el tratamiento de la base de datos de entrada para cada modelo. Poseer un marco común de datos de entrada nos ofrece la ventaja de poder comparar los resultados pero sin embargo, implica adaptar estos a cada método según los requisitos de la primera tarea. Para conseguirlo se emplean herramientas auxiliares de consulta que permiten conocer información adicional de las variaciones que ayudan a completar información necesaria en caso de que los datos provistos por Clinvar sean insuficientes para satisfacer los requerimientos del método.

5.5. Etapa 5: Implementación de los métodos

La última etapa del diseño de la investigación se centra en la propia implementación de los métodos. En este apartado se emplea todo el conocimiento adquirido durante el diseño para ser capaces de extraer resultados. Para lograr estos resultados es importante dominar los lenguajes de programación en los que se desarrollan, puesto que el objetivo de este trabajo es analizar, adaptar y comparar, no cambiar el código. En este trabajo se han usado dos lenguajes de programación: R y Python.

Estos lenguajes son los más comunes dentro del ámbito de la bioinformática, el análisis de datos y la estadística, puesto que poseen librerías muy específicas con funciones útiles en estos ámbitos. Estas características los hacen los candidatos ideales para ser empleados como lenguajes para este cometido. Es importante recalcar que se han adaptado los códigos de los métodos para poder hacer posible su implementación sin alterar las bases estadísticas de estos.

Capítulo 6

Ejecución de la investigación

En este capítulo se presentan los resultados derivados de la implementación del diseño de la investigación expuesto en el Capítulo 5. El propósito del capítulo es ofrecer una visión clara y detallada de cómo se han implementado cada una de las etapas así como los resultados de las mismas. Esta sección se estructura en 6 secciones principales. Las dos primeras son generales y se atienden a las dos primeras etapas del diseño de la investigación debido a que no son dependientes de los métodos estudiados.

Por otro lado, las cuatro secciones restantes se corresponden con los cuatro métodos seleccionados de la revisión bibliográfica, a partir de los cuales se ha implementado cada etapa del resto del diseño de la investigación. Esto es debido a que trabajar un método hasta al final del diseño una vez está seleccionado permite seguir el hilo del estudio y tener una visión secuencial de los resultados que se generan de cada método seleccionado.

6.1. Revisión bibliográfica

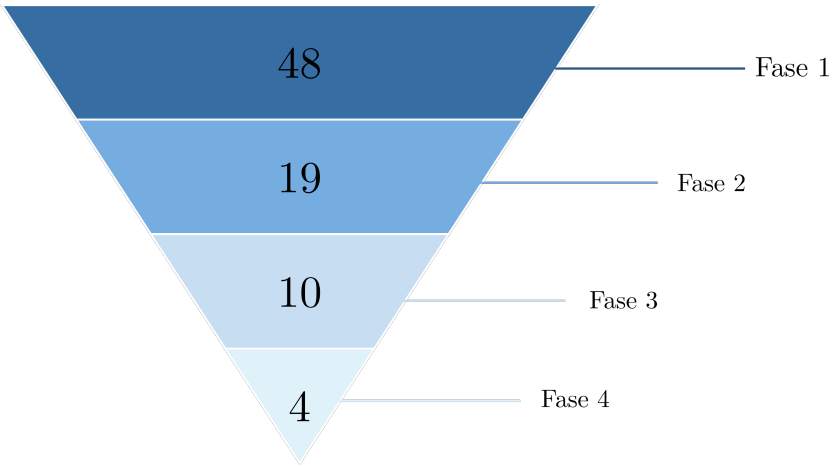

Figura 6.1: *Filtrado de artículos en cada una de las fases [Elaboración propia]*

En esta sección se muestran los resultados de la revisión bibliográfica después de aplicar las cuatro fases de filtrado expuestas en la Figura 6.1. En el apartado de Anexos 1.2, se puede observar una sucesión de tablas en la que se muestran los filtrados secuenciales de esta etapa así como los motivos de aceptación o rechazo de cada artículo en cada etapa. Atendiendo a la Figura 5.2, la primera tarea a realizar es la categorización de las técnicas de extracción de hotspots. Los resultados de esa categorización se muestran en la Figura 6.2.

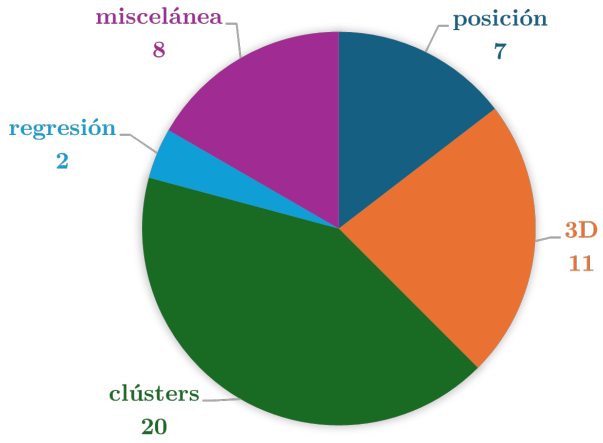

Figura 6.2: *Distribución de los artículos por el tipo de técnica de detección empleada [Elaboración propia]*

Tras esa categorización se descartaron los relativos a técnicas 3D puesto que se escapaban del propio interés del proyecto. El resto de artículos se seleccionaron para hacer una primera lectura preliminar y se descartaron 16 artículos. El motivo más recurrente fue en que realmente no detectaban hotspots mutacionales o que el cometido principal del método no era la propia detección de hotspots. Después de las lecturas se aceptaron 19 artículos y se rechazaron 29. Los artículos aceptados son: [Chang et al., 2017, Fijal et al., 2002, Wong et al., 2022, Waring et al., 2020, Chen et al., 2016, Baeissa et al., 2017, Piraino and Furney, 2017, Rhee et al., 2018, Trevino, 2020, Chang et al., 2015, Guo et al., 2018, Hess et al., 2019, Buisson et al., 2019, Rheinbay et al., 2017, Ye et al., 2010, Van den Eynden et al., 2015, Juul et al., 2021, Glazko et al., 1998]
y [Mullick et al., 2021].

En la fase 2 del filtrado, las dos principales tareas se basan en el estudio del índice JCR y en la disponibilidad del código. Tras hacer este filtrado, se aceptan 10 artículos y se rechazan 9. Es interesante comentar como el índice JCR no ha supuesto el descarte de ningún artículo, puesto que todos corresponden a revistas con índices Q1 y Q2. Los artículos aceptados son: [Chang et al., 2017],[Wong et al., 2022], [Waring et al., 2020], [Chen et al., 2016],[Ye et al., 2010], [Rhee et al., 2018], [Trevino, 2020], [Buisson et al., 2019], [Van den Eynden et al., 2015]
y [Juul et al., 2021].

En la fase 3, se revisa la disponibilidad de los datos de entrada. Tras realizar estas tareas, se aceptaron 6 artículos y se rechazaron los 4 restantes. El motivo del rechazo reside en la indisponibilidad de los datos o la falta de información para proporcionar una correcta reproducibi-

lidad. Los artículos aceptados son: [Chang et al., 2017], [Chen et al., 2016], [Rhee et al., 2018], [Trevino, 2020], [Ye et al., 2010] y [Juul et al., 2021].

Para concluir este filtrado secuencial, en la fase 4 se estudia la adaptabilidad de los datos disponibles en cada método. En este caso se descartaron 2 artículos, dejando 4 de ellos para implementar. Los motivos de rechazo de estos artículos reside en la incapacidad de adaptación de los datos debido al marco común de datos de entrada del que se dispone. Los artículos aceptados son: [Chang et al., 2017], [Chen et al., 2016], [Rhee et al., 2018] y [Trevino, 2020].

6.2. Generación de datos de entrada

El desarrollo de la generación de datos de datos de entrada se realiza con R empleando las librerias *vcfR* [Knaus and Grunwald, 2017], *tidyverse* [Wickham, 2017], *openxlsx* [Walker, 2020], *jsonlite* [Ooms, 2020], *readxl* [Wickham and Bryan, 2019], *stringr* [Wickham, 2019], *stats* [Team, 2023] y *tidyr* [Wickham and Henry, 2020]. Es importante destacar la librería vcfR, la cual permite abrir archivos VCF en R y estudiar su estructura de forma más clara. El código empleado para el filtrado se observa en el Anexo 1.4.

El VCF o *Variant Call Format* es un formato de datos con información precisa sobre posiciones en el genoma humano que es ampliamente utilizado en genética y bioinformática [HTSlib, 2016]. Esta popularidad se debe a su fácil integración con otras herramientas de análisis genéticos tales como VEP (Variant Effect Predictor), la cual permite añadir anotaciones a variaciones genéticas [McLaren et al., 2016]. El formato VCF se caracteriza por ser un fichero de texto tabulado y por tener un encabezado en el que se detalla el contenido de las columnas que existen en FORMAT [HTSlib, 2016]. Asimismo existen 8 columnas estandar y su descripción es:

- #CHROM indica el cromosoma.
- POS indica la posición donde comienza la variación.
- ID es una lista de identificadores específicos de Clinvar separados por punto y como cuando estan disponibles.
- REF indica el alelo de referencia, es decir, la base nucleotídica previa a la variación.
- ALT indica el alelo alterado, es decir, la base nucleotídica por la que se cambia.
- QUAL indica la calidad con un valoración del 1 al 100.
- FILTER indica el estado del filtro; este vale si ha superado el filtro de calidad o no.
- INFO indica información adicional acerca de la variación.
- FORMAT es un listado opcional en el que se describen mejor las muestras.

La generación de una base de datos que sirva como entrada común a los modelos se basa en los datos disponibles de Clinvar. En esta base de datos se organiza información sobre variaciones genéticas, su relación con enfermedades y las condiciones clínicas. Toda esta información proviene de grupos de investigación y laboratorios clínicos que analizan muestras de pacientes. Gracias a esta base de datos se puede obtener de forma sencilla información de millones de variaciones genéticas y descargarlas de forma libre. En este caso se descargaron los datos a través del FTP de Clinvar en un formato VCF. La ruta del archivo descargado tras acceder al FTP de clinvar es 'vcf_GRCh37_clinvar.vcf.gz'.

Este archivo hace referencia a los datos de Clinvar de la versión del genoma de referencia para el Homo sapiens GRCh37. Actualmente existen dos versiones, la GRCh37 y la GRCh38 siendo esta última la más reciente. El motivo de descarga de la versión antigua reside en la reproducibilidad de los modelos, ya que estos se han basado en esta versión del genoma humano para desarrollarse así como el assembly empleado en la base de referencia CardioHotspots.

El punto más destacado de este análisis es la división en lotes de la base de datos original puesto que la máquina empleada no es capaz de procesar tanta información a la vez. La máquina empleada para el desarrollo de esta tarea tiene un procesador Intel(R) Core(TM) i7-8565U con una CPU de 1.80GHz y una RAM de 8 GB. Para ello se divide la base de datos en lotes de 30 000 variaciones genéticas y se procesan una a una. En la Figura 6.3 se observa un esquema de la generación de esta fuente de datos.

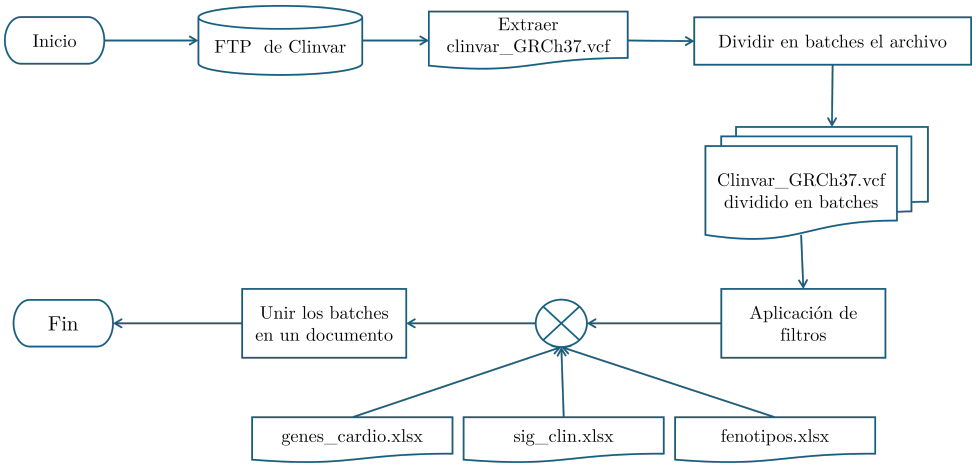

Figura 6.3: *Diagrama de flujo para la generación de los datos de entrada [Elaboración propia]*

Finalmente, el archivo de datos inicial está formado por 2690252 variaciones genéticas y 8 variables estándar y 36 relativas al campo FORMAT. Debido a que este trabajo se centra en la detección de hotspots relacionados con cardiopatias, se realiza un filtrado de la base de datos.

La necesidad del filtrado de la base reside en la focalización de la base al área clínica de interés. Por tanto, el filtrado de la base de datos se puede dividir en tres grandes apartados: (1) significado de las variaciones, (2) gen y (3) fenotipo asociado.

El primer filtro es el relativo al significado clínico de las variaciones. En este caso se aceptan las variaciones definidas como patogénicas, potencialmente patogénicas o de riesgo y todas sus combinaciones. Más concretamente, de las posibles anotaciones de significado se quedaron como válidas:

- "Likely pathogenic"
- "Likely pathogenic,low penetrance"
- "Likely pathogenic—risk factor"
- "Likely risk allele","Pathogenic"
- "Pathogenic/Likely pathogenic"
- "Pathogenic/Likely pathogenic/Pathogenic, low penetrance"
- "Pathogenic/Likely pathogenic—other"
- "Pathogenic/Likely pathogenic—risk factor"
- "Pathogenic—association—protective"
- "Pathogenic—other"
- "Pathogenic—protective"
- "Pathogenic — risk factor"

El segundo filtro es una lista de potenciales genes relacionados con cardiopatías provinente del Hospital La Fe de Valencia. Se conocen muchos genes y se sabe cuáles están relacionados con las cardiopatías, por lo que el estudio se enfoca únicamente en los potencialmente relacionados con las cardiopatías. La lista de genes de interés se puede observar en el Anexo1.5.

El tercer filtro es el relativo al fenotipo, ya que este trabajo se centra en cardiopatías familiares. Para determinar qué fenotipos se aceptaban se hizo una revisión curada manualmente de todos los posibles fenotipos ayudados de MeSH. MeSH es un vocabulario controlado y organizado jerárquicamente empleado generalmente para indexar, catalogar y buscar información biomédica relacionada con la salud [NCBI, 2024].

En este caso se emplea para estudiar los fenotipos relacionados con las cardiopatías y así poder determinar cuáles son los que se retienen para el estudio de predicción de hotspots relacionados con cardiopatías familiares. Para conseguir este cometido se emplea el archivo XML disponible en su web y se estudian las enfermedades de la rama C14, referente a enfermedades cardiovasculares y se descarta el resto. Los fenotipos aceptados se muestran en el Anexo 1.6.

Finalmente, la base de datos resultante consta de 13135 observaciones y 44 variables y la descripción de las variables se extrae del FTP de Clinvar [Clinvar, 2024] y esta se observa en el Anexo 1.3,. Sin embargo, las variables empleadas a lo largo del proyecto se definen en la tabla posterior.

Variable	Tipo	Descripción
CHROM	categórica	Número del cromosoma al que pertenece.
POS	numérica	Localización específica de la variación.
ID	categórica	Identificador de la variación interna de Clinvar.
REF	categórica	Base o conjunto de bases nitrogenadas de referencia.
ALT	categórica	Base o conjunto de bases nitrogenadas alternativas.
QUAL	numérica	Calidad. Esta columna siempre se rellena con '.'
FILTER	categórica	Estado del filtro. Esta columna siempre se rellena con '.'
ALLELEID	categórica	Identificador único del alelo.
CLNDN	categórica	Nombre de enfermedad preferido de ClinVar para el concepto especificado por los identificadores de enfermedad en CLNDISDB.
CLNDNINCL	categórica	Solo para variantes incluidas. Nombre de enfermedad preferido de ClinVar para el concepto especificado por los identificadores de enfermedad en CLNDISDBINCL.
CLNDISDB	categórica	Identificadores de los fenotipos en diferentes ontologías (MONDO, MedGen,OMIM, HP y Orphanet).
CLNDISDBINCL	categórica	Solo para variantes incluidas. Pares de etiqueta-valor del nombre y el identificador de la base de datos de enfermedades.
CLNHGVS	categórica	Expresión HGVS de nivel superior (ensamblaje primario, alt o parche).
CLNREVSTAT	categórica	Digital Object Identifier o identificador del artículo
CLNSIGCONF	categórica	Clasificación de línea germinal conflictiva para esta única variante; los valores múltiples están separados por una barra vertical.
CLNVC	categórica	Tipo de la variación.
CLNVCSO	categórica	Identificador relativo a CLNVC según la ontología SO.
DBVARID	categórica	Adquisiciones nsv desde dbVar para la variante.
GENEINFO	categórica	gen(es) para la variante informada como símbolo genético: NCBI GeneID. El símbolo y el ID del gen están delimitados por dos puntos y cada par está delimitado por una barra vertical.
MC	categórica	Lista separada por comas de consecuencias moleculares en forma de ID de ontología de secuencia | consecuencia$_{m}olecular$.

Tabla 6.1: *Descripción de las variables utilizadas en el análisis [Clinvar, 2024].*

6.3. Método 1: Accelerating discovery of functional mutant alleles in cancer.

6.3.1. Análisis teórico

Este artículo aborda la detección de hotspots definiéndolos como residuos mutados que surgen con más frecuencia de lo esperado en ausencia de selección, es decir, si no hubiera ninguna fuerza externa que favoreciera su aparición [Chang et al., 2017]. Es importante recalcar que esta metodología se basa en otra publicada por el mismo grupo de investigación y, por tanto, la explicación exhaustiva de esta metodología se encuentra en [Chang et al., 2015].

En la Figura 6.4 se puede observar un resumen de los pasos para extraer los hotspots mediante este método.

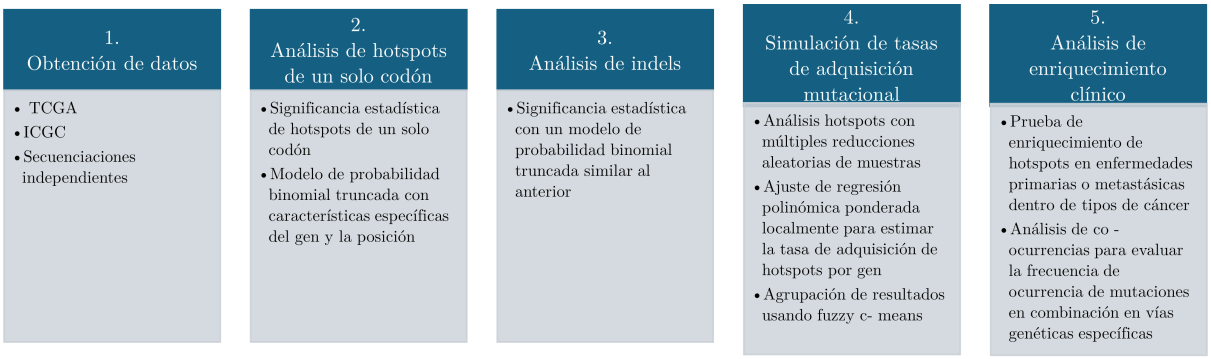

Figura 6.4: *Esquema de identificación de hotspot [Elaboración propia].*

Obtención de datos.

Para conseguir este cometido se extraen datos mutacionales de 3 fuentes de datos públicas (The Cancer Genome Atlas, TCGA; The International Cancer Genome Consortium; y secuenciaciones independientes publicadas en artículos científicos) y se escogen 10945 muestras de 10336 pacientes únicos. De toda esta información se analizan las mutaciones somáticas[1], inserciones[2] y deleciones[3] y alteraciones en el número de copias del ADN .

Análisis de hotspots de un solo codón.

La metodología empleada en este artículo se basa en el cálculo de una tasa de mutación de fondo, es decir, una tasa basal de mutaciones, y su comparación con la muestra estudiada a partir de una estrategia basada en la distribución binomial. Esto permitirá calcular la significación estadística de que una determinada región genómica se pueda clasificar como hotspot.

[1]**Mutación somática:** Alteración del ADN que ocurre después de la concepción, es decir, no pasan de generació en generación [NCI, 2024]

[2]**Inserción:** Tipo de mutación que implica la adición de uno o más nucleótidos en un segmento de ADN [Inserción, 2024].

[3]**Deleción:** Tipo de mutación que implica la oérdida de uno o más nucleótidos de un segmento de ADN [Delecion, 2024].

En concreto, se aplica un modelo de probabilidad binomial truncado que incorpora información sobre características subyacentes a las tasas de mutación específicas del gen analizado como la longitud del gen, su mutabilidad específica y posición, y la carga mutacional general del gen. Este modelo es útil ya que evita la sobreestimación de mutaciones recurrentes en un gen.

Como sabemos, el modelo de probabilidad binomial se define de la siguiente forma:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6.1}$$

En nuestro caso, X es la variable que mide el número de mutaciones que se producen en n muestras n será el número de muestras, k el número de mutaciones, y p la probabilidad de mutación en cualquier muestra. El modelo que se emplea es específico de cada gen, porque la probabilidad de que un gen mute depende de su tamaño y su tendencia a mutar. Este modelo calcula la probabilidad de que una variante de un gen mute comparándolo con las mutaciones de ese mismo gen. Asi se puede calcular cuán probable es que un gen particular mute en un lugar especíico, tomando en consideración las diferencias que hay entre los factores que afectan a las mutaciones en el cuerpo. Por este motivo, es crucial conocer el parámetro p de cada uno de los genes y eso es lo que se pretende conseguir.

Un modelo de probabilidad binomial truncado se caracteriza por ser un modelo binomial estándar, pero con un ajuste ya que ciertas observaciones (mutaciones en este caso) no son posibles o se excluyen del análisis. El modelo binomial tiene los siguientes requisitos aplicados a este caso:

- El experimento tiene un número fijo de muestras denominadas como n.
- En cada muestra puede aparecer una mutación o no.
- La probabilidad de mutación es constante para cada muestra y se denomina p.
- Las muestras son independientes entre sí.

Debido a que cada gen se compone de distintos nucleótidos y sus tasas de mutación de fondo son variables, se estima un coeficiente de probabilidad en una posición específica que junta la mutabilidad de los trinucleótidos en los que se formó la mutación, y la composición del trinucleótido afectado de cada gen. Se calcula la mutabilidad de un trinucleótodo como:

$$\mu_t = \frac{C_t}{F_t} \tag{6.2}$$

Dónde:

- C_t el número de mutaciones que afectan a la posición central del trinucleótido t de las muestras.
- F_t el número de apariciones del trinucleótido t en el genoma codificante.

Debido a que un codón[4] mutado en un gen se comprende de mutaciones en cualquiera de los tres trinucleótidos que forman el codón, se estima la mutabilidad de un codónc en el gen g:

$$m_{(c,g)} = \frac{\sum_{t \in c} m_t n_{(t,c)}}{n_c} \tag{6.3}$$

Donde:

- $n_{(t,c)}$ es el número de mutaciones en la posición central del trinucleótido t en el codón c
- n_c es el número de mutaciones del codón c en general
- $m_t = \frac{C_t}{F_t}$ es la mutabilidad del trinucleótido t con:

Asimismo, es importante conocer la mutabilidad del gen g y este se calcula como:

$$\mu_g = \frac{C_g}{nL_g} \tag{6.4}$$

Donde C_g es el número de mutaciones que afectan al gen sobre n muestras y L_g es la longitud del gen en aminoácidos.

De igual modo, la mutabilidad esperada para cada gen es:

$$m_g = \frac{\sum_t N_{t,g} m_t}{L_g} \tag{6.5}$$

Siendo $N_{t,g}$ el número de mutaciones del trinucleótido t en el gen g.

Tras conocer esto, ya se puede calcular el parámetro p de la distribución binomial para la detección de hotspots específicos de cada gen como:

$$p_{(c,g)} = r_{(c,g)} \mu_g \tag{6.6}$$

En la Figura 6.5 se observa de forma esquemática cómo obtener el parámetro p de la binomial para cada gen.

[4]**Codón:** Secuencia de ADN o ARN de tres nucleótidos (trinucleótido) que forma una unidad de información genómica que codifica para un aminoácido determinado o señaliza la terminación de una síntesis de proteínas [Codon, 2024].

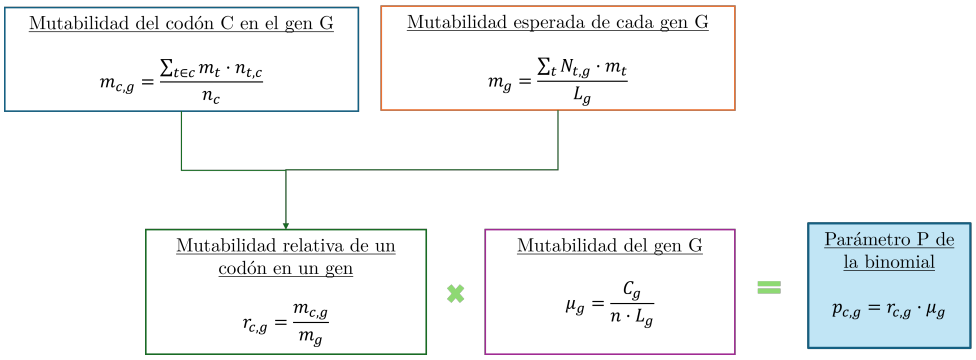

Figura 6.5: *Esquema de extracción del parámetro p de la binomial para cada gen.*

Para evitar la sobre estimación de la tasa de fondo mutacional para un gen con distintos hotspots se desarrolla el truncamiento. Este proceso consiste en eliminar las posiciones dentro del gen con una cantidad de mutaciones igual o mayor al percentil 99 de todas las mutaciones observadas en ese gen. Gracias a este proceso se evita que las zonas sobre estimadas influyan en la detección de hotspots.

En el caso de genes que mutan raramente, donde la probabilidad de encontrar un hotspot es muy baja, se limita la tasa de falsos descubrimientos o FDR. Esto se hace asegurando que los hotspots identificados no tengan una probabilidad menor que el percentil 20 de todo el conjunto de datos y la forma de ajustar esa probabilidad binomial es la que se observa en la Ecuación 6.7:

$$p''_{(c,g)} = \text{máx} \begin{cases} p'_{(c,g)} \\ 20\,\%\text{ile of all } p' \end{cases} \tag{6.7}$$

El FDR o *False Discovery Rate* es una técnica que conceptualiza la tasa de errores tipo I en las pruebas de contraste con múltiples comparaciones y el parámetro que devuelve esta técnica es el q – valor, que compara la tasa de falsos positivos con el total de falsos positivos y verdaderos positivos.

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{VP}} \tag{6.8}$$

Una vez calculados los p-valores unilaterales para todos los aminoácidos únicos en cada gen, se estudia la probabidad de observar más mutaciones de las esperadas por el azar. Para abordarlo, se corrigen estadísticamente los p-valores para así evitar falsos positivos, usando el método de Benjamini-Yekutieli para ajustar por múltiples hipótesis. En él, se ajustan los p–valores teniendo en cuenta las múltiples pruebas a través de un valor de α, siendo este el máximo rango de p-valor que se quiere controlar. En este caso, el método se caracteriza por tratar con la dependencia entre modelos.

$$P_{(k)} \leq \frac{k}{m \cdot C_{(m)}} \cdot \alpha \tag{6.9}$$

Siendo k el p - valor de estudio, m el número de pruebas y $C_{(m)}$ el factor de corrección para tratar la dependencia calculado como:

$$C_{(m)} = \sum_{i=1}^{m} \frac{1}{i} \tag{6.10}$$

Se consideraron estadísticamente significativos los hotpots mutacionales correspondientes a un valor de q < 0.1 (tasa de descubrimiento falso $< 10\,\%$).

Análisis de indels.

Después de ajustar los p-valores, se estudia la significación estadística de las inserciones y deleciones (indels), excluyendo las mutaciones que cambian el marco de lectura. Para estos indels se permite que la probabilidad disminuya por debajo del percentil 20, a diferencia de los codones individuales. Los indels se agruparon utilizando una región común máxima definida como la región genómica contigua abarcada por indels superpuestos. El recuento de mutaciones para cada una de estas regiones es la suma de todos los indels en el marco que abarcan.

Simulación de tasas de adquisición mutacional.

Tras el cálculo de la significación estadística de los codones individuales y los indels, se simularon las tasas de adquisición de mutaciones para así estimar la tasa de adquisición de hotspots dentro de los genes. En otras palabras, se estudia con qué frecuencia aparecen regiones de alta mutación en los genes.

Para ello, se seleccionan aleatoriamente conjuntos de pacientes (en muestras incrementales de 100 pacientes) y se estudia si las regiones continúan siendo significativas o no. En cada subconjunto aleatorio de pacientes, se ajusta una regresión polinómica local para cada gen y se evalúa cómo estas regiones de alta mutación aparecen o no.

Los resultados se agrupan mediante la técnica de clústering *fuzzy c-means* para ver si hay patrones generales. La técnica fuzzy c - medias es un tipo de clustering y fue creada por Pal, Bezdek y Hathaway en 1996. Este algoritmo específico de aprendizaje no supervisado se caracteriza por el hecho de que los individuos pueden pertenecer a más de un grupo ya que se obtiene para cada uno su probabilidad de pertenencia a cada uno de los clústeres formados [Diaz et al., 2009]. Así, el grado de pertenencia da lugar a una partición difusa ya que este indica la fuerza de relación entre el individuo y un grupo particular. Para determinar la partición se minimiza la función objetivo J, es decir, la suma ponderada de las distancias entre los puntos de los datos y los centros de los clústeres, donde los pesos son la pertenencia de cada punto a cada clúster. Para la determinación del número óptimo de clústeres se empleó el criterio de la reducción de la suma de cuadrados (variabilidad intra-clúster), probando entre 1 y 15 grupos. Se fijó el número óptimo de clústeres en 4.

Análisis de enriquecimiento clínico.

Tras la simulación de las tasas de adquisición mutacional se realiza un análisis de enriquecimiento para ver si los hotspots están relacionados con enfermedades primarias o metastásicas dentro de los tipos de cáncer. Este análisis se centra en la evaluación de la relación entre la presencia de hotspots y la presencia de ciertas enfermedades específicas. Para la determinación de si un hotspots está enriquecido en un tipo de cáncer en particular, se compara la frecuencia de muestras primarias que contienen un hotspot con la frecuencia de muestras metastásicas que también lo contienen. En este caso se usa una prueba exacta de Fisher bidireccional.

Los p-valores resultantes se ajustan para considerar las múltiples comparaciones a través del método de Benjamini y Hochberg [Benjamini and Hochberg, 1995]. Esta técnica ajusta los p-valores para controlar la tasa de falsos descubrimientas. Para conseguirlo se establece un nivel deseado de tasa de falsos descubrimientos α. A partir de este se determina el valor crítico del p-valor ajustado, asegurando que el p-valor crítico sea menor que el cociente entre el rango del p-valor ordenado y el número total de pruebas (m) multiplicado por α [Benjamini and Hochberg, 1995].

$$p_k \leq \frac{k}{m} \cdot \alpha \tag{6.11}$$

Tras estudiar el enriquecimiento se hace un análisis co-mutacional para poder analizar la frecuencia con la que ocurren las mutaciones en conjunto y para ello se recurre a un análisis de co–ocurrencias.

Un análisis de co–ocurrencias es una técnica empleada para estudiar cómo ocurren conjuntamente ciertos eventos. Para poder realizar este tipo de análisis hay que definir la matriz de co-ocurrencias, la cual muestra la frecuencia de aparición conjunta de diferentes pares de eventos.

Primero construimos una matriz binaria de tamaño $2 \times j$, llamada M, donde cada entrada m_{ij} se refería al estado del gen i en la muestra j y su valor era 1 si la muestra j tenía una alteración en el punto caliente del gen i. La matriz $M_{2\times j}$ se define como:

$$M_{2\times j} = \begin{bmatrix} m_{11} & ... & m_{1j} \\ m_{21} & ... & m_{2j} \end{bmatrix} \tag{6.12}$$

Donde cada elemento m_{ij} se define como:

$$m_{ij} = \begin{cases} 1 & \text{mutación} \\ 0 & \text{no mutación} \end{cases} \tag{6.13}$$

Se realiza un análisis de co-ocurrencia para todas las combinaciones únicas de genes dentro de una vía dada. Para conseguir este análisis se crea un modelo nulo de co-ocurrencia aleatoria permutando las alteraciones observadas conservando la frecuencia general de las alteraciones observadas en la cohorte.

Los p-valores se generaron como el número de veces que se observó la co-ocurrencia igual o más a menudo en esta distribución nula en comparación con los datos observados. Se realizó una corrección para múltiples hipótesis utilizando el enfoque de Benjamini y Hochberg, y se consideraron significativas las co-ocurrencias de pares de genes dentro de la vía con un q-valor $<$ 0.01. Tras la identificación de los hotspots, en este artículo se especifican distintas anotaciones que aportan más información acerca de esos hotspots relacionados con el cáncer. Sin embargo, dentro de este proyecto carece de sentido y, por tanto, no se especifican en este apartado.

6.3.2. Adaptación de los datos

Este método requiere de varios archivos para poder ser procesado y estos son:

- **archivo MAF** con una variación genética por cada fila y 94 columnas describiendo esa mutación. La descripción de estas variables se presenta en el Anexo 1.7.
- **objeto de R** con varios data frames con información general necesiaria para llevarse a cabo descrita a continuación.

  - **mu** es un dataframe con información de la mutabilidad de cada trinucleótido posible.

  - **p** es un dataframe con información acerca de la probabilidad p de cada trinucleótido.

  dmp es un data frame con 1358 observaciones y 6 variables con información acerca de mutaciones específicas.

  - **exacr0_2snps** es un data frame de 478010 observaciones y 5 variables con información acerca de la frecuencia alélica de cada uno de los cambios de base presentados.

  - **expressiontb** es un data frame de 20502 filas y 24 columnas con información acerca de la expresión de cada gen en cada tipo de cáncer.

  - **homopolymerbed** es un data frame de 118290 filas y 5 columnas con información acerca de regiones homopoliméricas.

En este caso, solo tenemos que generar un archivo .maf a partir de los datos de entrada generados. Tal y como se observa, la estructura de datos de entrada del modelo y la base de datos de la que se dispone no tiene la misma estructura. Es por ello por lo que se precisa de un preprocesado de la base de datos para así adecuarla al formato de este método. El código de este preprocesado se puede consultar en el Anexo 1.8 y el diagrama de flujo asociado al preprocesado de esta base de datos se puede ver en la Figura 6.6.

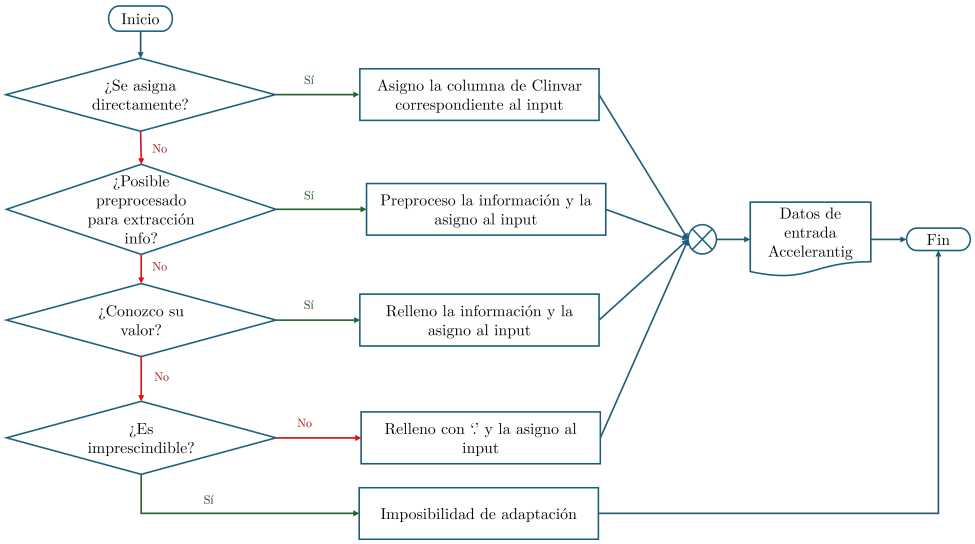

Figura 6.6: *Diagrama de flujo del pre-procesado de los datos de entrada para el método 1.*

En primer lugar, se asignan directamente las variables disponibles en la base a las columnas correspondientes del archivo .maf. Las columnas disponibles y sus asociaciones se observan en la Tabla 6.2.

Columna requerida	**Columna disponible**
Hugo_Symbol	GENEINFO
Start_Position	Start
End_Position	Stop
Variant_Classification	MC_mutation_subtype
Reference_Allele	REF
CLIN_SIG	CLNSIG
dbSNP_RS	RS
Tumor_Seq_Allele1	REF
Tumor_Seq_Allele2	ALT
Consequence	MC_mutation_subtype
HGVSp	ProtCh
SYMBOL	GENEINFO
Allele	ALT
Amino_Acid_Position	aapos

Tabla 6.2: *Relación entre las variables del .maf y de los datos de entrada.*

Tras esta asignación se pasa al preprocesado de las variables que requieren alguna ligera modificación de alguna de las columnas de la base de datos general. Las variables del archivo .maf que requieren el pre - procesado de alguna columna de la base de datos general son: Chromosome, Variant_Type, HGVSp_Short, Amino_Acids, Protein_position, Gene, Entrez_Gene_Id, HGNC_ID, Amino_Acid_Length y Protein_Length. Dado que el pre procesado de cada variable es diferente en función de las necesidades de cada uno, el proceso de ese cambio se refleja en el código en el Anexo 1.8.

Una vez se tiene rellenada esta información, se pasa a estudiar el resto de variables. Por la propia definición de las mismas expuesta en el Anexo 1.7, existen algunas columnas que se pueden rellenar y en la Tabla 6.3 se puede observar un resumen de la columna, la forma de completarse y el motivo del mismo.

Columna	Información	Motivo
Strand	+	Todas las variaciones se reportan como +.
ALLELE_NUM	1	Hace referencia al alelo alterado y, por definición, vale 1.
BIOTYPE	protein_coding	Solo se cogen mutaciones codificantes de proteína.
NCBI_Build	GRCh37	Se extrae la información solo de ese *assembly*.
CANONICAL	YES	Solo se contemplan los canónicos en la base.
Is_Ref	TRUE	Vale TRUE si hace referencia al *assembly* hg19 .

Tabla 6.3: *Motivo del completado de las columnas faltantes*

Para finalizar, se estudian las variables restantes estudiando si estas son imprescindibles para la detección de los hotspots. En este caso, se observa que la gran mayoría de las variables son necesarias para el anotado que se realiza después de la detección sin alterar el resultado de la misma, por tanto, se omite en esta proyecto. Por este motivo, se rellenan estas columnas con un '.'. En el resto, se estudia qué implicación tienen a la hora de la detección.

Concretamente, las variables asociadas a los códigos de barras de cada tipo de tumor y el identificador de los pacientes influyen en la creación de la variable Master_ID. Esta variable es de vital importancia en el algoritmo ya que, a partir de esta, se cuenta el total de muestras diferentes que se tienen por paciente y se calcula el p-valor asociado a cada una de ellas por separado. Este tipo de métodos pueden denominarse como poblacionales dado que necesitan información específica de pacientes para poder desarrollarse. Desafortunadamente, la base de datos que se tiene de entrada no contempla este tipo de información y, por tanto, se debe descartar el método en esta etapa al no poder reproducir de forma fidedigna los datos de entrada.

6.4. Método 2: Modeling and analysis of site-specific mutations in cancer identifies known plus putative novel hotspots and bias due to contextual sequences.

6.4.1. Análisis teórico

Este artículo define los hotspots como posiciones recurrentemente mutadas en el ADN asociadas al cáncer con un potencial impacto funcional importante. Por esta razón, el objetivo del artículo es generar un método de detección de hotspots basándose en datos públicos. En este caso se escogen datos del repositorio TCGA (`https://www.cancer.gov/tcga`), abarcando 33 tipos de cáncer, 10182 pacientes distintos y más de 3 millones de mutaciones, las cuales se agrupan por recurrencia dentro de cada gen.

Para ajustar de la mejor forma el reparto de las mutaciones, los autores probaron diferentes distribuciones de probabilidad para ver cual se ajustaba mejor a la distribución de las mutaciones recurrentes en posiciones específicas de aminoácidos dentro de genes. Las distribuciones probadas son la binomial, la geométrica, la beta–binomial y la zero–inflated beta–binomial (ZIBB).

La primera distribución que intenta ajustar es la binomial. Puesto que la definición queda explicada en el análisis teórico del método 1 y la función de probabilidad se expresa en la Ecuación 6.1, solo queda adaptarlo al contexto de este método. En este caso, la variable aleatoria X representa todas las posiciones mutadas de aminoácidos[5]de un gen, n las posiciones del gen y p la probabilidad de encontrar una mutación en esa posición.

La ***distribución geométrica*** modeliza los procesos en los que se repiten pruebas hasta que se obtiene el éxito esperado. En este tipo de distribución, al igual que en la anterior, existe una dicotomía entre los resultados (éxito o fracaso) y se asume independencia entre las pruebas. Una gran diferencia de este modelo con respecto al anterior es que este no concluye hasta que se obtiene por primera vez el éxito, que en este caso es hasta que se encuentra una mutación. Para definir una distribución geométrica se precisa de dos parámetros n y p y la forma de representarlo es:

$$P[X = x] = p(1 - p)^{x-1} \tag{6.14}$$

Siendo X el número de pruebas necesarias para encontrar la primera mutación y p la probabilidad de mutación, en este caso el éxito sería encontrar una mutación en una posición y el fracaso no encontrarla.

La ***distribución beta − binomial*** es un tipo de distribución binomial en la que la probabilidad de éxito en cada prueba no es fija, sino que sigue una distribución de probabilidad beta que a su vez, es una distribución apropiada para modelizar variables aleatorias continuas que toman valores dentro del rango 0–1. Por este motivo, resulta interesante para modelar la probabilidad de éxito del modelo binomial. La forma de definir este tipo de distribución beta es:

$$f_X(x) = \frac{x^{\alpha-1}(1 - x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \tag{6.15}$$

Siendo X la cantidad de posiciones mutadas de aminoácidos de un gen, n las posiciones del gen, p la probabilidad de encontrar una mutación que en este caso se pretende ajustar a través de una distribución beta y αy β los parámetros de forma de ajuste de esa probabilidad de encontrar la mutación p. Por un lado, el parámetro de forma α se encarga de controlar la inclinación de la distribución hacia el valor uno. Si $\alpha > 1$, la distribución tiende a concentrarse más hacia valores cercanos a uno, indicando así la tendencia a mutar de esa posición. Por el otro lado, el parámetro de forma β se encarga de controlar el fenómeno contrario, es decir, de la inclinación

[5]**Aminoácido:** Unidad basa que actúa como estructura fundamental de las proteínas [Aminoácido, NHGRI, 2024].

de la distribución hacia el valor cero.

La ***distribución zero–inflated beta–binomial*** o ZIBB es un tipo de distribución binomial que enfatiza la probabilidad de encontrar ceros, es decir, de no tener mutaciones en la secuencia. El motivo para incluirla en la comparación reside en que la mayoría de las posiciones no tienen mutaciones.

Una vez se conocen las principales características de cada una de las distribuciones seleccionadas para el ajuste de las distribuciones, se pasa a describir la metodología propuesta en el artículo. Esta metodología se basa en la separación de las mutaciones en función de los genes en los que se encontraban. Tras ello, se recuenta la cantidad de mutaciones por posición de aminoácido dependiendo de su correspondiente transcrito y proteína.

Para estimar la diferencia entre la distribución ajustada y los datos observados se realiza el test estadístico G, que se basa en la métrica de divergencia de Kullback – Leibler. El estadístico G se define según la siguiente ecuación:

$$G = 2 \sum o_i \log \left(\frac{o_i}{e_i} \right) \tag{6.16}$$

Donde o_i es el número de mutaciones observadas en la posición i y e_i el número de mutaciones esperadas en la posición i dadas por el modelo de probabilidad asumido.

En función de la distribución que se estuviera ajustando, se emplearon distintos paquetes estadísticos de R; para la distribución binomial y la geométrica se emplea el paquete *stat* [Team, 2023], para la ZIBB el paquete *gamlss*[Rigby et al., 2023] y para la beta–binomial el paquete *emd – book*[Bolker, 2023].

Una vez ajustado los datos a cada distribución, se comparan entre ellas y se concluye que, en general, el modelo beta-binomial parece ser el modelo que mejor se ajusta, ya que refleja la sobre-dispersión que se observa en los datos de mutación binomial. Por este motivo, se presenta un algoritmo de detección de estas posiciones basándose en un modelo estadístico de distribución beta–binomial dentro de cada gen. Esto implica que este algoritmo primero pasa por una fase de construcción del modelo beta–binomial específico de cada gen.

El modelo de regresión propuesto para cada gen añade una componente de efectos fijos, representando las posiciones con un exceso de mutaciones que presuntamente corresponden a localizaciones de hotspots. De forma esquemática, en la ecuación 6.17 se puede observar la estructura de un modelo beta-binomial con efectos fijos.

$$M = \text{BetaBin}(\alpha, \beta) + F \tag{6.17}$$

Donde M es el vector que almacena los números de posiciones con k mutaciones designados como M_k, F es el vector de efectos fijos de hotspots y α y β son los parámetros de forma de la distribución beta empleados para modelar el parámetro p de la distribución Binomial.

Para el ajuste del valor F de cada modelo de cada gen se emplea un algoritmo iterativo que en cada iteración evalúa las mejoras del ajuste añadiendo esos efectos fijos en ciertas posiciones y este se acaba cuando no detecta mejoras significativas o cuando alcanza un umbral determinado para evitar el sobreajuste del modelo. A este tipo de ajuste iterativo se le conoce por el nombre de *stepwise*. Profundizando en el algoritmo, el valor de cada celda o posición de la secuencia es el ratio de mejora entre el estadístico G sin efecto fijo y con él. Este algoritmo continúa hasta que el ratio más grande no es mayor que 1, indicando que no hay mejora, hasta que el número de pasos es más de dos veces el número máximo de las mutaciones o hasta que el estadístico G es menor que 1, para evitar el sobreajuste.

El resultado del algoritmo es el vector de efectos fijos F representando las mutaciones y la magnitud F_k que se mejora con los parámetros α y β, correspondientes al modelo beta–binomial.

La detección de hotspots se basa en si el valor ajustado del efecto fijo, F_k, es mayor al 50 % de las mutaciones en esa posición con un p–valor corregido – a través de la técnica FDR - menor o igual al 1 %, es decir un $q < 0.01$.

6.4.2. Adaptación de los datos

Los datos requeridos para poder implementar el modelo se centran en un vector de mutaciones observadas a lo largo de la secuencia y uno esperado resultante del estudio de la secuencia de cada gen. Por tanto, al reproducir los resultados en otro ámbito clínico, se debe poseer tanto información general de mutaciones reportadas, siendo estas empleadas para ajustar la regresión, como información de pacientes reales, siendo esta la referente a los datos observados. Acogiéndose a los datos de entrada de los que se dispone se debe descartar la adaptación de datos de este método, puesto que solo se dispone de los datos esperados de secuencia. Esto es debido a que los datos del proyecto se extraen de fuentes públicas y son mutaciones reportadas pero no disponemos de datos de pacientes para poder extraer el vector de mutaciones observadas.

Aunque este modelo es muy potente a nivel estadístico, la aplicación del mismo excede el marco de este estudio. Sin embargo, sería posible implementarlo si se tuviera datos de pacientes reales, pudiendo primero ajustar los modelos, y luego observar dónde se encuentran los hotspots en cada gen.

6.5. Método 3: Hotspots mutation delineating drives mutational signatures and biological utilities across cancer types.

6.5.1. Análisis teórico

El artículo define hotspot mutacional como la mutación que aparece en un conjunto de muestras tumorales significativamente con mayor frecuencia de lo esperado a partir de una frecuencia de referencia que depende del gen, tipo de cáncer y tipo de mutación .

Esta definición supone una diferencia con respecto a la base de datos de entrada de la que disponemos ya que, en el caso del TFM, cada observación representa una variación descrita en la base de datos Clinvar y, en el caso del artículo, representa una muestra tumoral. Sin embargo, en el algoritmo provisto por los autores no se emplea esta etiqueta para la detección del hotspot y, por tanto, esta diferencia no supone un cambio real en términos del proceso de detección.

Los datos empleados en este proyecto provienen de COSMIC, una base de datos de mutaciones somáticas, de la cual se escogieron 12250 muestras de información mutacional curada de diferentes fuentes. Asimismo, se escogió una lista de genes candidatos de cáncer y se definieron los tipos de variaciones que se contemplaban en el trabajo; estas fueron : (1) "missense", (2) "nonsense" , (3) "coding - silent", (4) "insertion" y (5) "deletion". De la misma forma, se generan subtipos de variaciones en función del cambio de base nucleotídica que se presente. Finalmente, se generaron 20 subtipos de mutaciones a considerar.

En la Figura 6.7 se muestra el esquema de flujo de trabajo de detección de hotspots de este artículo.

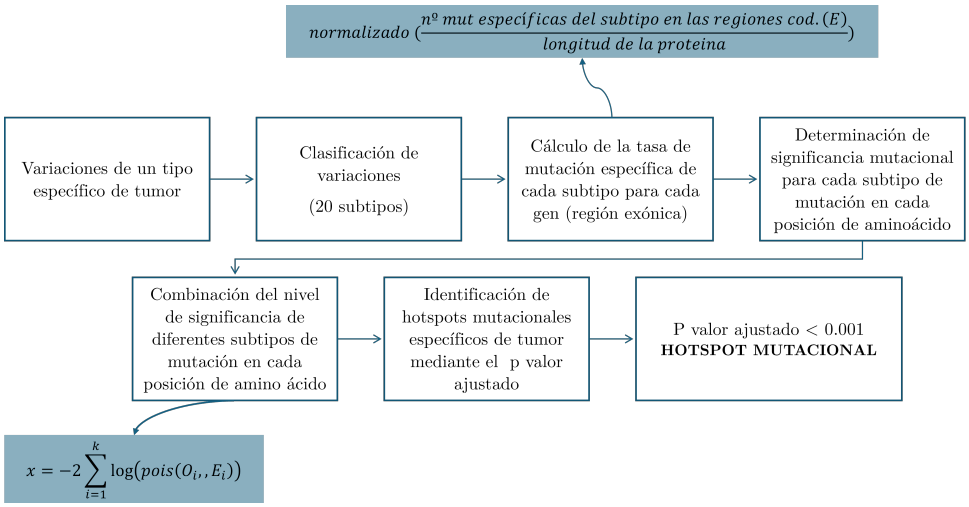

Figura 6.7: *Pasos de desarrollo de HotDriver [Chen et al., 2016]*

La detección de hotspots se basa en la posición de aminoácidos y, por tanto,el primer paso consiste en contar la cantidad de mutaciones en función del subtipo que se les ha asignado a través de toda la muestra para cada gen. Es decir, divido mi conjunto de datos en genes para,

después, contar la cantidad de mutaciones que hay de cada tipo para cada gen en mis datos. Tras realizar esta tarea, se calcula la tasa media de mutaciones específicas λ para cada subtipo en cada gen a través de la Ecuación 6.18.

$$\lambda = \frac{\text{nº de mutaciones}}{\text{longitud de la proteína}} \tag{6.18}$$

Esta fórmula normaliza el cociente de mutaciones de cada subtipo en regiones codificantes por la longitud de la proteína concreta.

Tras conocer este lambda, se calcula el p-valor de cada subtipo de mutación en cada posición de aminoácido a través de la función de probabilidad de Poisson expuesta en la Ecuación 6.19. Este cálculo se basa en la asunción de que las mutaciones específicas de cada subtipo siguen una distribución de Poisson. Este tipo de asunciones se toman para determinar si las mutaciones observadas en los datos son diferentes significativamente a lo esperado bajo una hipótesis nula. La hipótesis nula (H_0) en este caso es que las mutaciones siguen una distribución de Poisson con una tasa de ocurrencia esperada y la alternativa (H_1) que no la siguen.

El p-valor se calcula como el máximo entre esta probabilidad y un límite inferior establecido en 10^{-200}. Este límite inferior se establece debido a que la función de probabilidad de Poisson empleada tiene como condición que $\lambda > 0$ debido a que λ representa el número de veces que se espera que ocurra una mutación en un intervalo de posiciones concreta. La expresión que resume todo este proceso de forma matemática es:

$$\text{pois}(O_i, E_i) = \text{p-valor} = \text{máx}\left(10^{-200}, \frac{e^{-\lambda} \cdot \lambda^{O_i}}{O_i!}\right) \tag{6.19}$$

HotDriver combina los p-valores calculados para cada subtipo de mutación en una posición específica de un aminoácido (AA). Esta prueba sirve para encontrar una medida global de significancia cuando se hacen múltiples pruebas. En este caso se calcula un estadístico definido como se observa en 6.20 para enfrentarlo con una distribución chi – cuadrada con $2k$ grados de libertad y así conseguir el p – valor sin corregir. La expresión que combina los dos pasos anteriores es la mostrada en la Ecuación 6.20.

$$x = -2\sum_{i=1}^{k} \log(\text{pois}(O_i, E_i)) \tag{6.20}$$

En la Ecuación 6.20, k es el número de subtipos de mutaciones probadas, O_i son las mutaciones observadas de cada subtipo y E_i es el número de mutaciones esperadas de cada subtipo.

Finalmente, el algoritmo termina con la aplicación del *False Discovery Rate* o FDR para eliminar falsos positivos de nuestro conjunto de hotspots seleccionados. Además, en este artículo se emplea el método de Benjamini - Hochberg para ajustar los p valores extraídos de los z - scores. Esta corrección se encarga de ajustar los p valores para tener en cuenta la multiplicidad de las pruebas

y controlar de manera más efectiva el FDR. En este artículo se identifican las regiones como hotspots si su p – valor ajustado $< 0{,}001$.

6.5.2. Adaptación de los datos

Este método requiere de varios archivos para poder ser procesado y estos son:

- **hg_CpG_island.bed** ; el cual alberga las regiones CpG de la versión hg19 del genoma humano, es decir, las del *assembly* GRCh37.
- **gene_length.tsv**; alberga los genes y la longitud en aminoácidos de cada gen.
- **gene_list.tsv**; alberga una lista de genes candidatos.
- **mutation_data.tsv**; archivo que alberga información acerca de las mutaciones.

Los archivos .bed o Browser Extensible Data son un tipo de archivos de texto muy comunes en bioinformática para almacenar datos genómicos [Ensembl, 2024]. La estructura típica de estos archivos se compone de 3 columnas; la primera hace referencia al cromosoma de la región, la segunda a la posición de inicio y la tercera a la posición final de la región. En este caso, el archivo bed que se usa almacena información acerca de las islas CpG que existen en el genoma.

Las islas CpG se pueden definir como regiones del ADN con una alta frecuencia de aparición de la combinación citosina (C) y guanina (G) y son regiones muy importantes en la regulación de la expresión génica [Química.es, 2024]. Esto es debido a que son las encargadas de un proceso llamado metilación del ADN, proceso mediante el cual se puede alterar la expresión génica y por tanto, producir la inactivación de un gen. La inactivación de un gen da lugar a la falta de síntesis de su proteína correspondiente, por lo que puede ser precursora de algunas enfermedades.

Los archivos .tsv o Tab-Separated Values son archivos de texto usados para almacenar información separada por tabuladores [ReviverSoft, 2024]. En este método precisa de los 3 archivos .tsv descritos anteriormente. El archivo 'gene_length.tsv' es general y, por tanto, puede emplearse sin modificar. Sin embargo, se opta por no emplear el archivo 'gene_list.tsv', ya que es opcional para el modelo y en el trabajo previamente ya se ha filtrado por los genes de interés. En cuanto a la estructura del documento 'mutation_data.tsv', este tiene 8 columnas y su descripción se presenta a continuación:

Variable	Tipo	Descripción
Sample_ID	categórica	Identificador de la muestra de mutaciones.
Gene_ID	categórica	Símbolo del gen de la mutación.
Chromosome	numérica	Cromosoma al que pertenece.
Start	numérica	Posición de inicio de la mutación.
End	numérica	Posición de final de la mutación.
Ref_Allele	categórica	Alelo de referencia.
Alt_Allele	categórica	Alelo alterado.
Protein_Variant	categórica	Aminoácido de referencia en formato corto y posición de aminoácido.
Mutation_Subtype	categórica	Tipo de mutación.

Tabla 6.4: *Descripción de las variables utilizadas en mutation_data.tsv.*

Tal y como se observa en la tabla anterior, la estructura de datos de entrada del modelo y la base de datos de la que se dispone no es la misma. Es por ello que se precisa de un preprocesado de nuestra base de datos para así adecuarla al formato de este método. El código de este preprocesado se puede consultar en el Anexo 1.9 y el diagrama de flujo asociado al preprocesado de esta base de datos se puede ver en la Figura 6.8.

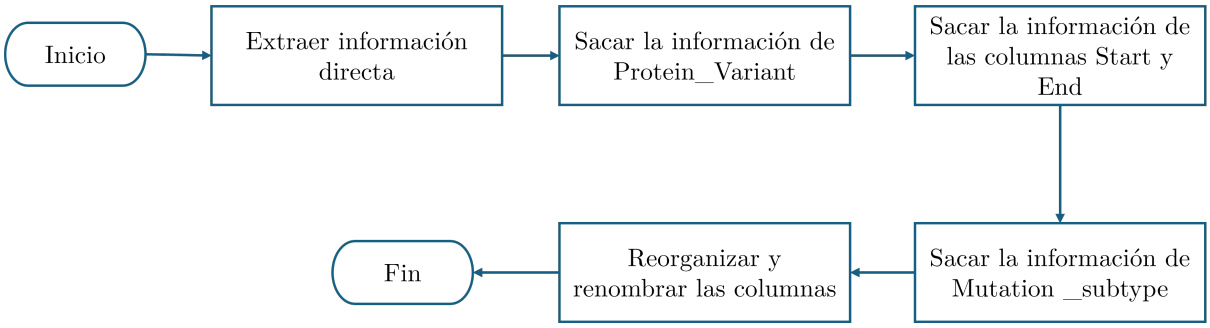

Figura 6.8: *Diagrama de flujo de adaptación de datos de entrada. [Elaboración propia]*

En primer lugar, es importante separar las columnas que pueden usarse sin precisar preprocesado de las que lo necesitan. En este caso, las columnas Sample_ID, Gene_ID,Chromosome, Ref_Allele y Alt_Allele se corresponden con las columnas ID, GENEINFO, CHROM, REF y ALT del fichero original de Clinvar obtenido en la sección 6.2. Por tanto, quedan por preprocesar las columnas Start, End, Mutation_subtype y Protein_variant.

Atendiendo al diagrama de flujo, el siguiente paso es la adecuación de la variable **Protein_variant**. Para conseguir este cometido se necesita volver a acceder al FTP de Clinvar para descargarse el archivo llamado 'hgvs4variation.txt'. Este documento reporta las expresiones HGVS (Human Genome Variation Society) por AlleleID así como los cambios asociados a cada expresión HGVS y los identificadores de secuencia de RefSeq a todos los niveles.

Las expresiones HGVS hacen referencia a un sistema estándar empleado para describir variaciones genéticas a través de sus diferentes dimensiones: secuencia genómica, cDNA [6],ADN no codificante, secuencia de ARN y secuencia de proteína. Gracias a estas expresiones, podemos

[6]**cDNA**: ADN codificante, es decir, que codifica una proteína [cDNA, NHGRI, 2024].

conocer el cambio proteico asociado a cada variante a través de la expresión de secuencia de proteína. Esta expresión tiene la estructura que se observa en la Figura 6.9.

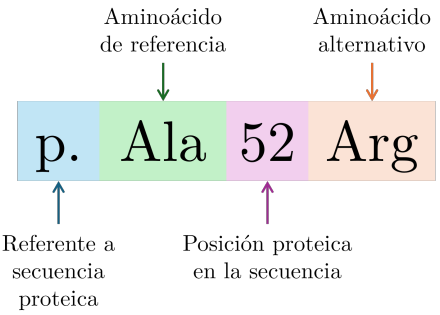

Figura 6.9: *Explicación de la estructura de las expresiones proteicas HGVS [Elaboración propia]*

Sin embargo, extraer esta expresión no es una tarea baladí y, por tanto, hay que tener en cuenta varias consideraciones.

En primer lugar, los cambios proteicos van en función de la isoforma que los padezca. Las isoformas son las diferentes variaciones de la misma proteína que surgen a partir de distintos procesos biológicos y estas pueden tener diferencias estructurales y funcionales aunque todas se originen a partir de un mismo gen o conjunto de genes [Eugenomic, 2024]. Para garantizar la homogeneización de la extracción del cambio proteico, se opta por emplear la isoforma canónica. La isoforma canónica es la variación de la proteína considerada de referencia, es decir, la forma principal.

La forma de conocer la isoforma canónica es a través de UniProtKB [Consortium, 2022]. Esta base de datos identifica los identificadores de secuencia proteica de RefSeq que hacen referencia a la isoforma canónica. Por este motivo se hace una consulta a UniProtKB para cada gen de forma automática. Sin embargo, no todos los genes tienen isoforma canónica, algunos de ellos solo tienen una y otros diversas pero no tienen determinada una como referencia. En este punto se toma la decisión de escoger la isoforma de la que UniProtKB tiene la secuenciación por posición de aminoácido. Esto es debido a que, en caso de no tener información acerca de los aminoácidos en ciertas posiciones, pueden buscarse en esta fuente de información.

Tras extraer los identificadores canónicos, se filtra la base de datos para así poder conocer las expresiones HGVS proteicas asociadas. Gracias a esta adaptación se consigue extraer la información necesaria para rellenar la variable Protein_variant.

Para obtener las variables **Start - End** atendiendo al *assembly* 37, se recurre al FTP de Clinvar. En este caso se descarga el archivo variant_summary.txt. Este archivo es un documento tabulado con las variaciones reportadas en Clinvar con su localización en el genoma. Este documento tiene 40 variables pero en este caso se trabaja con tan solo cuatro de ellas: AlleleID, Start,Stop y Assembly. En primer lugar, se filtra el documento para que solo queden las variaciones del *assembly* 37 y, posteriormente, se hace la alineación de esta base de datos con los datos de entrada a través del AlleleID para obtener el Start y el Stop, cuyos nombres pasan a llamarse

Start y End en el archivo .tsv de entrada.

Tras ello, se adecúa la variable **Mutation_subtype**. Esta variable es la más sencilla de adecuar puesto que la información proviene del arreglo de la columna MC de la base de datos principal. Tras la separación de la columna MC en MC_ID y MC_mutation_subtype, se adecua la última de ellas para que tenga las mismas categorías. En la Tabla 6.5 se puede observar la adecuación.

Subtipos de mutaciones inicial	Subtipos de mutaciones tras adecuar
missense_variant	Missense
nonsense	Nonsense
inframe_insertion	Insertion
inframe_deletion	Deletion

Tabla 6.5: Adecuación de las categorías de los tipos de mutaciones a las especificaciones de HotDriver.

Para concluir con la adaptación de los datos de entrada, se hacen unos pequeños ajustes relativos a la expresión de Mutation_subtype y de ordenación de columnas. Tras hacer todos estos ajustes se dispone de una base de datos con 7955 mutaciones y 9 columnas.

6.5.3. Implementación del método

Durante la implementación del modelo se encontraron problemas que provocaban la imposibilidad de ejecutar el código. Tras un análisis exhaustivo del código, se corrigieron esos errores, que estaban relacionados con la tabulación del propio script, la sobre-escritura de algunos objetos y la desactualización de algunas funciones debido a la versión de Python empleada.

Tras la corrección del código se procede a la aplicación del método, omitiendo en nuestro el ajuste del p-valor por tests múltiples, puesto que de lo contrario no se obtenía ningún resultado significativo. Los resultados se obtienen en un archivo con formato .tsv, que contiene las siguientes 6 variables:

- **Gene:** Gen donde hay un hotspot.
- **aaPos:** Posición proteica donde aparece ese hotspot. Primero aparece el aminoácido en su versión corta y luego la posición.
- **MutCount:** Número de mutaciones encontradas en esa posición proteica de ese gen.
- **MutSubtype:** Clasificación por subtipos de las variaciones encontradas.
- **MutPositions:** Posiciones genómicas a las que pertenece cada mutación. La estructura es cromosoma:inicio-fin base de referencia/base alterada (cantidad de veces que aparece).
- **Pvalue:** P-valor asociado a ese hotspot.

Se exponen a continuación los resultados encontrados para distintos umbrales del p-valor. La Figura 6.10 muestra los 17 hotspots con un nivel de confianza del 95 %, es decir, con un p-valor menor que 0,05 y en la Figuras 6.11 ,6.12 y 6.13 se muestran los 51 hotspots encontrados con un nivel de confianza del 90 %, es decir, con un p-valor menor que 0,10.

Gene	aaPos	MutCount	MutSubtypes	MutPositions	Pvalue
SOS1	R552	8	Missense_ATts(1),Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(4),Deletion(1)	2:39249915-39249915T/C(1),2:39249913-39249913C/G(1),2:39249914-39249914C/A(1),2:39249914-39249914C/G(1),2:39249914-39249914C/T(1),2:39249913-39249913C/A(1),2:39249915-39249915T/A(1),2:39249914-39249916CCTT/C(1)	0,00004249
IDH2	R172	5	Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(3)	15:90631837-90631837C/G(1),15:90631838-90631838C/A(1),15:90631838-90631838C/T(1),15:90631839-90631839T/A(1),15:90631837-90631837C/A(1)	0,0003051
RIT1	F82	7	Missense_ATts(2),Missense_ATtv(5)	1:155874285-155874285A/C(1),1:155874287-155874287A/C(1),1:155874287-155874287A/G(1),1:155874287-155874287A/T(1),1:155874285-155874285A/T(1),1:155874286-155874286A/C(1),1:155874286-155874286A/G(1)	0,001544
RAF1	P261	6	Missense_NoCpG_CGts(2),Missense_NoCpG_CGtv(4)	3:12645688-12645688G/A(1),3:12645688-12645688G/T(1),3:12645688-12645688G/C(1),3:12645687-12645687G/C(1),3:12645687-12645687G/A(1),3:12645687-12645687G/T(1)	0,001955
RAF1	S259	6	Missense_ATts(2),Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(2)	3:12645694-12645694A/T(1),3:12645693-12645693G/C(1),3:12645693-12645693G/A(1),3:12645693-12645693G/T(1),3:12645694-12645694A/G(1),3:12645691-12645694TGGA/CCCT(1)	0,002669
RBM20	R636	4	Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(3)	10:112572061-112572061C/A(1),10:112572062-112572062G/A(1),10:112572062-112572064GTA/TTG(1),10:112572062-112572062G/T(1)	0,02337
SOS2	M267	3	Missense_ATts(1),Missense_ATtv(2)	14:50649239-50649239A/T(1),14:50649239-50649239A/G(1),14:50649239-50649239A/C(1)	0,03064
NRAS	G12	5	Missense_NoCpG_CGts(2),Missense_NoCpG_CGtv(3)	1:115258747-115258747C/T(1),1:115258748-115258748C/G(1),1:115258747-115258747C/A(1),1:115258748-115258748C/T(1),1:115258747-115258747C/G(1)	0,0328
CACNA1D	I750	1	Missense_ATtv(1)	3:53764495-53764495A/T(1)	0,03547
ABCC9	R1154	2	Missense_NoCpG_CGts(2)	12:21995261-21995261G/A(1),12:21995260-21995260C/T(1)	0,03634
ABCC9	R1116	2	Missense_NoCpG_CGts(2)	12:21995374-21995374C/T(1),12:21995375-21995375G/A(1)	0,03634
SCN10A	F1400	1	Deletion(1)	3:38751047-38751049TGAA/T(1)	0,03719
CASZ1	V815	1	Deletion(1)	1:10713655-10713671GAGGCTGGGGGTGCCCAC/G(1)	0,03894
ANK2	Q3921	2	Nonsense_NoCpG_CGts(1),Nonsense_NoCpG_CGtv(1)	4:114294507-114294507C/T(1),4:114294506-114294507GC/TT(1)	0,04015
SOS2	T376	3	Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(1)	14:50628269-50628269G/C(1),14:50628270-50628270T/A(1),14:50628269-50628269G/A(1)	0,04245
OPA1	L534	1	Missense_ATtv(1)	3:193364865-193364865T/G(1)	0,04942
NEBL	N418	1	Deletion(1)	10:21129753-21129753AT/A(1)	0,04945

Figura 6.10: *Resultados de la implementación del método 3 para un* $\alpha = 0,05$

Gene	aaPos	MutCount	MutSubtypes	MutPositions	Pvalue
SOS1	R552	8	Missense_ATts(1),Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(4),Deletion(1)	2:39249915-39249915T/C(1),2:39249913-39249913C/G(1),2:39249914-39249914C/A(1),2:39249914-39249914C/G(1),2:39249914-39249914C/T(1),2:39249913-39249913C/A(1),2:39249915-39249915T/A(1),2:39249914-39249916CCTT/C(1)	0,00004249
IDH2	R172	5	Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(3)	15:90631837-90631837C/G(1),15:90631838-90631838C/A(1),15:90631838-90631838C/T(1),15:90631839-90631839T/A(1),15:90631837-90631837C/A(1)	0,0003051
RIT1	F82	7	Missense_ATts(2),Missense_ATtv(5)	1:155874285-155874285A/C(1),1:155874287-155874287A/C(1),1:155874287-155874287A/G(1),1:155874287-155874287A/T(1),1:155874285-155874285A/T(1),1:155874286-155874286A/C(1),1:155874286-155874286A/G(1)	0,001544
RAF1	P261	6	Missense_NoCpG_CGts(2),Missense_NoCpG_CGtv(4)	3:12645688-12645688G/A(1),3:12645688-12645688G/T(1),3:12645688-12645688G/C(1),3:12645687-12645687G/C(1),3:12645687-12645687G/A(1),3:12645687-12645687G/T(1)	0,001955
RAF1	S259	6	Missense_ATts(2),Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(2)	3:12645694-12645694A/T(1),3:12645693-12645693G/C(1),3:12645693-12645693G/A(1),3:12645693-12645693G/T(1),3:12645694-12645694A/G(1),3:12645691-12645694TGGA/CCCT(1)	0,002669
RBM20	R636	4	Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(3)	10:112572061-112572061C/A(1),10:112572062-112572062G/A(1),10:112572062-112572064GTA/TTG(1),10:112572062-112572062G/T(1)	0,02337
SOS2	M267	3	Missense_ATts(1),Missense_ATtv(2)	14:50649239-50649239A/T(1),14:50649239-50649239A/G(1),14:50649239-50649239A/C(1)	0,03064
NRAS	G12	5	Missense_NoCpG_CGts(2),Missense_NoCpG_CGtv(3)	1:115258747-115258747C/T(1),1:115258748-115258748C/G(1),1:115258747-115258747C/A(1),1:115258748-115258748C/T(1),1:115258747-115258747C/G(1)	0,0328
CACNA1D	I750	1	Missense_ATtv(1)	3:53764495-53764495A/T(1)	0,03547
ABCC9	R1154	2	Missense_NoCpG_CGts(2)	12:21995261-21995261G/A(1),12:21995260-21995260C/T(1)	0,03634
ABCC9	R1116	2	Missense_NoCpG_CGts(2)	12:21995374-21995374C/T(1),12:21995375-21995375G/A(1)	0,03634
SCN10A	F1400	1	Deletion(1)	3:38751047-38751049TGAA/T(1)	0,03719
CASZ1	V815	1	Deletion(1)	1:10713655-10713671GAGGCTGGGGGTGCCCAC/G(1)	0,03894
ANK2	Q3921	2	Nonsense_NoCpG_CGts(1),Nonsense_NoCpG_CGtv(1)	4:114294507-114294507C/T(1),4:114294506-114294507GC/TT(1)	0,04015
SOS2	T376	3	Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(1)	14:50628269-50628269G/C(1),14:50628270-50628270T/A(1),14:50628269-50628269G/A(1)	0,04245
OPA1	L534	1	Missense_ATtv(1)	3:193364865-193364865T/G(1)	0,04942
NEBL	N418	1	Deletion(1)	10:21129753-21129753AT/A(1)	0,04945

Figura 6.11: *Resultados de la implementación del método 3 para un* $\alpha = 0,10$. *Parte 1.*

Gene	aaPos	MutCount	MutSubtypes	MutPositions	Pvalue
RIT1	M90	5	Missense_ATts(1),Missense_ATtv(1),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(2)	1:155874261-155874261C/T(1),1:155874261-155874261C/G(1),1:155874261-155874261C/A(1),1:155874263-155874263T/C(1),1:155874263-155874263T/G(1)	0,05166
LARS2	R103	1	Missense_NoCpG_CGts(1)	3:45441810-45441810G/A(1)	0,05199
BRAF	F247	5	Missense_ATts(2),Missense_ATtv(3)	7:140501332-140501332A/G(1),7:140501333-140501333A/C(1),7:140501331-140501331A/C(1),7:140501333-140501333A/G(1),7:140501331-140501331A/T(1)	0,05527
NDUFS1	V253	1	Missense_ATtv(1)	2:207009730-207009730A/C(1)	0,05711
CPT2	P50	1	Missense_CpG_CGtv(1)	1:53662764-53662764C/A(1)	0,05963
KCND3	G371	1	Missense_NoCpG_CGts(1)	1:112329724-112329724C/T(1)	0,06052
FBXL4	R482	1	Missense_NoCpG_CGts(1)	6:99323549-99323549G/A(1)	0,06115
DOLK	Q331	1	Nonsense_NoCpG_CGts(1)	9:131708592-131708592G/A(1)	0,06506
EARS2	Q117	1	Nonsense_NoCpG_CGts(1)	16:23555971-23555971G/A(1)	0,06586
FARS2	Y144	1	Missense_ATts(1)	6:5369234-5369234A/G(1)	0,07022
PDSS2	Q322	1	Nonsense_NoCpG_CGts(1)	6:107531687-107531687G/A(1)	0,07404
HCN4	G482	3	Missense_NoCpG_CGts(2),Missense_NoCpG_CGtv(1)	15:73622060-73622060C/T(1),15:73622060-73622060C/G(1),15:73622059-73622059C/T(1)	0,07439
HCN4	G480	3	Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(2)	15:73622066-73622066C/G(1),15:73622065-73622065C/A(1),15:73622066-73622066C/T(1)	0,07765
MFF	Q64	1	Nonsense_NoCpG_CGts(1)	2:228195493-228195493C/T(1)	0,07914
DHDDS	S213	1	Missense_NoCpG_CGts(1)	1:26784377-26784377G/A(1)	0,07995
TSFM	Q286	1	Nonsense_NoCpG_CGts(1)	12:58190244-58190244C/T(1)	0,0809
BRAF	E501	5	Missense_ATts(1),Missense_ATtv(2),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(1)	7:140477807-140477807C/T(1),7:140477806-140477806T/C(1),7:140477806-140477806T/G(1),7:140477806-140477806T/A(1),7:140477807-140477807C/G(1)	0,08323
GATAD1	S102	1	Missense_ATts(1)	7:92078120-92078120T/C(1)	0,08778
MYOZ2	S48	1	Missense_ATts(1)	4:120072092-120072092T/C(1)	0,08849
PMPCA	G356	1	Missense_NoCpG_CGts(1)	9:139313082-139313082G/A(1)	0,08878
PMPCA	A377	1	Missense_NoCpG_CGts(1)	9:139313299-139313299G/A(1)	0,08878

Figura 6.12: *Resultados de la implementación del método 3 para un* $\alpha = 0,10$. *Parte 2.*

Gene	aaPos	MutCount	MutSubtypes	MutPositions	Pvalue
HCN4	G480	3	Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(2)	15:73622066-73622066C/G(1),15:73622065-73622065C/A(1),15:73622066-73622066C/T(1)	0,07765
MFF	Q64	1	Nonsense_NoCpG_CGts(1)	2:228195493-228195493C/T(1)	0,07914
DHDDS	S213	1	Missense_NoCpG_CGts(1)	1:26784377-26784377G/A(1)	0,07995
TSFM	Q286	1	Nonsense_NoCpG_CGts(1)	12:58190244-58190244C/T(1)	0,0809
BRAF	E501	5	Missense_ATts(1),Missense_ATtv(2),Missense_NoCpG_CGts(1),Missense_NoCpG_CGtv(1)	7:140477807-140477807C/T(1),7:140477806-140477806T/C(1),7:140477806-140477806T/G(1),7:140477806-140477806T/A(1),7:140477807-140477807C/G(1)	0,08323
GATAD1	S102	1	Missense_ATts(1)	7:92078120-92078120T/C(1)	0,08778
MYOZ2	S48	1	Missense_ATts(1)	4:120072092-120072092T/C(1)	0,08849
PMPCA	G356	1	Missense_NoCpG_CGts(1)	9:139313082-139313082G/A(1)	0,08878
PMPCA	A377	1	Missense_NoCpG_CGts(1)	9:139313299-139313299G/A(1)	0,08878
MAP2K1	P124	4	Missense_NoCpG_CGts(2),Missense_NoCpG_CGtv(2)	15:66729163-66729163C/A(1),15:66729163-66729163C/T(1),15:66729162-66729162C/T(1),15:66729162-66729162C/G(1)	0,09931
TMEM43	S358	1	Missense_NoCpG_CGts(1)	3:14183165-14183165C/T(1)	0,09983
TMEM43	E85	1	Missense_NoCpG_CGts(1)	3:14172412-14172412G/A(1)	0,09983
BRAF	L485	5	Missense_ATts(1),Missense_ATtv(1),Missense_NoCpG_CGtv(2),Deletion(1)	7:140477853-140477853C/G(1),7:140477854-140477854A/G(1),7:140477853-140477853C/A(1),7:140477854-140477854A/C(1),7:140477839-140477853AGGTGCTGTCACATTC/A(1)	0,05119

Figura 6.13: *Resultados de la implementación del método 3 para un* $\alpha = 0,10$*. Parte 3.*

6.6. Método 4: Identification of local clusters of mutation hotspots in cancer-related genes and their biological relevance.

6.6.1. Análisis teórico

En este artículo los hotspots mutacionales se definen como residuos aislados de aminoácidos o secuencias que muestran una alta frecuencia mutacional en genes relacionados con el cáncer, pero cuya prevalencia y relevancia biológica no se comprende completamente.

Actualmente se sabe que las mutaciones somáticas suelen estar en los residuos de los aminoácidos, hecho que sugiere que su acumulación en ciertos genes implique alta fuerza de selección. A partir de esta definición, se propone un método de detección de hotspots mutacionales que se basa en el algoritmo de Smith–Waterman llamado MustClustSW. El algoritmo Smith–Waterman se ha usado tradicionalmente para alinear secuencias de nucleotidos a un genoma de referencia, es decir, buscar similitudes entre dos secuencias de genoma [Smith and Waterman, 1981]. Para conseguir este objetivo se busca la submatriz con una similitud máxima a través de dos secuencias empleando programación dinámica. Este algoritmo de alineamiento local de secuencias se basa en dos fases: (1) generación de la matriz de puntuaciones y (2) proceso de *backtracking* desde el punto de valor máximo de la matriz.

El método MustClustSW para la identificación de hotspots también se divide en dos pasos. El primer paso es la discretización de la secuencia, en la cual la cadena de aminoácidos del gen de estudio se convierte en un vector unidimensional. Este vector recibe una puntuación basada en la frecuencia de aparición de las mutaciones en esa posición de aminoácido, convirtiéndolo así en un problema de subconjunto máximo [7]. Para conseguir ese vector, se preprocesan los datos, cuantificando el número de mutaciones en cada lugar de la proteína. Asimismo, en este caso al vector que se usó se le empleó la razón logarítmica en base dos normalizada.

El segundo paso es la implementación del algoritmo SW modificado para la detección de hotspot. En este caso, en lugar de alinear las secuencias para encontrar similitudes, se optimiza la detección de segmentos contiguos con una alta tasa de mutaciones. Para ello se inicializa una matriz de puntuación, donde cada celda se corresponde a una posible posición en el vector de mutaciones, y cada puntuación inicial es la propia de cada posición.

El método de puntuación del algoritmo es el presentado por la Ecuación **??**, el cual calcula una puntuación en función de si en esa posición existen mutaciones o no. Para ello se definen las siguientes variables:

- $R = (r_1, ..., r_L)$ es un vector de tamaño L indicando la presencia o ausencia de mutaciones así como la frecuencia de mutaciones de una posición de un gen G, donde L longitud en aminoácidos de un gen.

[7] **Problema de subconjunto máximo:** Tipo de problema de optimización combinatoria que busca el conjunto más grande que satisfaga las condiciones impuestas[Smith and Waterman, 1981]

- $m_{i,G}$ es el número total de mutaciones en la posición i-ésima del gen G.
- M_G es el número total de mutaciones en el gen G.
- N_G es el número total de posiciones no mutadas del gen G.

Y las ecuaciones de puntuación son, para residuos mutados la Ecuación 6.21 y para los residuos no mutados la Ecuación 6.22 .

$$+\frac{m_{i,G}}{M_G} \tag{6.21}$$

$$-\frac{1}{N_G} \tag{6.22}$$

Tras conocer esto, se completa la matriz de puntuaciones atendiendo a la ecuación 6.23. Después se extrae el valor $d = -1/L$ para que la suma de los scores sea -1.

$$H_{ij} = \text{máx} \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ \text{máx}_{k \geq 1}\{H_{i-k,j} - W_k\}, \\ \text{máx}_{l \geq 1}\{H_{i,j-l} - W_l\}, \\ 0 \end{cases} \quad (1 \leq i \leq n, 1 \leq j \leq m) \tag{6.23}$$

En este caso:

- $H_{ij} + s(a_i, b_j)$ es la puntuación de alinear a_i y b_j.
- a_i es la posición i - ésima de la secuencia i.
- b_j es la posición j - ésima de la secuencia j.
- W_k es el peso del tamaño de deleción de tamaño k en la secuencia.
- W_l es el peso del tamaño de deleción de tamaño l.

Tras esa inicialización, empieza la evaluación de las puntuaciones acumulativas de los subsegmentos contiguos para así encontrar la región con mayor concentración de mutaciones definida como $[a_{\max}, b_{\max}]$. La fórmula que se emplea es la expuesta en la Ecuación 6.23, la cual pretende maximizar la suma de las puntuaciones de mutaciones. La primera representa si las dos secuencias estan asociadas, la segunda si hay una deleción en la primera secuencia, la tercera si hay una deleción en la segunda secuencia y la cuarta indica que no hay relación entre las dos secuencias.

Tras el cálculo de la puntuación, se realiza un *backtracking* para encontrar los límites del segmento que constituye un hotspot. El *backtracking* es una estrategia algorítmica que se encarga de buscar todas las posibles soluciones dadas unas variables iniciales para encontrar un resultado. Este tipo de técnicas se apoyan en la recursividad para hacer esa búsqueda exhaustiva de todas las posibles soluciones [Peleato, 2024].

La búsqueda recursiva sirve para distinguir múltiples hotspots mutacionales que se han identificado juntos por estar muy cerca. En este modelo, los segmentos de alta puntuación se reevalúan y se extraen de nuevo los vectores R, M y L de forma que se denominan R' , M' y L'. Si la nueva región definida excede la puntuación del segmento que se está estudiando, se redefine el hotspot con los nuevos límites.

$$\left(\frac{S'_{\max}}{\sqrt{b'_{\max} - a'_{\max} - 1}}\right) > \left(\frac{S_{\max}}{\sqrt{b_{\max} - a_{\max} - 1}}\right) \rightarrow |a'_{\text{máx}}, b'_{\text{máx}}| \tag{6.24}$$

Donde:

- $S_{\max}$ es la puntuación máxima de la región que se está reevaluando.
- $S'_{\max}$ es la nueva puntuación máxima de la región tras la reevaluación.
- $a_{\max}$ y $b_{\max}$ son los límites de la región de la secuencia que se está reevaluando.
- $a'_{\max}$ y $b'_{\max}$ son los nuevos límites de la región que se está reevaluando.

En el vector R, los hotspots mutacionales se identifican como segmentos locales de alta puntuación cuya suma de puntuaciones no puede aumentarse mediante la reducción o expansión de los límites del segmento. Si hay más de un hotspot, al identificar el primero se pone el contaje a cero y luego se vuelve a iterar. Esa iteración continúa hasta que no se identifiquen segmentos positivos y el pseudo-código asociado a esta estrategia se describe en el Algoritmo 1.

Algorithm 1 Pseudcódigo de MustClustW para un gen G de tamaño L

```
k = 1
repeat
    S = 0, a = 1, S_max = 0
    for i = 1 to L do
        if la posición i tiene j mutaciones y no está enmascarada then
            S = S + j/M − 1/L
        else if la posición i no tiene mutaciones y no está enmascarada then
            S = S − 1/N − 1/L
        end if
        if S > S_max then
            S_max = S, a_max = a, b_max = i
        end if
        if S < 0 then
            S = 0, a = i + 1
        end if
    end for
    Reportar S_max, a_max, b_max
    Enmascarar la posición i desde a_max hasta b_max
    k = k + 1
until S_max = 0
```

En este caso, el segmento $[a_{max}, b_{max}]$ se reporta como un hotspot.

Tras aplicar esta técnica e identificar los hotspots, se analizan para determinar si esa región representa un verdadero hotspot o simplemente es el resultado de un sesgo. En este apartado se incluyen las penalizaciones por *gaps* y la redefinición de los límites de cada segmento además de aplicar filtros estadísticos para asegurar la no aleatoriedad de los hotspots.

El cálculo del p.valor en este caso se emplea para la evaluación de la significancia de la puntuación del segmento comparándolo con una distribución de puntuaciones obtenidos aleatoriamente. Para conseguir este cometido primero se calcula la puntuación normalizada o NES del segmento actual. Esto se consigue restando la media y dividiendo por la desviación estándar de las puntuaciones aleatorios. Gracias a esta normalización se consigue proporcionar una medida de cuántas desviaciones estándar por encima o por debajo se encuentra la puntuación que se está estimando. Tras esto, el p - valor se obtiene como la proporción de puntuaciones aleatorias que son iguales o mayores que la puntuación observada ajustada por el número total de permutaciones más uno. Esto indica la probabilidad de obtener una puntuación igual o mayor al observado.

El cálculo del p-valor puede verse sesgado por la elevada cantidad de evlauaciones de segmentos que se realizan en una secuencia, dando lugar a posibles acumulaciones de falsos positivos. Por este motivo se normalizan los p-valores calculados usando el paquete locfdr [T. P. B. Smith, 2021] y, tras esto, se opta por la técnica de Benjamini - Hochberg (Ecuación 6.11) para la corrección de los mismos. Finalmente, se consideran hotspots significativos a los segmentos que presentan un FDR $< 0{,}05$.

6.6.2. Adaptación de los datos

En este caso, se requiere de dos archivos: (1) un archivo de mutaciones por tipo con el símbolo del gen y la posición de aminoácido donde se encuentra la mutación y (2) un archivo con la longitud de gel gen en aminoácido.

Para generar el archivo de mutaciones partiendo de la base de datos de entrada, primero se extrae la información de las mutaciones *missense* y *nonsense*. Esto es debido a que el método solo se implementa para estos tipos y el umbral empleado es diferente para ambos tipos. Por este motivo, es importante generar dos archivos de mutaciones, uno para las mutaciones *missense* y otro para las mutaciones *nonsense*.

Después se seleccionan las columnas de interés de nuestra base de datos general de entrada, estudiando solo las columnas Mutation_subtype, Protein_variant y Gene_ID. Tras ello, se ajusta la información para que tenga la estructura que precisa el modelo y el código empleado para hacerlo se muestra en el Anexo 1.10 . De igual forma, se emplea el archivo gene_length.tsv empleado en el método anterior para obtener la información de la longitud de los genes.

6.6.3. Implementación del método

Los resultados de esta tabla se muestran en forma de archivo .txt de 8 variables cuya descripción se muestra a continuación:

- **Gene:** Gen donde hay un hotspot.
- **Start:** Inicio de la región donde hay un hotspot.
- **End:** Fin de la región donde hay un hotspot.
- **Scoresum:** Puntuación Smith-Waterman.
- **Recursion:** Y o N, dependiendo de si se ha recurrido a la recursión o no.
- **Mutation:** Número de mutaciones encontradas.
- **NES:** Valor estandarizado de la puntuación de Smith-Waterman.
- **Pvalue:** p-valor asociado.

Una vez definidas las variables, se exponen los resultados encontrados para distintos umbrales del p-valor. La Tabla 6.6 muestra los 47 hotspots obtenidos para un nivel de significación de 0,05 y en la Tabla 6.7 y en la Tabla 6.8 los 57 hotspots encontrados con un nivel de significación de 0,10.

Gene	Start	End	Scoresum	Recursion	Mutation	NES	Pvalue
RBM20	634	638	0,68741703	Y	9	11,4936299	0
NF1	1441	1447	0,1042055	Y	8	3,61169526	0
NF1	777	784	0,08998043	Y	7	2,58828476	0,01980198
KCNQ1	302	322	0,15053564	Y	39	15,4780066	0
KCNQ1	258	262	0,04996721	Y	12	2,72399007	0,01980198
KCNQ1	338	345	0,06207479	Y	15	4,25946394	0
HCN4	480	482	0,49750831	Y	8	6,49126953	0
ABCC6	1114	1138	0,24638042	Y	5	2,91174133	0,01980198
MAP2K2	56	64	0,42630685	Y	6	4,93534338	0
MAP2K2	126	134	0,34938377	Y	5	3,34483353	0,00990099
MYBPC3	490	502	0,20405166	Y	6	3,05501919	0,00990099
KCNJ2	299	302	0,13351135	Y	7	2,72428247	0,00990099
GATA4	292	296	0,31739549	Y	5	3,22420274	0
SCN5A	1763	1795	0,09199538	Y	9	3,44894089	0
SCN5A	353	376	0,07300186	Y	7	1,84826956	0,04950495
FHL1	150	153	0,39513252	Y	5	3,09751154	0,00990099
TNNT2	94	104	0,29408998	Y	16	10,0865105	0
TNNI3	182	192	0,16996537	Y	11	4,57020038	0
TNNI3	144	146	0,11911532	Y	6	1,87138721	0,03960396
SOS1	548	552	0,20748291	Y	10	7,68879173	0
RYR2	4090	4201	0,14755121	Y	31	20,9567327	0
RYR2	2291	2342	0,05316186	Y	12	4,53555072	0
RYR2	4742	4772	0,04940335	Y	10	3,88167222	0
RYR2	169	176	0,03978002	Y	7	2,20747586	0,03960396
RYR1	4804	4937	0,21262168	Y	49	38,6989338	0
RYR1	4631	4640	0,06089392	Y	12	7,42658098	0
PTPN11	56	63	0,16915299	Y	19	14,5647072	0
PTPN11	498	510	0,16490916	Y	20	14,0733094	0
PTPN11	69	73	0,09837861	Y	11	6,36965527	0
PTPN11	279	285	0,06821406	Y	9	2,87686483	0
MAP2K1	122	130	0,36396475	Y	10	6,65464244	0
NRAS	12	13	0,35789474	Y	7	5,29083199	0
NRAS	58	61	0,29473684	Y	6	3,68240464	0
RAF1	256	263	0,6448162	Y	23	22,1032848	0
BRAF	463	501	0,28048568	Y	28	19,5322552	0
BRAF	599	601	0,10135182	Y	8	4,36973929	0
BRAF	241	247	0,15917891	Y	13	9,26442664	0
BRAF	594	597	0,07373224	Y	6	2,03192244	0,02970297
KCNH2	609	635	0,1723585	Y	26	19,9312782	0
KCNH2	558	605	0,16396947	Y	30	18,7204444	0
KCNH2	28	58	0,07861294	Y	16	6,40047593	0
DSP	597	622	0,27736434	Y	5	3,20569669	0
RIT1	81	84	0,34848485	Y	11	7,62828007	0
RIT1	89	90	0,19090909	Y	6	2,55105281	0,01980198
LZTR1	230	248	0,42205849	Y	6	4,76309995	0
SOS2	264	267	0,59624624	Y	6	5,35773724	0
TCAP	15	25	0,39077381	Y	4	1,94772115	0,01980198

Tabla 6.6: *Resultados del método 4 para un nivel de confianza del 95 %.*

Gene	Start	End	Scoresum	Recursion	Mutation	NES	Pvalue
RBM20	634	638	0,68741703	Y	9	11,4936299	0
NF1	1441	1447	0,1042055	Y	8	3,61169526	0
NF1	777	784	0,08998043	Y	7	2,58828476	0,01980198
KCNQ1	302	322	0,15053564	Y	39	15,4780066	0
KCNQ1	258	262	0,04996721	Y	12	2,72399007	0,01980198
KCNQ1	241	243	0,03990366	Y	9	1,44773761	0,06930693
KCNQ1	338	345	0,06207479	Y	15	4,25946394	0
KCNQ1	272	284	0,04096034	Y	14	1,58174453	0,05940594
HCN4	480	482	0,49750831	Y	8	6,49126953	0
ABCC6	1114	1138	0,24638042	Y	5	2,91174133	0,01980198
TBX5	80	85	0,19298953	Y	4	1,42239227	0,06930693
MAP2K2	56	64	0,42630685	Y	6	4,93534338	0
MAP2K2	126	134	0,34938377	Y	5	3,34483353	0,00990099
CAV3	27	33	0,31922515	Y	4	1,39773623	0,08910891
MYBPC3	490	502	0,20405166	Y	6	3,05501919	0,00990099
KCNJ2	299	302	0,13351135	Y	7	2,72428247	0,00990099
KCNJ2	215	218	0,11059692	Y	6	1,64132539	0,07920792
GATA4	292	296	0,31739549	Y	5	3,22420274	0
SCN5A	1763	1795	0,09199538	Y	9	3,44894089	0
SCN5A	353	376	0,07300186	Y	7	1,84826956	0,04950495
SCN5A	1620	1632	0,07004403	Y	6	1,59899981	0,07920792
FHL1	150	153	0,39513252	Y	5	3,09751154	0,00990099
TNNT2	94	104	0,29408998	Y	16	10,0865105	0
TNNI3	182	192	0,16996537	Y	11	4,57020038	0
TNNI3	144	146	0,11911532	Y	6	1,87138721	0,03960396
KRAS	58	60	0,16269841	Y	5	1,43046615	0,0990099
SOS1	548	552	0,20748291	Y	10	7,68879173	0
SOS1	432	434	0,1041341	Y	5	1,85139231	0,06930693
RYR2	4090	4201	0,14755121	Y	31	20,9567327	0
RYR2	2291	2342	0,05316186	Y	12	4,53555072	0
RYR2	4742	4772	0,04940335	Y	10	3,88167222	0
RYR2	3860	3875	0,03692902	Y	7	1,71147924	0,05940594
RYR2	169	176	0,03978002	Y	7	2,20747586	0,03960396
RYR1	4804	4937	0,21262168	Y	49	38,6989338	0
RYR1	4631	4640	0,06089392	Y	12	7,42658098	0
PTPN11	56	63	0,16915299	Y	19	14,5647072	0
PTPN11	498	510	0,16490916	Y	20	14,0733094	0
PTPN11	69	73	0,09837861	Y	11	6,36965527	0
PTPN11	279	285	0,06821406	Y	9	2,87686483	0
MAP2K1	122	130	0,36396475	Y	10	6,65464244	0
NRAS	12	13	0,35789474	Y	7	5,29083199	0
NRAS	58	61	0,29473684	Y	6	3,68240464	0
RAF1	256	263	0,6448162	Y	23	22,1032848	0
BRAF	463	501	0,28048568	Y	28	19,5322552	0
BRAF	599	601	0,10135182	Y	8	4,36973929	0
BRAF	241	247	0,15917891	Y	13	9,26442664	0
BRAF	594	597	0,07373224	Y	6	2,03192244	0,02970297

Tabla 6.7: *Resultados del método 4 con un nivel de confianza del 90 %.Parte 1.*

Gene	Start	End	Scoresum	Recursion	Mutation	NES	Pvalue
KCNH2	609	635	0,1723585	Y	26	19,9312782	0
KCNH2	558	605	0,16396947	Y	30	18,7204444	0
KCNH2	28	58	0,07861294	Y	16	6,40047593	0
DSP	597	622	0,27736434	Y	5	3,20569669	0
RIT1	81	84	0,34848485	Y	11	7,62828007	0
RIT1	89	90	0,19090909	Y	6	2,55105281	0,01980198
RIT1	77	79	0,14815225	N	5	1,17339054	0,0990099
LZTR1	230	248	0,42205849	Y	6	4,76309995	0
SOS2	264	267	0,59624624	Y	6	5,35773724	0
TCAP	15	25	0,39077381	Y	4	1,94772115	0,01980198

Tabla 6.8: *Resultados del método 4 para un nivel de confianza del 90 %.Parte 2*

Capítulo 7

Análisis de resultados

En este capítulo se analizan los resultados obtenidos de los modelos aplicados para identificar hotspots. En la primera sección se definen los parámetros que se van a emplear para evaluar los modelos y en la segunda se presentan y analizan los resultados de estos parámetros.

Los algoritmos de detección de hotspot son métodos de clasificación binaria, ya que se analiza la capacidad de los modelos para identificar las regiones que son hotspots en la base de datos de referencia, CardioHotspots. La capacidad de un modelo para detectar estos hotspots es fundamental para evaluar su precisión y utilidad en aplicaciones prácticas y por tanto es el principal cometido de este capítulo.

7.1. Definición de los parámetros de evaluación

Matriz de confusión

La matriz de confusión es una tabla usada para la evaluación del rendimiento de un modelo predictivo de clasificación [Datasource.ai, 2023]. En este caso tenemos dos clases definidas como hotspot (1) o no hotspot (0). La estructura de una matriz de confusión estándar se puede observar en la Figura 7.5.

		Valores de referencia	
		Positivo	**Negativo**
Valores predichos por modelos	**Positivo**	Verdaderos Positivos VP	Falsos Positivos FP
	Negativo	Falsos Negativos FN	Verdaderos Negativos VN

Figura 7.1: *Estructura general de una matriz de confusión. [Elaboración propia]*

Recapitulando en la definición de hotspot empleada y definida en el Capítulo 4, un hotspot es una región con una alta frecuencia de mutación. Por tanto, en este capítulo se pretende estudiar cuántas de las regiones que vienen definidas en CardioHotspots se detectan por el modelo. Es

por ello que, el enfoque que se plantea es recorrer la base de datos de referencia buscando las regiones definidas por cada uno de los modelos y en base a esto, definir cuatro tipos hotspots:(1) verdaderos positivos o VP, (2) verdaderos negativos o VN, (3) falsos negativos o FN y (4) falsos positivos o FP.

Los verdaderos positivos o VP son los hotspots definidos en la referencia que detecta el modelo. Los verdaderos negativos o VN se consideran a las regiones definidas como no hotspots en la referencia que el modelo no detecta. Los falsos negativos o FN son regiones definidas como hotspots en la referencia que no detecta el modelo. Los falsos positivos o FP se refieren a las regiones detectadas como hotspots por el modelo pero que no aparecen en la referencia. En este punto aparece una consideración que debe puntualizarse. Se considera que existen dos tipos de falsos positivos poniéndoles la etiqueta de asumidos y contrastados.

Por un lado, los falsos positivos asumidos son aquellos hotspots que detecta el modelo pero que no aparecen en la base de datos de referencia. Es importante recalcar que CardioHotspots es una base de datos de hotspots reportados por la literatura y por tanto es posible que no contenga todos los hotspots de cardiopatías existentes. Por otro lado, tenemos los falsos positivos contrastados, que son aquellos que aparecen en la base de referencia como regiones no-hotspot pero que el modelo sí los detecta como hotspots. La diferencia entre ambas categorías es pequeña pero muy importante puesto que tienen naturalezas distintas; la primera no cierra la posibilidad de que esas regiones sean hotspots puesto que asume que la base de datos puede no estar completa mientras que la segunda afirma de forma fidedigna que esas regiones no son hotspots y el modelo sí las considera. En este caso, se consideran para la matriz de confusión a los falsos positivos contrastados puesto que toda la matriz se construye en base a CardioHotspots.

La función de esta matriz reside en el análisis detallado de las predicciones del modelo y de esta surgen diferentes métricas de evaluación: Sensibilidad, especificidad, precisión, exactitud y F1 - score.

Sensibilidad o *recall*

La sensibilidad o *recall* se define como la proporción de verdaderos positivos sobre los casos de referencia positivos [Barrios, 2024]. Esta métrica nos permite evaluar la capacidad discriminativa entre clases del modelo y la fórmula para calcularlo es:

$$\text{Sensibilidad} = \frac{VP}{VP + FN} \tag{7.1}$$

Esta métrica también es conocida como tasa de verdaderos positivos ya que su resultado es la probabilidad de que un resultado positivo real se prediga como positivo [Barrios, 2024]. Ajustando esa definicion al proyecto implica que la sensibilidad mide la probabilidad de detectar correctamente las regiones definidas como hotspots en la base de datos de referencia.

Especificidad o *specificity*

La especificidad o *specificity* se define como la proporción de verdaderos negativos sobre los casos de referencia negativos y la fórmula es:

$$\text{Especificidad} = \frac{VN}{VN + FP} \tag{7.2}$$

Esta métrica también es conocida como tasa de verdaderos negativos ya que su resultado es la probabilidad de que un resultado negativo real se prediga como negativo [Barrios, 2024]. Ajustando esa definición al proyecto implica que la especificidad mide la probabilidad de detectar correctamente las regiones definidas como no - hotspots.

Precisión o *precision*

La precisión o *precision* se define como la proporción de verdaderos positivos dividido entre los resultados positivos del modelo [Barrios, 2024]. Esta métrica mide la dispersión del modelo a partir de mediciones repetidas de una magnitud y su fórmula es:

$$\text{Precisión} = \frac{VP}{VP + FP} \tag{7.3}$$

Exactitud o *accuracy*

La exactitud se define como la proporción de verdaderos positivos sobre todas las predicciones y su fórmula es:

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN} \tag{7.4}$$

En términos estadísticos, esta métrica está relacioanda con el sesgo de la estimación, es decir, la cantidad de predicciones positivas que fueron correctas [Barrios, 2024]. Conforme aumente la precisión, mejor clasifica el modelo los hotspots.

F1 - Score

El F1-Score es una métrica que calcula la media armónica entre la precisión y la sensibilidad [Barrios, 2024], útil para clases desbalanceadas y su fórmula es:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Sensibilidad}}{\text{Precision} + \text{Sensibilidad}} \tag{7.5}$$

En este caso, la base de referencia tiene una distribución de clases 40 % de regiones denominadas hotspots frente al 60 % de regiones denominadas como no - hotspots.

7.2. Comparativa de cada método frente a la referencia

Método 3: Hotspots mutation delineating drives mutational signatures and biological utilities across cancer types.

La matriz de confusión resultante de este método para un nivel de confianza del 95 % es:

		Valores de referencia	
		Hotspot	**No - Hotspot**
Valores predichos por modelos	**Hotspot**	3	2
	No - Hotspot	108	165

Figura 7.2: *Matriz de confusión resultado del análisis del método 3 con un valor de* $\alpha = 0,05$ *[Elaboración propia].*

Y las métricas son:

Sensibilidad	**Especificidad**	**Exactitud**	**Precisión**	**F1 - Score**
0.02727	0.98802	0.60431	0.6	0.05172

Tabla 7.1: *Métricas del análisis del método 3 con un valor de* $\alpha = 0{,}05$.

En la tabla 7.1 se observan unas métricas que muestran un rendimiento razonable en cuanto a precisión y exactitud aunque la sensibilidad es muy baja. Esta métrica indica que el modelo no identifica los hotspots de nuestra base de datos aunque es muy bueno para detectar las regiones clasificadas como no-hotspots.

La matriz de confusión resultante de este método para un nivel de confianza del 90 % es:

		Valores de referencia	
		Hotspot	**No - Hotspot**
Valores predichos por modelos	**Hotspot**	3	3
	No - Hotspot	108	164

Figura 7.3: *Matriz de confusión resultado del análisis del método 3 con un valor de* $\alpha = 0{,}10$ *[Elaboración propia].*

Y las métricas son:

Sensibilidad	**Especificidad**	**Exactitud**	**Precisión**	**F1 - Score**
0.02703	0.98204	0.60072	0.5	0.05128

Tabla 7.2: *Métricas del análisis del método 3 con un valor de* $\alpha = 0{,}10$.

Los resultados de la Tabla 7.2 muestran una disminución de un 10 % en la precisión al aumentar el nivel de significación al 10 %. Esto implica un empeoramiento del modelo con respecto a

la capacidad clasificativa de casos positivos. Sin embargo, la exactitud no ha sufrido muchos cambios aumentando la confianza. Asimismo, la sensibilidad y la especificidad siguen con valores similares al anterior.

Estos resultados muestran que el modelo predictivo de hotspots adaptado al ámbito de las cardiopatías familiares no es muy bueno. Los motivos subyacentes a estos malos resultados pueden deberse a diversos motivos. En primer lugar, la base de datos de referencia es pequeña y, por tanto, no hay muchos datos de evaluación. Esto es debido a que no se publican muchos análisis exploratorios de laboratorios que analizan el genoma buscando estas regiones.

Método 4: Identification of local clusters of mutation hotspots in cancer-related genes and their biological relevance.

La matriz de confusión resultante de este método para un nivel de significación del 5 % es:

		Valores de referencia	
		Hotspot	**No - Hotspot**
Valores predichos por modelos	**Hotspot**	37	13
	No - Hotspot	74	154

Figura 7.4: *Matriz de confusión resultado del análisis del método 4 con un valor de* $\alpha = 0,05$ *[Elaboración propia].*

Y las métricas son:

Sensibilidad	**Especificidad**	**Exactitud**	**Precisión**	**F1 - Score**
0.33333	0.92215	0.68705	0.74	0.45963

Tabla 7.3: *Métricas del análisis del método 4 con un valor de* $\alpha = 0,05$.

Los resultados de la Tabla 7.3 muestran un aumento significativo de las métricas de evaluación comparando este modelo con el anterior. En este caso la precisión es del 74 %, implicando que casi 3 de cada 4 hotspots predichos por el modelo están en la base de datos de referencia. Asimismo la exactitud se encuentra en 68.7 %, significando que el modelo funciona de una forma más equilibrada que el anterior.

Sin embargo, la sensibilidad sigue siendo muy baja, tan solo del 33.33 %. Esto implica que solo uno de cada tres hotspots identificados en CardioHotspots son predichos por este modelo. La especificidad es elevada, de un 92.22 %, indicando que el modelo es capaz de detectar correctamente las regiones denominadas como no-hotspots.

La matriz de confusión resultante de este método para un nivel de significación del 10 % es:

		Valores de referencia	
		Hotspot	**No - Hotspot**
Valores predichos por modelos	**Hotspot**	38	15
	No - Hotspot	73	152

Figura 7.5: *Matriz de confusión resultado del análisis del método 4 con un valor de* $\alpha = 0,10$ *[Elaboración propia].*

Y las métricas son:

Sensibilidad	**Especificidad**	**Exactitud**	**Precisión**	**F1 - Score**
0.34234	0.91017	0.68345	0.71698	0.46341

Tabla 7.4: *Métricas del análisis del método 4 con un valor de* $\alpha = 0,10$.

Los resultados de la Tabla 7.4 muestran una disminución de un 3 % en la precisión al aumentar la confianza en un 5 %. Esto implica un empeoramiento del modelo con respecto a la capacidad clasificativa de casos positivos. Sin embargo, este valor sigue siendo superior al azar. Al analizar la sensibilidad y la especificidad es similar a la que nos brindan los resultados con un nivel de confianza mayor.

Capítulo 8

Conclusiones

En este TFM, se han seleccionado, estudiado, adaptado y aplicado métodos estadísticos para evaluar si una determinada región genómica se puede o no considerar hotspot. Originalmente, los métodos estudiados se acogían únicamente al ámbito clínico oncológico, centrándose así en la detección de hotspots relacionados con el cáncer y, por esto, se precisa una adaptación de los mismos para poder emplearlos para la detección de hotspots relacionados con cardiopatías familiares.

Después de la realización del proyecto, queda patente que la tarea de desarrollar herramientas predictivas para identificar este concepto genético en este área clínica sigue siendo un desafío . Por esta razón, queda evidenciado que el campo requiere una mayor investigación para lograr mejorar y optimizar las metodologías actuales. Estas mejoras se plantean para aumentar la precisión de los hotspots predichos , puesto que las métricas que se han obtenido no han sido satisfacorias. El área de la bioinformática y el análisis de datos genéticos requiere de conocimientos de biología, estadística, informática e ingeniería y, sin alguno de estos cuatro pilares, la tarea de reunir el conocimiento y ser capaces de generar este tipo de herramientas sería imposible. Por esta razón, este Trabajo Final de Máster representa una aportación a los grandes esfuerzos que la comunidad científica está haciendo para aunar estos dominios tan diferentes, pero a la vez tan complementarios.

Desafortunadamente, como ya se ha comentado anteriormente, no existen herramientas bioinformáticas que sean capaces de predecir hotspots en un dominio general, pero sí existen dentro del dominio oncológico. Es por ello por lo que este Trabajo de Final de Máster ha tenido como principal objetivo analizar, adaptar y comparar distintos métodos predictivos de hotspots en el área clínica de las cardiopatías familiares. Para conseguir este cometido se realiza un ciclo empírico de conocimiento enmarcado dentro de la metodología del Design Science para poder darle respuesta.

Debido a que la metodología que se emplea en el trabajo es el Design Science y los objetivos se aclaran como preguntas de investigación (Capítulo 2), a continuación se muestra la respuesta a las preguntas planteadas:

1. Análisis de investigación del problema

1.1. ¿Qué es la genética?

La genética es la rama de la biología encargada de la herencia, incluyendo la interacción ntre genes, las variaciones del ADN y la interacción con otros facores ambientales.

1.2. ¿Qué es un hotspot?

Un hotspot es una secuencia de ADN muy susceptible de ser mutada debido a su inestabilidad inherente, su tendencia al entrecruzamiento desigual o su predisposición química a sustituciones de nucleótidos simples. De forma simple se podría definir como una región del ADN en el que se observan mutaciones con más frecuencia de lo habitual.

1.3. ¿Qué fuentes de datos existen sobre hotspots relacionados con cardiopatías familiares?

La fuente de datos existente que reporta los hotspots de cardiopatías familiares es CardioHotspots.

2. Diseño de investigación e inferencia

2.1. ¿Cómo identifico los métodos predictivos existentes?

A través de una revisión bibliográfica estratégica se identifican los métodos predictivos existentes en la literatura.

2.2. ¿Cómo selecciono los métodos más importantes?

La selección de los métodos más importantes es el resultado de la determinación de unos filtros de cribado (expuesta en el apartado 6.1) de métodos predictivos en función de las necesidades de este trabajo.

2.3. ¿Qué fundamentos teóricos sustentan los métodos?

Cada método se ve apoyado por un fundamento teórico que queda explicado en el Capítulo 6.

2.4. ¿Cómo generalizo la entrada de datos?

Para generalizar la entrada de datos se parte de la misma información para todos los métodos para así garantizar la comparativa entre los resultados, tal y como se muestra en la sección 6.2.

2.5. ¿Cómo aplico estos métodos a las cardiopatías familiares?

Estudiando los requisitos de cada método se puede adecuar cada método a las cardiopatías familiares. Los ítems más importantes a tener en cuenta son el análisis teórico del método y el estudio de los requisitos de los datos de entrada.

3. Ejecución de la investigación

3.1. ¿Cuáles son los resultados?

Los resultados varían en función del método empleado y los resultados se observan en el Capítulo 6 pero se destaca la rpecisión del 75 % observada en el método acuñado como MustClustSW y que se basa en el algoritmo Smith - Waterman.

4. Análisis de resultados

4.1. ¿Qué enfoque se emplea para la comparativa?

El enfoque empleado es el de un problema de clasificación comparándolo con CardioHotspots, base de datos usada como referencia.

4.2. ¿Qué parámetros son clave para la comparativa?

Los parámetros clave son: sensibilidad, especificidad, precisión, exactitud y F1- Score.

4.3. ¿Qué método ofrece mejores parámetros de evaluación con respecto a la base de referencia?

El método que ofrece mejores parámetros con respecto a la base de referencia es el de " Identification of local clusters of mutation hotspots in cancer-related genes and their biological response".

8.1. Limitaciones de la investigación

Al llevar a cabo esta investigación, se han identificado diversas limitaciones que han dificultado la consecución de los objetivos. Una de las principales limitaciones es la acotación del trabajo al dominio de variaciones reportadas públicamente. Esto se deriva en el descarte de métodos predictivos potencialmente buenos por el hecho de obtener sus resultados utilizando datos experimentales de pacientes. El motivo principal de imponer esa restricción reside en la necesidad de encontrar métodos estandarizados e independientes de una muestra de pacientes. Asimismo, la necesidad de tener un marco de datos de entrada común para poder comparar los resultados con una referencia imposibilita el empleo de datos experimentales para algunos modelos y los de repositorios públicos para otras. Sin embargo, ha implicado dejar como válidos únicamente 2 de los 48 métodos existentes.

Otra de las limitaciones de este trabajo ha sido la extracción de los datos de un único repositorio público. Clinvar es la fuente más conocida y la que ofrece mayor cobertura y, por ese motivo, es la seleccionada para este trabajo. Sin embargo, esta no garantiza la concentración de toda la información disponible. Esto es debido a la falta de estandarización y almacenamiento de la información genética. La integración de esa información no es una tarea sencilla y excede los objetivos de este trabajo, por tanto, se escoge solo Clinvar para abordarlos.

Además, otra de las grandes limitaciones del trabajo es la falta de información reportada en la literatura acerca de las regiones que se pretenden predecir. CardioHotspots es una base de datos que recoge información de más de 100 artículos donde se reportan estas regiones y tan solo se han extraído unas 111 regiones denominadas como hotspots y unas 167 denominadas como no

hotspots. Al tener tan poca información denominada como *ground truth*, la evaluación de los resultados ha sido deficiente. Esta limitación se produce como resultado de la falta de análisis de estas regiones cuando se secuencia genoma ya que la definición de este concepto es muy general. Esto también puede deberse a que el concepto de hotspot apareció por primera vez en 2013 y se concretó dentro del área de la genética unos años más tarde. Este hecho denota la juventud del término y, por tanto, limita la capacidad de la comunidad científica para reportar resultados acerca del mismo.

8.2. Trabajo futuro

En esta sección se van a detallar dos posibles líneas de trabajo que no solo garantizan la posible continuidad de esta investigación sino que también establecen guías de seguimiento atendiendo a las limitaciones del mismo.

Una de las posibles líneas de investigación sería realizar este mismo trabajo pero empleando datos de pacientes. Con este nuevo enfoque, se amplía el número de posibles métodos de detección de hotspots a utilizar y permite el acercamiento a la ya conocida medicina de precisión.

Asimismo, otra posible línea de investigación pasaría por la generación de métodos predictivos desde el origen a partir de los datos de Clinvar de los que se dispone. Con esta nueva investigación seríamos capaces de generar un método que no precisara adaptación sino que más bien los datos fueran los que generaran el modelo en función de sus características. Esta línea es interesante debido a la potencia que Clinvar posee en cuanto a cobertura de información. Si bien no garantiza un 100 % de la información disponible, sí constituye el marco más empleado por la comunidad científica actual. Estos nuevos modelos podrían generarse a partir de nuevas asunciones estadísticas o ayudándose de metologías de aprendizaje estadístico o inteligencia artificial.

Bibliografía

[Ackerman et al., 2013] Ackerman, M. J., Marcou, C. A., and Testera, D. J. (2013). Medicina personalizada: diagnóstico genético de cardiopatías/canalopatías hereditarias. *Revista Española de Cardiología (Ed. Impr.)*, pages 298–307.

[ADN, NHGRI, 2024] ADN, NHGRI (2024). Ácido desoxirribonucleico (adn). National Human Genome Research Institute.

[Almeida et al., 2020] Almeida, R. C., Bernardes, J., and Ding, Y. (2020). Computational methods for detecting cancer hotspots. *Computational and Structural Biotechnology Journal*, 18:1856–1866.

[Aminoácido, NHGRI, 2024] Aminoácido, NHGRI (2024). Aminoácido. National Human Genome Research Institute.

[ARN, 2024] ARN, N. (2024). Ácido ribonucleico (arn). National Human Genome Research Institute.

[Baeissa et al., 2017] Baeissa, H., Benstead-Hume, G., Richardson, C., and Pearl, F. (2017). Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. *Oncotarget*, 8.

[Barrios, 2024] Barrios, J. (2024). La matriz de confusión y sus métricas.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple hypothesis testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57:289–300.

[Bolker, 2023] Bolker, B. (2023). *emdbook: Ecological Models and Data in R.* R package version 1.3.12.

[Buisson et al., 2019] Buisson, R., Langenbucher, A., Bowen, D., Kwan, E., Benes, C., Zou, L., and Lawrence, M. (2019). Passenger hotspot mutations in cancer driven by apobec3a and mesoscale genomic features. *Science*, 364:eaaw2872.

[cDNA, NHGRI, 2024] cDNA, NHGRI (2024). Copy dna (cdna). National Human Genome Research Institute.

[Chang et al., 2015] Chang, M., Asthana, S., Gao, P., Lee, B., Chapman, J., Kandoth, C., Gao,

J., Socci, N., Solit, D., Olshen, A., Schultz, N., and Taylor, B. (2015). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology*, 34.

[Chang et al., 2017] Chang, M., Shrestha-Bhattarai, T., Schram, A., Bielski, C., Donoghue, M., Jonsson, P., Chakravarty, D., Phillips, S., Kandoth, C., Penson, A., Gorelick, A., Shamu, T., Patel, S., Harris, C., Gao, J., Sumer, S., Kundra, R., Razavi, P., Bob, T., and Taylor, B. (2017). Accelerating discovery of functional mutant alleles in cancer. *Cancer Discovery*, 8:CD–17.

[Chen et al., 2016] Chen, T., Wang, Z., Zhou, W., Chong, Z., Meric-Bernstam, F., Mills, G., and Chen, K. (2016). Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types. *BMC Genomics*, 17.

[Choi and Henderson, 2015] Choi, K. and Henderson, I. R. (2015). Meiotic recombination hotspots - a comparative view. *Plant Journal*, 83(1):52–61.

[Clarivate, 2024] Clarivate (2024). Journal citation reports.

[Clinvar, 2024] Clinvar, N. (2024). Clinvar database. `https://ftp.ncbi.nlm.nih.gov/pub/clinvar/`. National Center for Biotechnology Information.

[Codon, 2024] Codon, N. (2024). Codon - genetics glossary. National Human Genome Research Institute.

[Consortium, 2022] Consortium, T. U. (2022). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

[Datasource.ai, 2023] Datasource.ai (2023). Métricas de evaluación de modelos en el aprendizaje automático.

[Delecion, 2024] Delecion (2024). Deleción en el diccionario de genética.

[Diaz et al., 2009] Diaz, J., Rojas, J. C. C., and Moreno Laverde, R. (2009). Técnicas de lógica difusa aplicadas a la minería de datos. *Scientia et Technica*, 3(40).

[EBI, 2024] EBI (2024). European Bioinformatics Institute.

[Ensembl, 2024] Ensembl (2024). Uploading bed files.

[Eugenomic, 2024] Eugenomic (2024). Isoforma.

[Fenotipo, NHGRI, 2024] Fenotipo, NHGRI (2024). Fenotipo. National Human Genome Research Institute.

[Fijal et al., 2002] Fijal, B., Idury, R., and Witte, J. (2002). Analysis of mutational spectra: Locating hotspots and clusters of mutations using recursive segmentation. *Statistics in medicine*, 21:1867–85.

[Fokkema et al., 2011] Fokkema, I. F. A., Taschner, P. E. M., Schaafsma, G. C., Celli, J., Laros, J. F. J., and den Dunnen, J. T. (2011). Lovd v.2.0: the next generation in gene variant

databases. *Human Mutation*, 32(5):557–563.

[García Zarzoso, 2023] García Zarzoso, A. (2023). *Diseño y desarrollo de una fuente de datos sobre hotspots asociados al criterio PM1 de las guías ACMG-AMP 2015 aplicado a cardiopatías familiares*. PhD thesis, Universitat Politècnica de València.

[GCECGH, sf] GCECGH ((s.f).). Guía de aplicación clínica de la secuenciación masiva en síndromes mielodisplásicos y leucemia mielomonocítica crónica.

[Genética, NHGRI, 2024] Genética, NHGRI (2024). Genetics - genetics glossary. National Human Genome Research Institute.

[Glazko et al., 1998] Glazko, G., Milanesi, L., and Rogozin, I. (1998). The subclass approach for mutational spectrum analysis: Application of the sem algorithm. *Journal of theoretical biology*, 192:475–87.

[Guo et al., 2018] Guo, Y., Chang, M., Huang, W., Ooi, R., Xing, M., Tan, P., and Jacobsen Skanderup, A. (2018). Mutation hotspots at ctcf binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nature Communications*, 9.

[Hess et al., 2019] Hess, J., Bernards, A., Kim, J., Miller, M., Taylor-Weiner, A., Haradhvala, N., Lawrence, M., and Getz, G. (2019). Passenger hotspot mutations in cancer. *Cancer Cell*, 36:288–301.e14.

[HTSlib, 2016] HTSlib (2016). *VCF Version 4.1 Specification*.

[Human Phenotype Ontology (HP),] Human Phenotype Ontology (HP). The human phenotype ontology. `https://hpo.jax.org/`.

[Hurtado, 2022] Hurtado, C. (2022). Medicina de precisión: conceptos, aplicaciones y proyecciones. *Revista Médica Clínica Las Condes*, 33(1):7–16. TEMA CENTRAL: Medicina de precisión: hacia una terapia individualizada - Parte I.

[Inserción, 2024] Inserción, N. (2024). National Human Genome Research Institute.

[Juul et al., 2021] Juul, R., Nielsen, M., Juul, M., Feuerbach, L., and Pedersen, J. (2021). The landscape and driver potential of site-specific hotspots across cancer genomes. *npj Genomic Medicine*, 6.

[Knaus and Grunwald, 2017] Knaus, B. J. and Grunwald, N. (2017). *vcfR: Manipulate and Visualize VCF Data*. R package version 1.8.0.

[Landrum et al., 2014] Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. (2014). Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985.

[Mapa Genético, NHGRI, 2024] Mapa Genético, NHGRI (2024). Mapa genético. National Human Genome Research Institute.

[McLaren et al., 2016] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann,

A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biology*, 17(1):122.

[MONDO Consortium, sf] MONDO Consortium ((s.f).). Monarch disease ontology (mondo). `https://monarchinitiative.org/monarch-disease-ontology`.

[Mullick et al., 2021] Mullick, B., Magar, R., Jhunjhunwala, A., and Barati Farimani, A. (2021). Understanding mutation hotspots for the sars-cov-2 spike protein using shannon entropy and k-means clustering. *Computers in Biology and Medicine*, 138:104915.

[NCBI, 2005] NCBI (2005). Gene help. Gene Frequently Asked Questions.

[NCBI, 2024] NCBI (2024). Mesh (medical subject headings). National Center for Biotechnology Information.

[NCI, 2024] NCI (2024). Mutación somática. National Cancer Institute.

[NLM(US), NCBI, 2002] NLM(US), NCBI (2002). The ncbi handbook. `http://www.ncbi.nlm.nih.gov/books/NBK21091`. National Library of Medicine (US), National Center for Biotechnology Information.

[Olivé, 2007] Olivé, A. (2007). *Conceptual modeling of information systems*. Springer, Berlin, Heidelberg.

[OMIM, sf] OMIM ((s.f).). Online mendelian inheritance in man, omim®. `https://omim.org/`. Johns Hopkins University, Baltimore, MD.

[Ooms, 2020] Ooms, J. (2020). *jsonlite: A Simple and Robust JSON Parser and Generator for R*. R package version 1.7.1.

[Paul, 2016] Paul, P., N. D. . C. S. (2016). Recombination hotspots: Models and tools for detection. *DNA Repair*, 40:47–56.

[Peleato, 2024] Peleato, J. J. (2024). Backtracking.

[Piraino and Furney, 2017] Piraino, S. and Furney, S. (2017). Identification of coding and non-coding mutational hotspots in cancer genomes. *BMC Genomics*, 18.

[PubMed, NCBI, sf] PubMed, NCBI ((s.f)). Pubmed. `https://pubmed.ncbi.nlm.nih.gov/`.

[Química.es, 2024] Química.es (2024). Islas cpg.

[ReviverSoft, 2024] ReviverSoft (2024). Extensiones de archivo tsv. Accedido: [Fecha de acceso].

[Rhee et al., 2018] Rhee, J.-K., Yoo, J., Kim, K., Kim, J., Lee, Y.-J., Cho, B., and Kim, T.-M. (2018). Identification of local clusters of mutation hotspots in cancer-related genes and their biological relevance. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP:1–1.

[Rheinbay et al., 2017] Rheinbay, E., Parasuraman, P., Grimsby, J., Grace, T., Engreitz, J., Kim, J., Lawrence, M., Taylor-Weiner, A., Rodríguez-Cuevas, S., Rosenberg, M., Hess, J., Stewart, C., Maruvka, Y., Stojanov, P., Cortes, M., Seepo, S., Cibulskis, C., Tracy, A., Pugh,

T., and Getz, G. (2017). Recurrent and functional regulatory mutations in breast cancer. *Nature*, 547.

[Richards et al., 2015] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., and Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5):405–424.

[Rigby et al., 2023] Rigby, M. D., Stasinopoulos, D. M., Gilchrist, R. A., and Voudouris, W. (2023). *GAMLSS: Generalized Additive Models for Location Scale and Shape*. R package version 6.0-2.

[Roche, sf] Roche ((s.f).). Glosario de genética - puntos calientes de mutación. `https://www.institutoroche.es/recursos/glosario/Puntos+calientes+de+mutaci%C3%B3n`. Fundación Instituto Roche.

[Rodrigues, 2013] Rodrigues, A. S. L. (2013). Hotspots. In *Encyclopedia of Biodiversity: Second Edition*, pages 127–136. Elsevier.

[Rogozin and Pavlov, 2003] Rogozin, I. B. and Pavlov, Y. I. (2003). Theoretical analysis of mutation hotspots and their dna sequence context specificity. *Mutation Research/Reviews in Mutation Research*, 544(1):65–85.

[Sayers et al., 2022] Sayers, E. W., Sherry, S. M., Kattman, B. L., Chetvernin, V., Karsch-Mizrachi, I., A., L., L., K., and M., G. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1):D20–D26.

[ScienceDirect, sf] ScienceDirect ((s.f).). Genome assembly.

[S.Garcia et al., 2024] S.Garcia, A., Costa, M., García-Zarzoso, A., and Pastor, O. (2024). Cardiohotspots: a database of mutational hotspots for cardiac disorders. *Database : the journal of biological databases and curation*, 2024.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

[Sollis et al., 2022] Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorff, L., Cunningham, F., Lambert, S. A., Inouye, M., Parkinson, H., and Harris, L. W. (2022). The nhgri-ebi gwas catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 50(D1):D957–D965. Epub ahead of print.

[T. P. B. Smith, 2021] T. P. B. Smith, S. E. J. (2021). *locfdr: Local False Discovery Rate*. R package version 1.0.0.

[Team, 2023] Team, R. C. (2023). *stats: The R Stats Package*. R package version 4.3.0.

[Trevino, 2020] Trevino, V. (2020). Modeling and analysis of site-specific mutations in cancer identifies known plus putative novel hotspots and bias due to contextual sequences. *Computational and Structural Biotechnology Journal*, 18.

[Van den Eynden et al., 2015] Van den Eynden, J., Fierro, A., Verbeke, L., and Marchal, K. (2015). Sominaclust: Detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC bioinformatics*, 16:125.

[Walker, 2020] Walker, A. (2020). *openxlsx: Read, Write and Edit XLSX Files*. R package version 4.1.5.

[Waring et al., 2020] Waring, A., Harper, A., Salatino, S., Kramer, C., Neubauer, S., Thomson, K., Watkins, H., and Farrall, M. (2020). Data-driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomyopathy. *Journal of Medical Genetics*, 58:jmedgenet–2020.

[Wickham, 2017] Wickham, H. (2017). *Tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.3.0.

[Wickham, 2019] Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.

[Wickham and Bryan, 2019] Wickham, H. and Bryan, J. (2019). *readxl: Read Excel Files*. R package version 1.3.1.

[Wickham and Henry, 2020] Wickham, H. and Henry, L. (2020). *tidyr: Tidy Messy Data*. R package version 1.1.2.

[Wieringa, 2014] Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.

[Wong et al., 2022] Wong, J., Aichmüller, C., Schulze, M., Hlevnjak, M., Elgaafary, S., Lichter, P., and Zapatka, M. (2022). Association of mutation signature effectuating processes with mutation hotspots in driver genes and non-coding regions. *Nature Communications*, 13.

[Ye et al., 2010] Ye, J., Pavlicek, A., Lunney, E., Rejto, P., and Teng, C.-H. (2010). Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC bioinformatics*, 11:11.

Apéndice A

Anexos

1.1. Anexo 1: ODS

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ANEXO I. RELACIÓN DEL TRABAJO CON LOS OBJETIVOS DE DESARROLLO SOSTENIBLE DE LA AGENDA 2030

Anexo al Trabajo de Fin de Grado y Trabajo de Fin de Máster: Relación del trabajo con los Objetivos de Desarrollo Sostenible de la agenda 2030.

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**	X			
ODS 4. **Educación de calidad.**				X
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**				X
ODS 9. **Industria, innovación e infraestructuras.**		X		
ODS 10. **Reducción de las desigualdades**.				X
ODS 11. **Ciudades y comunidades sostenibles**.				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**				X

Descripción de la alineación del TFG/TFM con los ODS con un grado de relación más alto.

***Utilice tantas páginas como sea necesario.

A lo largo de este trabajo se potencia de forma evidente la ODS nº3, ya que la misión de este Trabajo Final de Máster es la aportación a la mejora de la medicina de precisión y al diagnóstico precoz de enfermedades a través de la identificación de hotspots relacionados con cardiopatías familiares. A lo largo de todo el trabajo se expone de forma extensa cómo se puede combinar, biología, estadística e informática para generar conocimiento y para extraer información valiosa de los datos genómicos y así contribuir a la detección de hotspots a nivel computacional. Asimismo también se considera una aportación media al ODS nº 9 relativo a la innovación. Actualmente la detección de hotspots sigue siendo un reto científico del cual solo se ha probado con enfermedades del ámbito oncológico. Por tanto, adaptar los métodos existentes al ámbito cardiológico supone una tarea tanto innovativa como necesaria para comenzar a investigar estos métodos en este ámbito clínico.

Editor: Secretaria General / UPV · D.L.: V-5092-2006 · ISSN: 1887-2298 · www.upv.es/secgen

1.2. Anexo 2: Revisión bibliográfica.

Título	ID	Tipo de técnica	Lectura	Fase 2	Motivo de aceptación o rechazo
3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets	28115009	3D	NO		Técnica no adecuada
A site specific model and analysis of the neutral somatic mutation rate in whole - genome cancer data	https://doi.org/10.1186/s12859-018-2141-2	regresión	SI		No cumple los requisitos
A spatial simulation approach to account for ptotein structure when identifying non - random somatic mutation	24990767	3D	NO		Técnica no adecuada
Accelerating discovery of functional mutant alleles in cancer	29247016	posición	SI		Cumple requisitos
Analysis of mutational spectra: locating hotspots and clusters of mutations using recursive segmentation	12111894	clusters	SI		Cumple requisitos
Artificial intelligence based methods for hot spot prediction	https://doi.org/10.1016/j.sbi.2021.11.003	miscelánea	SI		No es un método de interés
Association of mutation signature effectuating processes with mutation hotspots in driver genes and non - coding regions	https://doi.org/10.1038/s41467-021-27792-6	regresión	SI		Cumple requisitos
Cancer3D: understanding cancer mutations through protein structures	25392415	3D	NO		Técnica no adecuada
Chromatin structure-based prediction of recurrent noncoding mutations in cancer.	27723759	clusters	SI		No predice hotspots
Comprehensive assessment of cancer missense muttion clustering in protein structures	26392535	3D	NO		Técnica no adecuada
Data Driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomiopathy	10.1136/jmedgenet-2020-106922	posición	SI		Cumple requisitos
DMCM: a Data-adaptive Mutation Clustering Method to identify cancer-related mutation clusters	30010784	clusters	SI		No cumple todos los requisitos
e-Driver: a novel method to identify protein regions driving cancer.Bioinformatics	25064568	clusters	SI		No predice hotspots
Exome - Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure	27197156	3D	NO		Técnica no adecuada
Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types	27356755	posición	SI		Cumple requisitos
Identification and analysis of mutational hotspots in oncogenes and tumour suppressors	28423505	clusters	SI		Cumple requisitos
Identification of coding and non-coding mutational hotspots in cancer genomes.	28056774	clusters	SI		Cumple requisitos
Identification of Local Clusters of Mutation Hotspots in Cancer-Related Genes and Their Biological Relevance	29993813	clusters	SI		Cumple requisitos
Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations	26691984	clusters	SI		No cumple los requisitos
Identifying mutation specific cancer pathways using a structurally resolved protein interaction network	25592571	3D	NO		Técnica no adecuada
Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity	26619011	posición	SI		Cumple requisitos
LARVA: An integrative framework for large-scale analysis of recurrent variants in noncoding annotations	26304545	clusters	SI		No predice hotspots
Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures	31462496	3D	NO		Técnica no adecuada
Leveraging protein quaternary structure to identify oncogenic driver mutations	27001666	3D	NO		Técnica no adecuada
Modeling and analysis of site-specific mutations in cancer identifies known plus putative novel hotspots and bias due to contextual sequences	32670506	posición	SI		Cumple requisitos
MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis	25348067	clusters	SI		No se atiende a la definición objetivo

Figura A.1: *Análisis de la revisión bibliográfica de la Fase 1 parte 1*

Título	ID	Tipo de técnica	Lectura	Fase 2	Motivo de aceptación o rechazo
Multiscale mutation clustering algorithm identifies pan-cancer mutational clusters associated with pathway-level changes in gene expression	28170390	clústeres	SI		No cumple requisitos
Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers	29670109	clústeres	SI		Cumple requisitos
Mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome	26841357	3D	NO		Técnica no adecuada
Mutational fingerprint induced by the antineoplastic drug chloroethyl - cyclohexilnitrosoure in mammalian cells	[CANCER RESEARCH55, 4658-4663, October 15, 1995]	miscelánea	SI		No se atiende a la definición objetivo
Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate	28362259	posición	SI		No cumple requisitos
OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes	23884480	clústeres	SI		No predice hotspots
OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers	31228182	clústeres	SI		No predice hotspots
OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations	27311963	miscelánea	SI		No cumple requisitos
Passenger Hotspot Mutations in Cancer	31526759	posición	SI		Cumple requisitos
Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features	31249028	miscelánea	SI		Cumple requisitos
Predicting the recurrence of noncoding regulatory mutations in cancer.	27912731	clústeres	SI		No predice hotspots
Protein - structure - guided discovery of functional mutations across 19 cancer types	27294619	3D	NO		Técnica no adecuada
Recurrent and functional regulatory mutations in breast cancer	28658208	clústeres	SI		Cumple requisitos
Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma	28481342	clústeres	SI		No predice hotspots
Recurrent somatic mutations in regulatory regions of human cancer genomes	26053494	clústeres	SI		No predice hotspots
SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering	25903787	clústeres	SI		Cumple requisitos
Statistical method on nonrandom clustering with application to somatic mutation in cancer	20053295	clústeres	SI		Cumple requisitos
The landscape and driver potential of site - specific hotspots acorss cancer genomes	https://doi.org/10.1038/s41525-021-00197-6	miscelánea	SI		Cumple requisitos
The subclass approach for mutational spectrum analysis: application of the SEM Algorithm	0022-5193/98/120475 + 13 $ 25.00/0/jt980668	miscelánea	SI		Cumple requisitos
Theoretical analysis of mutation hotspots and their DNA sequence context specificity	doi:10.1016/S1383-5742(03)00032-2	miscelánea	SI		No hay método predictivo, solo teórico
Understanding mutation hotspots for the SARS -CoV-2 spike protein using Shannon Entropy and K-means clustering	https://doi.org/10.1016/j.compbiomed.2021.104915	miscelánea	SI		Cumple requisitos
Utilizing protein structure to identify non - random somatic mutation	23758891	3D	NO		Técnica no adecuada

Figura A.2: *Análisis de la revisión bibliográfica de la Fase 1 parte 2*

Título	ID	REVISTA	ÍNDICE JCR	DISP CÓDIGO	FASE 3	MOTIVO ACEPTACIÓN/ RECHAZO
Accelerating discovery of functional mutant alleles in cancer	29247016	CANCER DISCOVERY -2018	Q1 (ONCOLOGY)	SI		Cumple todos los requisitos
Analysis of mutational spectra: locating hotspots and clusters of mutations using recursive segmentation	12111894	STATISTICS IN MEDICINE -2002	N/A (MATHEMATICAL AND COMPUTATIONAL BIOLOGY)| Q1 (MEDICAL INFORMATICS) |Q1 (STATISTICS AND PROBABILITY)	NO		No se puede reproducir el método
Association of mutation signature effectuating processes with mutation hotspots in driver genes and non - coding regions	https://doi.org/10.1038/s41467-021-27792-6	NATURE COMMUNICATIONS - 2022	Q1 (MULTIDISCIPLINARY SCIENCES)	SI		Cumple todos los requisitos
Data Driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomiopathy	10.1136/jmedgenet-2020-106922	BMJ OPEN -2020	Q2 (MEDICINE. GENERAL AND INTERNAL)	SI		Cumple todos los requisitos
Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types	27356755	BMC GENOMICS - 2016	Q1 (BIOTECHNOLOGY & APPLIED MICROBIOLOGY) |Q2 (GENETICS & HEREDITARY)	SI		Cumple todos los requisitos
Identification and analysis of mutational hotspots in oncogenes and tumour suppressors	28423505	ONCOTARGET -2017 (SOLO INFO HASTA 2016)	Q1 (CELL BIOLOGY)| Q1 (ONCOLOGY)	NO		No se puede reproducir el método
Identification of coding and non-coding mutational hotspots in cancer genomes.	28056774	BMC GENOMICS -2017	Q1 (BIOTECHNOLOGY & APPLIED MICROBIOLOGY) |Q2 (GENETICS & HEREDITARY)	NO		No se puede reproducir el método
Identification of Local Clusters of Mutation Hotspots in Cancer-Related Genes and Their Biological Relevance	29993813	IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS -2019	Q2 (BIOCHEMICAL RESEARCH METHODS)| Q2 (COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS) |Q1 (MATHEMATICS, INTERDISCIPLINARY APPLICATIONS)| Q1 (STATISTICS & PROBABILITY)	SI		Cumple todos los requisitos
Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity	26619011	NATURE BIOTECHNOLOGY -2016	Q1 (BIOTECHNOLOGY AND APPLIED MICROBIOLOGY)	SI		Código no original
Modeling and analysis of site-specific mutations in cancer identifies known plus putative novel hotspots and bias due to contextual sequences	32670506	COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL	Q1 (BIOCHEMISTRY & MOLECULAR BIOLOGY)	SI		Cumple todos los requisitos
Passenger Hotspot Mutations in Cancer	31526759	CANCER CELL - 2019	Q1 (ONCOLOGY) |Q1 (CELL BIOLOGY)	NO		Es muy complejo y no usa técnicas informáticas para hacer los cálculos. Habría que trasladar los cálculos estadísticos a código

Figura A.3: *Análisis de la revisión bibliográfica de la Fase 2 parte 1*

Título	ID	REVISTA	ÍNDICE JCR	DISP CÓDIGO	FASE 3	MOTIVO ACEPTACIÓN/ RECHAZO
Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features	31249028	SCIENCE -2019	Q1 (MULTIDISCIPLINARY SCIENCES)	SI		Cumple todos los requisitos
Recurrent and functional regulatory mutations in breast cancer	28658208	NATURE -2017	Q1 (MULTIDISCIPLINARY SCIENCES)	NO		Al no haber código habría que generar uno. Usa datos de secuenciación de pacientes reales
SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering	25903787	BMC INFORMATICS -2015	Q1 (MATHEMTICS) | Q2 (BIOTECHNOLOGY AND MICROBIOLOGY)| Q3 (BIOCHEMICAL RESEARCH JOURNALS)	SI		Cumple todos los requisitos
Statistical method on nonrandom clustering with application to somatic mutation in cancer	20053295	BMC BIOINFORMATICS- 2O10	Q2 (BIOCHEMICAL RESEARCH MEATHODS)| Q2(BIOTECHNOLOGY AND APPLIED MICROBIOLOGY)|Q1 (MATHEMATICAL AND COMPUTATIONAL BIOLOGY)	SI		Cumple todos los requisitos
The landscape and driver potential of site - specific hotspots acorss cancer genoems	https://doi.org/10.1038/s41525-021-00197-6	NPJ GENOMIC MEDICINE -2021	Q1 (GENETICS AND HEREDITY)	SI		Cumple todos los requisitos
The subclass approach for mutational spectrum analysis: application of the SEM Algorithm	0022-5193/98/120475 + 13 $ 25.00/0/jt980668	JOURNAL OF THEORETICAL BIOLOGY -1998	Q2(BIOLOGY)| N/A (MATHEMATICAL AND COMPUTATIONAL BIOLOGY)| Q2 (BIOLOGY, MISCELLANEOUS)	NO		Hay web server pero no funciona → no reproducible
Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers	29670109	NATURE COMMUNICATIONS - 2018	Q1 (MULTIDISCIPLINARY SCIENCES)	NO		No hay código
Understanding mutation hotspots for the SARS -CoV-2 spike protein using Shannon Entropy and K-means clustering	https://doi.org/10.1016/j.compbiomed.2021.104915	COMPUTERS IN BIOLOGY AND MEDICINE -2021	Q1(BIOLOGY)| Q1(COMPUTATIONAL SCIENCE, INTERDISCIPLINARY APPLICATIONS)| N/A (BIOLOGY, MISCELLANEOUS)	NO		No se puede reproducir el método ni se puede acceder a la información suplementaria

Figura A.4: *Análisis de la revisión bibliográfica de la Fase 2 parte 2*

TITULO	ID	LENGUAJE	DISP DATOS O ESTRUCTURA	FASE 4	MOTIVO ACEPTACION/RECHAZO
Accelerating discovery of functional mutant alleles in cancer	29247016	R + PYTHON	SI		Cumple todos los requisitos
Association of mutation signature effectuating processes with mutation hotspots in driver genes and non - coding regions	https://doi.org/10.1038/s41467-021-27792-6	R	NO		no input data
Data Driven modelling of mutational hotspots and in silico predictors in hypertrophic cardiomiopathy	10.1136/jmedgenet-2020-106922	BMJ OPEN -2020	NO		Modela un modelo de regresion para cada gen. Necesitariamos info de antes para poder luego mirar. Demasiada información
Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types	27356755	R	SI		Cumple todos los requisitos
Identification of Local Clusters of Mutation Hotspots in Cancer-Related Genes and Their Biological Relevance	29993813	R	SI		Cumple todos los requisitos
Modeling and analysis of site-specific mutations in cancer identifies known plus putative novel hotspots and bias due to contextual sequences	32670506	MUY ALTO	SI		Cumple todos los requisitos
Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features	31249028	MATLAB	NO		no input data
SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering	25903787	R	NO		No se puede reproducir el input
Statistical method on nonrandom clustering with application to somatic mutation in cancer	20053295	R	SI		Cumple todos los requisitos
The landscape and driver potential of site - specific hotspots acorss cancer genoems	https://doi.org/10.1038/s41525-021-00197-6	ALTO	SI		Cumple todos los requisitos

Figura A.5: *Análisis de la revisión bibliográfica de la Fase 3*

TÍTULO	ID	ADAPTABILIDAD DE LOS DATOS	SELECCIÓN	MOTIVO DE SELECCIÓN O DESCARTE
Accelerating discovery of functional mutant alleles in cancer	29247016	SI		Cumple todos los requisitos
Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types	27356755	SI		Cumple todos los requisitos
Identification of Local Clusters of Mutation Hotspots in Cancer-Related Genes and Their Biological Relevance	29993813	SI		Cumple todos los requisitos
Modeling and analysis of site-specific mutations in cancer identifies known plus putative novel hotspots and bias due to contextual sequences	32670506	SI		Cumple todos los requisitos
Statistical method on nonrandom clustering with application to somatic mutation in cancer	20053295	NO		Falta información para adaptar los datos
The landscape and driver potential of site - specific hotspots acorss cancer genoems	https://doi.org/10.1038/s41525-021-00197-6	NO		Imposibilidad de adaptar los datos por su naturaleza

Figura A.6: *Análisis de la revisión bibliográfica de la Fase 4*

1.3. Anexo 3: Descripción de variables de Clinvar.

Variable	Tipo	Descripción
ONC	categórica	Clasificación de oncogenicidad agregada para esta única variante; los valores múltiples están separados por una barra vertical.
ONCINCL	categórica	Clasificación de oncogenicidad para un haplotipo o genotipo que incluye esta variante. Se informa como pares de VariationID:classification; los valores múltiples están separados por una barra vertical .
ONCREVSTAT	categórica	Estado de la revisión de ClinVar sobre la clasificación de oncogenicidad para la ID de variación .
ONCCONF	categórica	Clasificaciones de oncogenicidad conflictivas para esta única variante; los valores múltiples están separados por una barra vertical.
ORIGIN	categórica	Origen del alelo informado a ClinVar. Uno o más de los siguientes valores: 0 - desconocido; 1 - línea germinal; 2 - somático; 4 - heredado; 8 - paterno; 16 - materno; 32 - de novo; 64 - biparental; 128 - uniparental; 256 - no probado; 512 - probado - no concluyente; 1073741824 - otro .
RS	numérica	ID de dbSNP (es decir, número rs) de dbSNP build 155.
SCIDN	categórica	Nombre de enfermedad preferido de ClinVar para el concepto especificado por los identificadores de enfermedad en SCIDISDB.
SCIDNINCL	categórica	Para la variante incluida: nombre de enfermedad preferido de ClinVar para el concepto especificado por los identificadores de enfermedad en SCIDISDBINCL.
SCIDISDB	categórica	Pares de etiquetas y valores de nombres e identificadores de bases de datos de enfermedades enviados para clasificaciones de impacto clínico somático.
SCIDISDBINCL	categórica	Para la variante incluida: pares de etiqueta-valor del nombre de la base de datos de enfermedades y el identificador para las clasificaciones del impacto clínico somático .
SCIREVSTAT	categórica	Estado de la revisión de ClinVar sobre el impacto clínico somático de la Variación ID .
SCI	categórica	Impacto clínico somático agregado para esta única variante; los valores múltiples están separados por una barra vertical .
SCIINCL	categórica	Clasificación del impacto clínico somático para un haplotipo o genotipo que incluye esta variante. Se informa como pares de VariationID:classification; los valores múltiples están separados por una barra vertical .
CLNDN split	categórica	Enfermedad asociada.

Tabla A.1: *Descripción de las variables utilizadas en el análisis [[Clinvar, 2024]].*

Variable	Tipo	Descripción
CHROM	categórica	Número del cromosoma al que pertenece.
POS	numérica	Localización específica de la variación.
ID	categórica	Identificador de la variación interna de Clinvar.
REF	categórica	Base o conjunto de bases nitrogenadas de referencia.
ALT	categórica	Base o conjunto de bases nitrogenadas alternativas.
QUAL	numérica	Calidad. Esta columna siempre se rellena con '.'
FILTER	categórica	Estado del filtro. Esta columna siempre se rellena con '.'
ALLELEID	categórica	Identificador único del alelo.
CLNDN	categórica	Nombre de enfermedad preferido de ClinVar para el concepto especificado por los identificadores de enfermedad en CLNDISDB.
CLNDNINCL	categórica	Solo para variantes incluidas. Nombre de enfermedad preferido de ClinVar para el concepto especificado por los identificadores de enfermedad en CLNDISDBINCL.
CLNDISDB	categórica	Identificadores de los fenotipos en diferentes ontologías (MONDO, MedGen,OMIM, HP y Orphanet).
CLNDISDBINCL	categórica	Solo para variantes incluidas. Pares de etiqueta-valor del nombre y el identificador de la base de datos de enfermedades.
CLNHGVS	categórica	Expresión HGVS de nivel superior (ensamblaje primario, alt o parche).
CLNREVSTAT	categórica	Digital Object Identifier o identificador del artículo.
CLNSIGCONF	categórica	Clasificación de línea germinal conflictiva para esta única variante; los valores múltiples están separados por una barra vertical.
CLNVC	categórica	Tipo de la variación.
CLNVCSO	categórica	Identificador relativo a CLNVC según la ontología SO.
DBVARID	categórica	Adquisiciones nsv desde dbVar para la variante.
GENEINFO	categórica	gen(es) para la variante informada como símbolo genético: NCBI GeneID. El símbolo y el ID del gen están delimitados por dos puntos y cada par está delimitado por una barra vertical.
MC	categórica	Lista separada por comas de consecuencias moleculares en forma de ID de ontología de secuencia | consecuencia$_{m}olecular$.
ONCDN	categórica	Nombre de enfermedad preferido de ClinVar para el concepto especificado por los identificadores de enfermedad en ONCDISDB.
ONCDNINCL	categórica	Para variaciones incluidas: El nombre de la enfermedad usado por Clinvar para el concepto especificado por los identificadores de la enfermedad en ONCDISDBINCL .
ONCDISDB	categórica	Par de etiquetas para el nombre de la base de datos de la enfermedad y el identificador usado para la clasificaicón de oncogenicidad.

Tabla A.2: *Descripción de las variables utilizadas en el análisis [[Clinvar, 2024]].*

1.4. Anexo 4: Código de filtrado de la base de datos de Clinvar.

```
    # 1. Carga de librerias

Se cargan las librerias necesarias para el filtrado

library(vcfR)
library(tidyverse)
library(readxl)
library(stats)
library(jsonlite)
library(tidyr)
library(openxlsx)
library(stringr)


# 2. Carga de inputs para el filtrado

datos\_clinvar= base de datos de clinvar
datos\_vcfR = datos de pacientes
genes = lista de genes de cardio de interés
fenotipos = lista de fenotipos de CLINVAR

datos\_clinvar=read.vcfR('clinvar.vcf')
genes=read\_excel('genes\_cardio.xlsx')

fenotipos = read\_excel("tiposclinvar.xlsx")

fenotipos= fenotipos$fenotipos


Tras la carga hay que preprocesar los inputs

# 3. Preprocesado de los inputs
## 3.1. objetos vcfR

# Definir el número de variantes genómicas por lote
variantes\_por\_lote <- 30000   # Puedes ajustar este valor según tus necesidades

# Calcular el número total de variantes genómicas en el archivo VCF
```

```
total\_variantes <- nrow(datos\_clinvar@fix)

# Calcular el número total de lotes
total\_lotes <- ceiling(total\_variantes / variantes\_por\_lote)

# Crear una lista para almacenar los lotes
lotes <- vector("list", total\_lotes)

# Dividir en lotes
inicio <- 1
for (i in 1total\_lotes) {
  fin <- min(inicio + variantes\_por\_lote - 1, total\_variantes)
  lotes[[i]] <- datos\_clinvar[iniciofin, ]
  inicio <- fin + 1
}

# 4. Filtros

# nos quedamos con los significados de interés
significados.filt<- c("Likely\_pathogenic", "Likely\_pathogenic,\_low\_penetrance",

"Likely\_pathogenic|risk\_factor","Likely\_risk\_allele","Pathogenic",

"Pathogenic/Likely\_pathogenic",

"Pathogenic/Likely\_pathogenic/Pathogenic,\_low\_penetrance",

"Pathogenic/Likely\_pathogenic|other","Pathogenic/Likely\_pathogenic|risk\_factor",

"Pathogenic|association|protective","Pathogenic|other","Pathogenic|protective",

"Pathogenic|risk\_factor")

# nos quedamos solo con los genes de interés
genes.filt <- paste(genes$gene\_name)

# cogemos los fenotipos de interés

fenotipos.filt <- paste(fenotipos)

# 5. Conversión de obejto vcfR a df con filtrado de gen / significación clínica
```

```
# Crear un marco de datos vacío para almacenar los resultados combinados
combined.df.v2 <- data.frame()

# Recorrer cada entrada en la lista lotes
for (i in 1length(lotes)) {
  # Extracción de datos de INFO y pasar a un marco de datos
  df.lote <- INFO2df(lotes[[i]])

  # Extracción de datos de FIX y pasar a un marco de datos
  fix.df.lote <- as.data.frame(getFIX(lotes[[i]]))

  # Combinación de los marcos de datos INFO y FIX
  combined.df.lote <- cbind(fix.df.lote, df.lote)

  ### FILTRADO DE POR SI POR GEN/FENOTIPO/SIGNIFICADO

  # 1 - POR SIGNIFICADO DE VARIANTE
  #combined.df.lote <- combined.df.lote[combined.df.lote$CLNSIG %in%
  significados.filt, ]

  # 2- POR GEN CON LA LISTA DE INTERÉS
  #combined.df.lote$GENEINFO <- sub(".*", "", combined.df.lote$GENEINFO)
  #combined.df.lote <- combined.df.lote[combined.df.lote$GENEINFO %in%

  genes.filt, ]

  # # 3 - POR FENOTIPOS DE LA LISTA DE MESH
  #
  # # combined.df.lote <- combined.df.lote[combined.df.lote$CLNDN %in%

  fenotipos.filt, ]
  # combined.df.lote <- combined.df.lote[combined.df.lote$CLNDN %in%

  fenotipos2.filt, ]
  #
  ### AL ACABAR EL FILTRADO DE CADA LOTE LO UNO, ASÍ FILTRO Y CONVIERTO EN DF

  SOLO UNO
  # Agregar el resultado combinado al marco de datos principal
  combined.df.v2 <- rbind(combined.df.v2, combined.df.lote)
```

```
  rm(df.lote,fix.df.lote,combined.df.lote)
  # print("fin")
}

# Ahora combined.df contendrá todos los datos combinados para cada entrada en lotes


# 6. Filtrado por fenotipo

# Descomponer la columna CLNDN en una lista de fenotipos
combined.df.v2$CLNDN\_split <- strsplit(as.character(combined.df.v2$CLNDN), "\\|")

# Crear una función para verificar si algún fenotipo en CLNDN

está en fenotipos.filt
check\_fenotipo <- function(fenotipos) {
  any(fenotipos %in% fenotipos.filt)
}

# Aplicar la función a cada fila y filtrar el dataframe
combined.df.final <- combined.df.v2[sapply(combined.df.v2$CLNDN\_split,


# 7. Guardar el archivo en excel

write.xlsx (combined.df.final, "clinvarlimpio.xlsx")


#8. Guardarlo en formato vcf


combined.df.final<- filtered\_df2
combined.df.final$INFO <- apply(combined.df.final[, which(names(combined.df.final)

%in%  names(combined.df.final)

[which(names(combined.df.final) == "FILTER") +

1ncol(combined.df.final)])], 1, function(x) paste(x, collapse = "|\_|"))
```

```
write\_vcf <- function(data, file\_name) {
  # Crear encabezado VCF
  header <- c(
    "##fileformat=VCFv4.2",
    "##INFO=<ID=AF,Number=A,Type=Float,Description=\"Allele Frequency\">",
    "#CHROM\tPOS\tID\tREF\tALT\tQUAL\tFILTER\tINFO"
  )

  # Escribir encabezado al archivo
  writeLines(header, con = file\_name)

  # Convertir el data.frame a formato VCF (tabulado)
  write.table(data, file = file\_name, append = TRUE,

  sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)
}

# Llamar a la función para escribir el archivo VCF
write\_vcf(combined.df.final,

"D/TFM\_ORDENADO/2-IMPLEMENTACIONDEMETODOS/3 - ACCELERATING/hotdriverinput.vcf")
```

1.5. Anexo 5: Lista de genes de interés cardiológico.

- ADSL
- AFG3L2
- AGL
- AGPAT2
- AIFM1
- AKAP9
- AKT1
- A2ML1
- AARS
- ABAT
- ABCC6
- ABCC9
- ACADM
- ACADS
- ACADSB
- ACADVL
- ACAT1
- ACO2
- ACTA1
- ACTC1
- ACTN2
- ANK2
- ANKRD1
- ANO5
- APOA1
- ATAD3A
- ATP5A1
- ATP6AP2
- AUH
- BAG3
- ALG6
- ALMS1
- BCS1L
- BRAF
- BSCL2
- BTD
- C10orf2
- C12orf65
- CACNA1C
- CACNA1D
- CACNB2
- CALM1
- CALM2
- CALR
- CALR3
- CASQ2
- CASZ1
- CAV3
- CBL
- CDH2
- COQ2
- COQ5
- COQ6
- COX10
- COX14

- COX15
- COX6B1
- CEP89
- CHRM2
- CNBP
- COG4
- COG5
- COG6
- COL7A1
- DLAT
- DLD
- DMD
- DMPK
- CPT1A
- CPT2
- CRYAB
- CSRP3
- CYCS
- D2HGDH
- DES
- DHDDS
- DSC2
- DSG2
- DSP
- DTNA
- FAH
- ECHS1
- ELAC2
- EMD
- ETFA
- ETFB
- ETFDH
- EYA4
- DNAJB6
- DNAJC19
- DNM1L
- DPAGT1
- DPYD
- GDAP1
- GFER
- GFPT1
- GJA5
- FARS2
- FASTKD2
- FBXL4
- FH
- FHL1
- FHL2
- FHOD3
- FKRP
- FKTN
- FLNC
- FOXRED1
- FXN
- G6PC
- GAA
- GAMT
- GARS
- GATA4
- GATA5
- GATM
- GBA
- GBE1
- GNE
- GNPTAB
- GPD1L
- HRAS
- GRHPR
- GSK3B
- GUSB
- GYG1
- HADH
- HADHA
- HADHB
- HCN4
- HEXA
- HFE
- HIBCH
- HLCS
- HMGCL
- HMGCS2
- GLA
- GLB1
- HSD17B10
- HSPD1
- HTRA2
- HTT
- IARS2
- IDH2
- IDH3B
- IDS
- IDUA
- ILK
- KCNQ1
- LZTR1
- KRAS
- L2HGDH
- LAMA2
- LAMA4
- LAMP2
- LDB3
- LETM1
- LMNA
- JPH2
- JUP
- KCNA5
- KCND3
- KCNE1
- KCNE2
- KCNE3
- KCNH2
- KCNJ2
- KCNJ5
- KCNJ8
- MTFMT
- MAN1B1
- MAP2K1
- MAP2K2
- MEF2C
- MFF
- MFN2
- MGAT2
- MGME1
- MIB1
- MPC1
- MPDU1
- MPV17
- MRPS22
- MYLK2
- MYO6
- MYOM1
- MYOT
- MYPN
- NARS2
- OPA1
- OPA3
- NDUFA2
- NDUFAF2
- NDUFAF5
- NDUFAF6
- NDUFS1
- NDUFS2
- NDUFS3
- NDUFS4
- NDUFS7
- NDUFS8
- NDUFV1
- NDUFV2
- NEBL
- NEXN

- NF1
- NKX2-5
- NOS1AP
- NPC1
- NPC2
- NRAS
- NUBPL
- MYBPC3
- MYH6
- MYH7
- MYL2
- MYL3
- PKD2
- PKP2
- OXCT1
- PANK2
- PC
- PCK1
- PDHA1
- PDHB
- PDHX
- PDSS2
- PGM1
- PHKA1
- PHYH
- PPOX
- PRDM16
- PRKAG2
- PRKCSH
- PSEN1
- PSEN2
- PTPN11
- PTRF
- PUS1
- PLN
- PMM2
- PNPT1
- POLG
- SACS
- SAMHD1
- RAF1
- RBM20
- RIT1
- RRM2B
- RYR1
- RYR2
- SEMA3A
- SERAC1
- SLC6A8
- SGCA
- SGCB
- SGCD
- SGCG
- SHOC2
- SLC19A3
- SLC22A5
- SLC25A10
- SLC25A12
- SLC25A24
- SLC25A3
- SLC25A38
- SLC25A4
- SLC35D1
- SCN10A
- SCN1B
- SCN3B
- SCN4B
- SCN5A
- SCO1
- SCO2
- SDHA
- SDHAF1
- SDHAF2
- SDHB
- SDHC
- SDHD
- SEC23B
- SOS1
- SOS2
- SPG7
- SRD5A3
- TRMU
- COX7B
- KLF10
- ST3GAL3
- WRN
- XK
- YARS
- AARS2
- APTX
- DARS2
- DGUOK
- ETHE1
- GATAD1
- KCNE5
- MARS2
- STAT2
- STT3A
- SUCLA2
- SUCLG1
- MLYCD
- NADK2
- PDK3
- SLC25A19
- MURC
- NDUFA13
- NDUFA4
- NDUFA6
- NDUFA8
- NDUFB1
- NDUFB6
- NDUFC2
- NDUFV3
- SURF1
- SYNE1
- SYNE2
- NDUFB9
- TACO1
- TAZ
- TNNI3K
- TBX20
- TBX5
- TCAP
- XPNPEP3
- ALG11
- CA5A
- MTO1
- ACAD9
- AGK
- COQ4
- TGFB3
- NGLY1
- MYOZ2
- PDLIM3
- POLG2
- TK2
- ATP5E
- ATPAF2
- COA5
- DNA2
- SLC25A1
- KARS
- DPM2
- EARS2
- ISCU
- LRPPRC
- MRPL3
- NDUFA1
- NDUFA10
- NDUFA11

- NDUFA12
- NDUFA7
- NDUFA9
- NDUFAB1
- NDUFAF1
- NDUFAF3
- NDUFAF4
- NDUFB3
- NDUFS6
- NFU1
- ALG8
- DPM1
- TMEM43
- TMEM70
- TMPO
- RARS2
- RMND1
- SLC35A2
- TPK1
- UQCRB
- UQCRQ
- AGXT
- BOLA3
- DOLK
- HCCS
- CARS2
- TNNC1
- TNNI3
- TNNT2
- MRPL44
- NAA10
- SLC19A2
- B4GALT1
- COX4I2
- SLC35A3
- CLPB
- TPM1
- SARS2
- SLC35A1
- YARS2
- ALG12
- ALG13
- ALG2
- TRPM4
- ALG3
- ALG9
- COG1
- COG7
- COG8
- COQ9
- CYC1
- DDOST
- DPM3
- DPYS
- GFM1
- GTPBP3
- GYS1
- GYS2
- HOGA1
- CALM3
- PMPCA
- TSFM
- TTN
- TTR
- LIAS
- LYRM4
- MOGS
- MPI
- MRPS16
- PCK2
- PDP1
- PDSS1
- RFT1
- RPIA
- SFXN4
- SLC25A32
- SLC35C1
- TARS2
- TMEM165
- VARS2
- COX4I1
- COX7A1
- COX7A2
- TUFM
- TUSC3
- TXN2
- TXNRD2
- TYMP
- PGM3
- TRNT1
- HARS2
- UPB1
- LARS2
- ATP6V0A2
- TIMM8A
- UQCRC2
- UQCRH
- VCL
- SMPD1
- SNTA1
- LONP1

1.6. Anexo 6: Lista de fenotipos aceptados.

- Abnormal_aortic_valve_morphology
- Abnormal_cardiovascular_system_morphology
- Abnormal_morphology_of_left_ventricular_trabeculae
- Abnormal_ventricular_septum_morphology
- Abnormality_of_the_cardiovascular_system
- Aborted_sudden_cardiac_death
- ACTA1_gene_related_myopathy

- ACTA1-related_condition
- ACTA1-related_myopathies
- Actin_accumulation_myopathy
- ACTN2-related_disorders
- Acute_myocardial_infarction
- Agenesis_of_corpus_callosum,_cardiac,_ocular,_and_genital_syndrome
- AGK-Related_Disorders
- AGL-related_condition
- ALMS1-related_condition
- Alstrom_syndrome
- Amyloid_Cardiomyopathy,_Transthyretin-related
- Amyloidosis,_cardiac_and_cutaneous
- Andersen_Tawil_syndrome
- ANK2-related_condition
- Aortic_dilatation
- Aortic_valve_disease_1
- Aortic_valve_disease_2
- Arrhythmogenic_cardiomyopathy
- Arrhythmogenic_cardiomyopathy_with_wooly_hair_and_keratoderma
- Arrhythmogenic_right_ventricular_cardiomyopathy
- Arrhythmogenic_right_ventricular_dysplasia,_familial ,_11,_with_mild_palmoplantar_keratoderma_and_woolly_hair
- ARRHYTHMOGENIC_RIGHT_VENTRICULAR_DYSPLASIA,_FAMILIAL,_11, _WITH_OR_WITHOUT_MILD_PALMOPLANTAR_KERATODERMA
- Arrhythmogenic_right_ventricular_dysplasia,_familial,_14
- Arrhythmogenic_right_ventricular_dysplasia,_familial,_15
- Arrhythmogenic_right_ventricular_dysplasia_1
- Arrhythmogenic_right_ventricular_dysplasia_10
- Arrhythmogenic_right_ventricular_dysplasia_11
- Arrhythmogenic_right_ventricular_dysplasia_12
- Arrhythmogenic_right_ventricular_dysplasia_2

- Arrhythmogenic_right_ventricularvdysplasia_5
- Arrhythmogenic_right_ventricular_dysplasia_8
- Arrhythmogenic_right_ventricular_dysplasia_9
- Arrhythmogenic_ventricular_cardiomyopathy
- Arterial_calcification,_generalized,_of_infancy,_2
- Arteriovenous_malformation
- Asymmetric_septal_hypertrophy
- Atrial_conduction_disease
- Atrial_fibrillation
- Atrial_fibrillation,_familial,_1
- Atrial_fibrillation,_familial,_10
- Atrial_fibrillation,_familial,_11
- Atrial_fibrillation,_familial,_13
- Atrial_fibrillation,_familial,_16
- Atrial_fibrillation,_familial,_17
- Atrial_fibrillation,_familial,_3
- Atrial_fibrillation,_familial,_7
- Atrial_fibrillation,_familial,_9
- Atrial_fibrillation,_somatic
- Atrial_septal_defect
- Atrial_septal_defect_1
- Atrial_septal_defect_2
- Atrial_septal_defect_3
- Atrial_septal_defect_4
- Atrial_septal_defect_5
- Atrial_septal_defect_7
- Atrial_standstill_1,_digenic
- Atrioventricular_block
- Atrioventricular_septal_defect,_somatic
- Atrioventricular_septal_defect_4

- Atypical_coarctation_of_aorta
- BAG3-related_condition
- Benign_scapuloperoneal_muscular_dystrophy_with_cardiomyopathy
- Bicuspid_aortic_valve
- Biventricular_noncompaction_cardiomyopathy
- Bradycardia
- Brugada_syndrome
- Brugada_syndrome_(shorter-than-normal_QT_interval)
- Brugada_syndrome_1
- Brugada_syndrome_3
- Brugada_syndrome_4
- Brugada_syndrome_5
- Brugada_syndrome_8
- Brugada_syndrome_9
- CACNA1C-related_condition
- CACNA1C-Related_Disorder
- CACNA1C-Related_Disorders
- CAP-congenital_myopathy_with_arthrogryposis_multiplex_congenita_without_heart_involvement
- Capillary_malformation-arteriovenous_malformation_1
- Cardiac_arrest
- Cardiac_arrhythmia
- Cardiac_arrhythmia,_ankyrin-B-related
- Cardiac_conduction_defect,_nonprogressive
- Cardioencephalomyopathy,_fatal_infantile,_due_to_cytochrome_c_oxidase_deficiency_1
- Cardioencephalomyopathy,_fatal_infantile,_due_to_cytochrome_c_oxidase_deficiency_2
- Cardioencephalomyopathy,_fatal_infantile,_due_to_cytochrome_c_oxidase_deficiency_3
- Cardio-facio-cutaneous_syndrome
- Cardiofaciocutaneous_syndrome_1
- Cardiofaciocutaneous_syndrome_2
- Cardiofaciocutaneous_syndrome_3

- Cardiofaciocutaneous_syndrome_4
- Cardiomyopathy
- Cardiomyopathy,_dilated,_1LL
- Cardiomyopathy,_dilated,_2E
- Cardiomyopathy,_dilated,_with_wooly_hair,_keratoderma,_and_tooth_agenesis
- Cardiomyopathy,_familial_hypertrophic,_23,_with_or_without_ventricular_noncompaction
- Cardiomyopathy,_familial_hypertrophic,_28
- Cardiomyopathy,_familial_restrictive,_1
- Cardiomyopathy,_familial_restrictive,_3
- Cardiomyopathy,_familial_restrictive,_4
- Cardiomyopathy,_familial_restrictive,_5
- Cardiomyopathy-hypotonia-lactic_acidosis_syndrome
- Cardiovascular_phenotype
- Catecholaminergic_polymorphic_ventricular_tachycardia
- Catecholaminergic_polymorphic_ventricular_tachycardia_1
- Catecholaminergic_polymorphic_ventricular_tachycardia_2
- Catecholaminergic_polymorphic_ventricular_tachycardia_4
- CBL-related_disorder
- Central_core_disease,_autosomal_recessive
- Central_core_myopathy
- Centronuclear_myopathy
- Coenzyme_Q10_deficiency
- Coenzyme_Q10_deficiency,_primary,_1
- Coenzyme_Q10_deficiency,_primary,_3
- Coenzyme_q10_deficiency,_primary,_9
- Conduction_disorder_of_the_heart
- Congenital_heart_defects,_multiple_types,_5
- Congenital_heart_disease
- Congenital_heart_disease_(variable)
- Congenital_long_QT_syndrome

- Congenital_multicore_myopathy_with_external_ophthalmoplegia
- Congenital_myopathy
- Congenital_myopathy_2b,_severe_infantile,_autosomal_recessive
- Congenital_myopathy_2c,_severe_infantile,_autosomal_dominant
- Congenital_myopathy_with_fiber_type_disproportion
- Congenital_titinopathy
- Conotruncal_heart_malformations
- Costello_syndrome
- Costello_syndrome,_severe
- CRYAB-related_condition
- Desmin-related_myofibrillar_myopathy
- DHDDS-related_condition
- Dilated_Cardiomyopathy,_Dominant
- Dilated_cardiomyopathy_1A
- Dilated_cardiomyopathy_1AA
- Dilated_cardiomyopathy_1BB
- Dilated_cardiomyopathy_1C
- Dilated_cardiomyopathy_1CC
- Dilated_cardiomyopathy_1D
- Dilated_cardiomyopathy_1DD
- Dilated_cardiomyopathy_1E
- Dilated_cardiomyopathy_1EE
- Dilated_cardiomyopathy_1FF
- Dilated_cardiomyopathy_1G
- Dilated_cardiomyopathy_1GG
- Dilated_cardiomyopathy_1HH
- Dilated_cardiomyopathy_1I
- Dilated_cardiomyopathy_1II
- Dilated_cardiomyopathy_1J
- Dilated_cardiomyopathy_1KK

- Dilated_cardiomyopathy_1L
- Dilated_cardiomyopathy_1M
- Dilated_cardiomyopathy_1NN
- Dilated_cardiomyopathy_1O
- Dilated_cardiomyopathy_1P
- Dilated_cardiomyopathy_1R
- Dilated_cardiomyopathy_1S
- Dilated_cardiomyopathy_1U
- Dilated_cardiomyopathy_1X
- Dilated_cardiomyopathy_1Y
- Dilated_cardiomyopathy_1Z
- Dilated_cardiomyopathy_2A
- Dilated_cardiomyopathy_2B
- Dilated_cardiomyopathy_3B
- Dilated_cardiomyopathy-hypergonadotropic_hypogonadism_syndrome
- Distal_myopathy,_Tateyama_type
- Distal_myopathy_with_posterior_leg_and_anterior_hand_involvement
- DMD-related_condition
- DNAJC19-related_condition
- DSC2-related_condition
- DSP-related_arrhythmogenic_cardiomyopathy
- DSP-related_cardiomyopathy
- DSP-related_condition
- DSP-Related_Disorders
- Duchenne_and_Becker_muscular_dystrophy
- Duchenne_muscular_dystrophy
- Dysplastic_pulmonary_valve
- Dystrophin_deficiency
- Early-onset_myopathy_with_fatal_cardiomyopathy
- EARS2-Related_Disorders

- ECHS1-related_condition
- Effort-induced_polymorphic_ventricular_tachycardia
- Elevated_diastolic_blood_pressure
- Elevated_systolic_blood_pressure
- EMD-related_condition
- Emery-Dreifuss_muscular_dystrophy
- Emery-Dreifuss_muscular_dystrophy_1,_X-linked
- Emery-Dreifuss_muscular_dystrophy_3,_autosomal_recessive
- Emery-Dreifuss_muscular_dystrophy_4,_autosomal_dominant
- Emery-Dreifuss_muscular_dystrophy_5,_autosomal_dominant
- Emery-Dreifuss_muscular_dystrophy_6
- Emery-Dreifuss_muscular_dystrophy_7,_autosomal_dominant
- Fabry_disease,_cardiac_variant
- Familial_atrioventricular_septal_defect
- Familial_cardiomyopathy
- Familial_Hypertrophic_Cardiomyopathy_with_Wolff-Parkinson-White_Syndrome
- Familial_isolated_arrhythmogenic_right_ventricular_dysplasia
- Familial_isolated_dilated_cardiomyopathy
- Familial_thoracic_aortic_aneurysm_and_aortic_dissection
- Fatal_infantile_hypertonic_myofibrillar_myopathy
- Fatal_infantile_mitochondrial_cardiomyopathy
- FLNC-associated_cardiomyopathy
- GATA4-related_condition
- Gillessen-Kaesbach-Nishimura_syndrome
- GNE_myopathy
- HADHA-related_condition
- HADHA-Related_Disorders
- HCN4-related_condition
- HCN4-related_disorder
- Heart,_malformation_of

- Heart_block
- Heart_block,_nonprogressive
- Heart_murmur
- Heart-hand_syndrome,_Slovenian_type
- Holt-Oram_syndrome
- HRAS-related_condition
- Hypertrophic_cardiomyopathy
- Hypertrophic_cardiomyopathy_1
- Hypertrophic_cardiomyopathy_10
- Hypertrophic_cardiomyopathy_11
- Hypertrophic_cardiomyopathy_12
- Hypertrophic_cardiomyopathy_13
- Hypertrophic_cardiomyopathy_14
- Hypertrophic_cardiomyopathy_15
- Hypertrophic_cardiomyopathy_16
- Hypertrophic_cardiomyopathy_17
- Hypertrophic_cardiomyopathy_18
- Hypertrophic_cardiomyopathy_2
- Hypertrophic_cardiomyopathy_20
- Hypertrophic_cardiomyopathy_25
- Hypertrophic_cardiomyopathy_26
- Hypertrophic_cardiomyopathy_3
- Hypertrophic_cardiomyopathy_4
- Hypertrophic_cardiomyopathy_6
- Hypertrophic_cardiomyopathy_7
- Hypertrophic_cardiomyopathy_8
- Hypertrophic_cardiomyopathy_9
- Hypoplastic_left_heart
- Hypoplastic_left_heart_syndrome_2
- Hypoplastic_right_heart_syndrome

- Inappropriate_sinus_tachycardia
- Inborn_mitochondrial_myopathy
- Infantile_hypertrophic_cardiomyopathy_due_to_MRPL44_deficiency
- Interstitial_cardiac_fibrosis
- Intrinsic_cardiomyopathy
- Isolated_Noncompaction_of_the_Ventricular_Myocardium
- Jervell_and_Lange-Nielsen_syndrome
- Jervell_and_Lange-Nielsen_syndrome_1
- Jervell_and_Lange-Nielsen_syndrome_2
- Juvenile_myopathy,_encephalopathy,_lactic_acidosis_AND_stroke
- Kahrizi_syndrome
- KCNH2-related_condition
- KCNH2-related_disorders
- KCNJ2-related_condition
- KCNQ1-related_condition
- KCNQ1-Related_Disorders
- King_Denborough_syndrome
- Kleefstra_syndrome_1
- KRAS-related_condition
- KRAS-related_disorders
- KRAS-related_RASopathy
- Laminopathy
- LCHAD_deficiency_with_maternal_acute_fatty_liver_of_pregnancy
- Left_ventricular_hypertrophy
- Left_ventricular_noncompaction
- Left_ventricular_noncompaction_1
- Left_ventricular_noncompaction_10
- Left_ventricular_noncompaction_2
- Left_ventricular_noncompaction_4
- Left_ventricular_noncompaction_5

- Left_ventricular_noncompaction_7
- Left_ventricular_noncompaction_8
- Left_ventricular_noncompaction_9
- Left_ventricular_noncompaction_cardiomyopathy
- LEOPARD_syndrome_1
- LEOPARD_syndrome_2
- LEOPARD_syndrome_3
- LMNA-associated_condition
- LMNA-related_condition
- LMNA-related_disease
- LMNA-Related_Disorders
- Long_QT_syndrome
- Long_QT_syndrome,_bradycardia-induced
- Long_QT_syndrome_1
- Long_QT_syndrome_1,_recessive
- Long_QT_syndrome_1/2,_digenic
- Long_QT_syndrome_10
- Long_QT_syndrome_11
- Long_QT_syndrome_14
- Long_QT_syndrome_15
- Long_QT_syndrome_16
- Long_QT_syndrome_2
- Long_QT_syndrome_3
- Long_QT_syndrome_3/6,_digenic
- Long_QT_syndrome_5
- Long_qt_syndrome_8
- Long_QT_syndrome_9
- Low-output_congestive_heart_failure
- Malformation_of_the_heart_and_great_vessels
- Mitochondrial_DNA_depletion_syndrome_12A_(cardiomyopathic_type),_autosomal_dominant

- Mitochondrial_DNA_depletion_syndrome_12B_(cardiomyopathic_type),_autosomal_recessive
- Mitochondrial_DNA_depletion_syndrome_14_(cardioencephalomyopathic_type)
- Mitochondrial_encephalomyopathy
- Mitochondrial_hypertrophic_cardiomyopathy_with_lactic_acidosis_due_to_MTO1_deficiency
- Mitochondrial_trifunctional_protein_deficiency_2_with_myopathy_and_neuropathy
- Mitral_regurgitation
- Mitral_valve_prolapse
- Multiminicore_myopathy
- MYBPC3-related_cardiomyopathies
- MYBPC3-related_condition
- MYBPC3-related_disorder
- MYBPC3-Related_Disorders
- MYH7-related_condition
- MYH7-related_disease
- MYH7-Related_Disorders
- MYH7-related_skeletal_myopathy
- MYL2-related_condition
- Myocarditis
- Myofibrillar_myopathy
- Myofibrillar_myopathy_2
- Myofibrillar_myopathy_3
- Myofibrillar_myopathy_4
- Myofibrillar_myopathy_5
- Myofibrillar_myopathy_6
- Myopathy
- Myopathy,_congenital,_with_excess_of_muscle_spindles
- Myopathy,_congenital,_with_structured_cores_and_z-line_abnormalities
- Myopathy,_distal,_6,_adult-onset,_autosomal_dominant
- Myopathy,_lactic_acidosis,_and_sideroblastic_anemia
- Myopathy,_lactic_acidosis,_and_sideroblastic_anemia_1

- Myopathy,_lactic_acidosis,_and_sideroblastic_anemia_2
- Myopathy,_myofibrillar,_12,_infantile-onset,_with_cardiomyopathy
- Myopathy,_myofibrillar,_9,_with_early_respiratory_failure
- Myopathy,_myosin_storage,_autosomal_recessive
- Myopathy,_reducing_body,_X-linked,_childhood-onset
- Myopathy,_reducing_body,_X-linked,_early-onset,_severe
- Myopathy,_RYR1-associated
- Myosin_storage_myopathy
- Nemaline_myopathy
- Nemaline_myopathy_3,_autosomal_dominant_or_recessive
- Neonatal_encephalomyopathy-cardiomyopathy-respiratory_distress_syndrome
- Noncompaction_cardiomyopathy
- Noonan_syndrome
- Noonan_syndrome_1
- Noonan_syndrome_10
- Noonan_syndrome_2
- Noonan_syndrome_3
- Noonan_syndrome_4
- Noonan_syndrome_5
- Noonan_syndrome_6
- Noonan_syndrome_7
- Noonan_syndrome_8
- Noonan_syndrome_9
- Noonan_syndrome_and_Noonan-related_syndrome
- Noonan_syndrome_with_multiple_lentigines
- Noonan_syndrome-like_disorder_with_juvenile_myelomonocytic_leukemia
- Noonan_syndrome-like_disorder_with_loose_anagen_hair
- Noonan_syndrome-like_disorder_with_loose_anagen_hair_1
- PKP2-related_condition
- Polyglucosan_body_myopathy_type_2

- PRDM16-related_congenital_heart_disease
- Premature_ventricular_contraction
- Primary_dilated_cardiomyopathy
- Primary_familial_dilated_cardiomyopathy
- Primary_familial_hypertrophic_cardiomyopathy
- Progressive_familial_heart_block
- Progressive_familial_heart_block,_type_1A
- Progressive_familial_heart_block_type_IB
- Prolonged_QT_interval
- PTPN11_Related_Disorders
- PTPN11-related_condition
- PTPN11-related_disorder
- RAF1-related_condition
- RAF1-related_disorders
- Rafiq_syndrome
- RBM20-related_condition
- Reduced_left_ventricular_ejection_fraction
- Restrictive_cardiomyopathy
- Right_ventricular_cardiomyopathy
- Right_ventricular_hypertrophy
- RIT1-related_condition
- RYR1-related_condition
- RYR1-Related_Disorders
- RYR1-related_myopathy
- RYR2-related_disorder
- SCN5A-related_condition
- SCN5A-related_conditions
- SCN5A-related_disease
- SCN5A-related_disorder
- SCN5A-Related_Disorders

- Secundum_atrial_septal_defect
- Severe_X-linked_mitochondrial_encephalomyopathy
- Short_QT_syndrome
- Short_QT_syndrome_type_1
- Short_QT_syndrome_type_2
- Short_QT_syndrome_type_3
- Sick_sinus_syndrome
- Sick_sinus_syndrome_1
- Sick_sinus_syndrome_2,_autosomal_dominant
- Sinoatrial_node_disorder
- Sinoatrial_node_dysfunction_and_deafness
- Sinus_tachycardia
- SOS1-related_condition
- SOS2-related_condition
- Sudden_cardiac_arrest
- Sudden_cardiac_death
- Supraventricular_tachycardia
- Syncope
- Tachycardia
- Tetralogy_of_Fallot
- Third_degree_atrioventricular_block
- Timothy_syndrome
- Titinopathy
- TNNI3-related_condition
- TNNT2_-related_cardiomyopathies
- Tricuspid_regurgitation
- TRPM4-related_condition
- TTN-related_condition
- TTN-related_disease
- TTN-Related_disorder

- TTN-Related_Disorders
- TTN-related_myopathy
- Two-raphe_bicuspid_aortic_valve
- Vascular_dilatation
- Vascular_malformation
- Ventricular_arrhythmia
- Ventricular_arrhythmias_due_to_cardiac_ryanodine_receptor_calcium_release_deficiency_syndrome
- Ventricular_fibrillation
- Ventricular_fibrillation,_paroxysmal_familial,_type_1
- Ventricular_hypertrophy
- Ventricular_septal_defect
- Ventricular_septal_defect_1
- Ventricular_septal_defect_3
- Ventricular_tachycardia

1.7. Anexo 7: Descripción de las 94 columnas que constituyen un archivo .maf.

1. **Hugo_Symbol** Símbolo HUGO para el gen (los símbolos HUGO siempre están en mayúsculas). "Desconocido"se utiliza para regiones que no corresponden a un gen.
2. **Entrez_Gene_Id** ID del gen Entrez (un número entero). "0"se utiliza para regiones que no corresponden a una región genética o ID de conjunto.
3. **Center** Uno o más centros de secuenciación del genoma que informan la variante.
4. **NCBI_Build** El genoma de referencia utilizado para el alineamiento (GRCh38).
5. **Chromosome** El cromosoma afectado (chr1).
6. **Start_Position** Posición numérica más baja de la variante informada en la secuencia de referencia genómica. Coordenada de inicio de mutación.
7. **End_Position** Posición genómica numérica más alta de la variante informada en la secuencia de referencia genómica. Coordenada de fin de mutación.
8. **Strand** Cadena genómica del alelo reportado. Actualmente, todas las variantes reportarán la cadena positiva '+'.
9. **Variant_Classification** Efecto traslacional del alelo variante

10. **Variant_Type** Tipo de mutación. El TNP (polimorfismo de trinucleótidos) es análogo al DNP (polimorfismo de dinucleótidos) pero para tres nucleótidos consecutivos. ONP (polimorfismo oligonucleótido) es análogo a TNP pero para ejecuciones consecutivas de cuatro o más (SNP, DNP, TNP, ONP, INS, DEL o Consolidado).
11. **Reference_Allele** El alelo de referencia de la cadena positiva en esta posición. Incluye la secuencia eliminada para una eliminación o "para una inserción.
12. **Tumor_Seq_Allele1** Genotipo de datos primarios para la secuenciación (descubrimiento) del tumor alelo 1. Un símbolo "para una eliminación representa una variante. Un símbolo "para una inserción representa un alelo de tipo salvaje. La nueva secuencia insertada para inserción no incluye bases de referencia flanqueantes.
13. **Tumor_Seq_Allele2** Secuenciación (descubrimiento) del tumor alelo 2.
14. **dbSNP_RS** Los rs-ID de la base de datos dbSNP, "novel"si no se encuentra en ninguna base de datos utilizada, o nulo si no hay ningún registro dbSNP, pero se encuentra en otras bases de datos.
15. **dbSNP_Val_Status** El estado de validación de dbSNP se informa como una lista de estados separados por punto y coma. La unión de todos los rs-ID se toma cuando hay múltiples.
16. **Tumor_Sample_Barcode** Código de barras alícuota para la muestra de tumor.
17. **Matched_Norm_Sample_Barcode** Código de barras alícuota para la muestra normal coincidente.
18. **Match_Norm_Seq_Allele1** Datos primarios del genotipo. Alelo 1 de secuenciación normal coincidente. Un símbolo "para una eliminación representa una variante. Un símbolo "para una inserción representa un alelo de tipo salvaje. La nueva secuencia insertada para inserción no incluye bases de referencia flanqueantes (eliminadas en MAF somático).
19. **Match_Norm_Seq_Allele2** Alelo 2 de secuenciación normal coincidente.
20. **Tumor_Validation_Allele1** Datos secundarios de tecnología ortogonal. Genotipado (validación) del tumor para el alelo 1. Un símbolo "para una eliminación representa una variante. Un símbolo "para una inserción representa un alelo de tipo salvaje. La nueva secuencia insertada para inserción no incluye bases de referencia flanqueantes.
21. **Tumor_Validation_Allele2** Datos secundarios de tecnología ortogonal. Genotipado tumoral (validación) para el alelo 2.
22. **Match_Norm_Validation_Allele1** Datos secundarios de tecnología ortogonal. Genotipado normal coincidente (validación) para el alelo 1. Un símbolo "para una eliminación representa una variante. Un símbolo "para una inserción representa un alelo de tipo salvaje. La nueva secuencia insertada para inserción no incluye bases de referencia flanqueantes (eliminadas en MAF somático).

23. **Match_Norm_Validation_Allele2** Datos secundarios de tecnología ortogonal. Genotipado normal coincidente (validación) para el alelo 2 (aclarado en MAF somático).
24. **Verification_Status** La segunda pasada resulta de un intento independiente que utiliza los mismos métodos que la fuente de datos principal. Generalmente reservado para la secuenciación 3730 Sanger.
25. Validation_Status Resultados del segundo pase de la tecnología ortogonal.
26. **Mutation_Status** Una evaluación de la mutación como somática, línea germinal, LOH, modificación postranscripcional, desconocida o ninguna. Los valores permitidos en este campo están restringidos por el valor en el campo Validation_Status.
27. **Sequencing_Phase** Fase de secuenciación TCGA (si corresponde). La fase debe cambiar bajo cualquier circunstancia en que cambien los objetivos bajo consideración.
28. **Sequence_Source** Tipo de ensayo molecular utilizado para producir los analitos utilizados para la secuenciación. Los valores permitidos son un subconjunto de los valores del campo biblioteca_estrategia de SRA 1.5. Este subconjunto coincide con los utilizados en CGHub.
29. **Validation_Method** Las plataformas de ensayo utilizadas para la llamada de validación.
30. **Puntuación** No en uso.
31. **BAM_File** No en uso.
32. **Sequence** Instrumento utilizado para producir datos de secuencia primaria.
33. **Tumor_Sample_UUID** UUID alícuota de GDC para muestra de tumor.
34. **Matched_Norm_Sample_UUID** UUID alícuota de GDC para muestra normal coincidente.
35. **HGVSc** secuencia de codificación de la variante en el formato recomendado por HGVS.
36. **HGVSp** la secuencia de proteínas de la variante en el formato recomendado por HGVS.
37. **HGVSp_Short** igual que HGVSp, pero usando códigos de aminoácidos de 1 letra.
38. **Transcript_ID** transcripción en la que se ha asignado la consecuencia de la variante.
39. **Exon_Number** el número de exón (del número total).
40. **t_depth** lee la profundidad en este locus en el tumor BAM.
41. **t_ref_count** profundidad de lectura que respalda el alelo de referencia en el tumor BAM.
42. **t_alt_count** profundidad de lectura que respalda el alelo variante en el tumor BAM.
43. **n_depth** lee la profundidad en este lugar en BAM normal.
44. **n_ref_count** profundidad de lectura que respalda el alelo de referencia en BAM normal.
45. **n_alt_count** profundidad de lectura que respalda la variante del alelo en BAM normal.

46. **all_effects** una lista delimitada por punto y coma de todos los posibles efectos variantes, ordenados por prioridad.
47. **Allele** la variante del alelo utilizada para calcular la consecuencia.
48. **Gen** ID de conjunto estable del gen afectado.
49. **Función** ID de conjunto estable de la función.
50. **Feature_type** tipo de característica. Actualmente uno de Transcript, RegulatoryFeature, MotifFeature.
51. **Consequence** - tipo de consecuencia de esta variación.
52. **cDNA_position** - posición relativa del par de bases en la secuencia de cDNA.
53. **CDS_position** posición relativa del par de bases en la secuencia codificante.
54. **Protein_position** posición relativa del aminoácido en la proteína.
55. **Amino_Acids** solo se dan si la variación afecta la secuencia codificante de proteínas.
56. **Codons** los codones alternativos con la base variante en mayúsculas.
57. **Existing_variation** identificador conocido de la variación existente.
58. **ALLELE_NUM** número de alelo de la entrada; 0 es referencia, 1 es primera alternativa, etc.
59. **DISTANCE** distancia más corta desde la variante hasta la transcripción.
60. **STRAND** la cadena de ADN (1 o -1) en la que se encuentra la transcripción/característica.
61. **SYMBOL** - el símbolo del gen.
62. **SYMBOL_SOURCE** la fuente del símbolo genético.
63. **HGNC_ID** identificador de gen del Comité de Nomenclatura de Genes de HUGO.
64. **BIOTYPE** - biotipo de transcripción.
65. **CANONICAL** una bandera que indica si la transcripción se indica como transcripción canónica para este gen.
66. **CCDS** el identificador CCDS para esta transcripción, cuando corresponda.
67. **ENSP** el identificador de la proteína Ensembl de la transcripción afectada.
68. **SWISSPROT** - UniProtKB/Swiss-Prot identificador de producto proteico.
69. **TREMBL** - Identificador UniProtKB/TrEMBL de producto proteico.
70. **UNIPARC** - Identificador UniParc de producto proteico.
71. **RefSeq** identificador de RefSeq para esta transcripción.

72. **SIFT** la predicción y/o puntuación de SIFT, ambas dadas como predicción (puntuación).
73. **PolyPhen** la predicción y/o puntuación de PolyPhen.
74. **EXON** el número de exón (del número total).
75. **INTRON** - el número de intrón (del número total).
76. **DOMAINS** la fuente y el identificador de cualquier dominio proteico superpuesto.
77. **GMAF** alelo de no referencia y frecuencia de la variante existente en 1000 genomas.
78. **AFR_MAF** alelo de no referencia y frecuencia de la variante existente en la población africana combinada de 1000 genomas.
79. **AMR_MAF** alelo de no referencia y frecuencia de la variante existente en la población estadounidense combinada de 1000 genomas.
80. **ASN_MAF** alelo de no referencia y frecuencia de la variante existente en la población asiática combinada de 1000 genomas.
81. **EAS_MAF** alelo de no referencia y frecuencia de la variante existente en 1000 genomas combinados en la población de Asia oriental.
82. **EUR_MAF** alelo de no referencia y frecuencia de la variante existente en la población europea combinada de 1000 genomas.
83. **SAS_MAF** alelo de no referencia y frecuencia de la variante existente en 1000 genomas combinados en la población del sur de Asia.
84. **AA_MAF** alelo de no referencia y frecuencia de la variante existente en la población afroamericana del NHLBI-ESP.
85. **EA_MAF** alelo de no referencia y frecuencia de la variante existente en la población europea americana del NHLBI-ESP.
86. **CLIN_SIG** importancia clínica de la variante de dbSNP.
87. **SOMATIC** estado somático de las variaciones existentes.
88. **TUMORTYPE** tipo de tumor de muestra.
89. **Is_Ref** indicador TRUE/FALSE de si Reference_Allele coincide con la referencia hg19.
90. **Ref_Tri** - Contexto de mutación de trinucleótido de referencia.
91. **Amino_Acid_Length** longitud de los aminoácidos de la proteína.
92. **Amino_Acid_Position** para mutaciones que codifican proteínas, codón en el que reside la mutación.
93. **ccf** fracción de células cancerosas calculada.
94. **Master_ID** identificador único de muestra .Tumor_Sample_Barcode, origen de la muestra y tipo de cáncer delimitado por '.'.

1.8. Anexo 8: Código de preprocesado de los datos de entrada del método 1.

```
library(vcfR)
library(readxl)
library(readxl)
library(dplyr)
library(readr)
library(tidyr)
library(httr)
library(jsonlite)
library(openxlsx)
library(tidyverse)
library(stringr)

#####################CARGAMOS LOS DATOS##########################################
load("datossinprocesar.RData")
limpio <- filtered_df2[!is.na(filtered_df2$ProtCh), ]
rm(filtered_df2)
#########################DEFINIMOS EL ENCABEZADO ##############################
header <- c(
    "Hugo_Symbol", "Entrez_Gene_Id", "Center", "NCBI_Build", "Chromosome",
    "Start_Position", "End_Position", "Strand", "Variant_Classification",
    "Variant_Type", "Reference_Allele", "Tumor_Seq_Allele1", "Tumor_Seq_Allele2",
    "dbSNP_RS", "dbSNP_Val_Status", "Tumor_Sample_Barcode",
    "Matched_Norm_Sample_Barcode", "Match_Norm_Seq_Allele1", "Match_Norm_Seq_Allele2",
    "Tumor_Validation_Allele1", "Tumor_Validation_Allele2",
    "Match_Norm_Validation_Allele1", "Match_Norm_Validation_Allele2",
    "Verification_Status", "Validation_Status", "Mutation_Status",
    "Sequencing_Phase", "Sequence_Source", "Validation_Method", "Score",
    "BAM_File", "Sequencer", "Tumor_Sample_UUID", "Matched_Norm_Sample_UUID",
    "HGVSc", "HGVSp", "HGVSp_Short", "Transcript_ID", "Exon_Number",
    "t_depth", "t_ref_count", "t_alt_count", "n_depth", "n_ref_count",
    "n_alt_count", "all_effects", "Allele", "Gene", "Feature", "Feature_type",
    "Consequence", "cDNA_position", "CDS_position",
    "Protein_position", "Amino_acids", "Codons", "Existing_variation",
    "ALLELE_NUM", "DISTANCE", "STRAND", "SYMBOL", "SYMBOL_SOURCE",
    "HGNC_ID", "BIOTYPE", "CANONICAL", "CCDS", "ENSP", "SWISSPROT", "TREMBL",
    "UNIPARC", "RefSeq", "SIFT", "PolyPhen", "EXON", "INTRON", "DOMAINS", "GMAF",
    "AFR_MAF", "AMR_MAF", "ASN_MAF", "EAS_MAF", "EUR_MAF", "SAS_MAF", "AA_MAF",
```

```
    "EA_MAF", "CLIN_SIG", "SOMATIC", "TUMORTYPE","Is_Ref","Ref_Tri",
    "Amino_Acid_Length","Amino_Acid_Position","ccf","Master_ID"
  )

inputacc <- setNames(data.frame(matrix(ncol = length(header), nrow = 7955)), header)

############### COLUMNAS QUE SE CORRESPONDEN TAL CUAL#############################

inputacc$Hugo_Symbol<-limpio$GENEINFO
inputacc$Start_Position<-limpio$Start
inputacc$End_Position<-limpio$Stop
inputacc$Variant_Classification<-limpio$MC_mutation_subtype # NO ESTOY SEGURA
inputacc$Reference_Allele<-limpio$REF
inputacc$CLIN_SIG <- limpio$CLNSIG
inputacc$dbSNP_RS <-limpio$RS
inputacc$Tumor_Seq_Allele1<-limpio$REF
inputacc$Reference_Allele<-limpio$REF
inputacc$Tumor_Seq_Allele2<-limpio$ALT
inputacc$Consequence <-limpio$MC_mutation_subtype
inputacc$HGVSp<-limpio$ProtCh
inputacc$SYMBOL<- limpio$GENEINFO
inputacc$Allele<-limpio$ALT
inputacc$Amino_Acid_Position<-limpio$aapos
inputacc$Chromosome<-limpio$CHROM

###############  COLUMNAS QUE HAY QUE PREPROCESAR ################################

inputacc$Variant_Type<-limpio$CLNVC
inputacc$Variant_Type <- gsub("single_nucleotide_variant", "SNP", inputacc$Variant_Type)
inputacc$Variant_Type <- gsub("Deletion", "DEL", inputacc$Variant_Type)
inputacc$Variant_Type <- gsub("Insertion", "INS", inputacc$Variant_Type)

inputacc$Variant_Type <- gsub("Duplication", "DUP", inputacc$Variant_Type)
inputacc$Variant_Type <- gsub("Indel", "IND", inputacc$Variant_Type)
inputacc$Variant_Type <- gsub("Inversion", "INV", inputacc$Variant_Type)
inputacc$Variant_Type <- gsub("Microsatellite", "MICRO", inputacc$Variant_Type)
inputacc$HGVSp_Short<-limpio$ProtCh

dic_aa <- data.frame(Acronimo = c("Val","Leu","Thr","Lys","Trp", "His", "Phe",
                                  "Ile", "Arg", "Met", "Ala", "Pro", "Gly",
                                  "Ser", "Cys", "Asn", "Gln", "Tyr", "Asp",
```

```
                                  "Glu"),
                      Letra=c("V", "L", "T", "K","W", "H", "F", "I","R","M","A",
                              "P", "G","S", "C","N", "Q","Y", "D","E"))
list_dic_aa <- as.list(dic_aa)
for (i in 1:nrow(dic_aa)) {
  inputacc$HGVSp_Short <-
  str_replace_all(inputacc$HGVSp_Short,dic_aa$Acronimo[i],dic_aa$Letra[i])
}

limpio$tmp <- gsub("p\\.", "", limpio$ProtCh)

limpio$aaref <- substr(limpio$tmp, 1, 3)
limpio$aapos <- substr(limpio$tmp, 4, nchar(limpio$tmp) - 3)
limpio$aaalt <- substr(limpio$tmp, nchar(limpio$tmp) - 2, nchar(limpio$tmp))
for (i in 1:nrow(dic_aa)) {
  limpio$aaref <- str_replace_all(limpio$aaref,dic_aa$Acronimo[i],dic_aa$Letra[i])
  limpio$aaalt <- str_replace_all(limpio$aaalt,dic_aa$Acronimo[i],dic_aa$Letra[i])
}
inputacc$Amino_acids <- paste(limpio$aaref, limpio$aaalt, sep = "/")
inputacc$Protein_position <- paste(limpio$aapos, limpio$aalength, sep = "/")

# INTERMEDIO.R
inputacc$Gene<-limpio$ENSG
inputacc$Entrez_Gene_Id<-limpio$Entrez_Gene_ID
inputacc$HGNC_ID<-limpio$HGNC_ID
inputacc$Amino_Acid_Length <- limpio$aalength
inputacc$Protein_Length<-inputacc$Amino_Acid_Length

############### COLUMNAS QUE NO SON DE INTERÉS Y SE RELLENAN CON "." ############

inputacc$Matched_Norm_Sample_Barcode <-"."
inputacc$Tumor_Validation_Allele1 <-"."
inputacc$Tumor_Validation_Allele2 <-"."
inputacc$Verification_Status <-"."
inputacc$Match_Norm_Seq_Allele1<-"."
inputacc$Match_Norm_Validation_Allele1<-"."
inputacc$Sequencing_Phase <-"."
inputacc$BAM_File <-"."
inputacc$Match_Norm_Seq_Allele2 <-"."
inputacc$Matched_Norm_Sample_UUID <-"."
inputacc$dbSNP_Val_Status <-"."
```

```
inputacc$TUMORTYPE<-"."
inputacc$Master_ID <-"."
inputacc$ccf<-"."
inputacc$Tumor_Sample_Barcode<-"."
inputacc$Tumor_Sample_UUID<-"."
inputacc$Match_Norm_Seq_Allele1<-"."
inputacc$Match_Norm_Validation_Allele2<-"."
inputacc$t_depth<-"."
inputacc$n_alt_count<-"."
inputacc$UNIPARC<-"."
inputacc$AFR_MAF<-"."
inputacc$AMR_MAF<-"."
inputacc$ASN_MAF<-"."
inputacc$EAS_MAF<-"."
inputacc$EUR_MAF<-"."
inputacc$SAS_MAF<-"."
inputacc$AA_MAF<-"."
inputacc$EA_MAF<-"."
inputacc$Validation_Method<-"."
inputacc$Validation_Status<-"."
inputacc$Sequencer<-"."
inputacc$Sequence_Source<-"."
inputacc$t_ref_count<-"."
inputacc$Center<-"."
inputacc$dbSNP_Val_Status<-"."
inputacc$Verification_Status<-"."
inputacc$Mutation_Status<-"."
inputacc$Sequencing_Phase<-"."
inputacc$Score<-"."
inputacc$t_alt_count<-"."
inputacc$n_depth<-"."
inputacc$n_ref_count<-"."
inputacc$n_alt_count<-"."
inputacc$all_effects<-"."
inputacc$CCDS<-"."
inputacc$SIFT<-"."
inputacc$PolyPhen<-"."
inputacc$EXON<-"."
inputacc$INTRON<-"."
inputacc$DOMAINS<-"."
inputacc$GMAF<-"."
```

```
inputacc$SOMATIC<-"."
inputacc$HGVSc<-"."
inputacc$Transcript_ID<-"."
inputacc$Exon_Number<-"."
inputacc$Feature <-"."
inputacc$Feature_type<-"."
inputacc$cDNA_position<-"."
inputacc$CDS_position<-"."
inputacc$Existing_variation<-"."
inputacc$RefSeq<-"."
inputacc$Ref_Tri<-"."
inputacc$Codons<-"."
inputacc$Center<-"."

############### COLUMNAS CURADAS A MANO POR DEFINICIÓN DE LA PROPIA COL #########

inputacc$Strand<-'+'

inputacc$ALLELE_NUM<-1# allele number from input; 0 is reference, 1 is first alternate etc

inputacc$CANONICAL<-'YES'
inputacc$BIOTYPE<-"protein_coding"

inputacc$NCBI_Build<-"GRCh37"

inputacc$Is_Ref <- TRUE

############################## guardarlo ##########################################

# Suponiendo que tu DataFrame se llama df_maf
write.table(inputacc, file =

"D:/TFM_ORDENADO/2-IMPLEMENTACIONDEMETODOS/3 - ACCELERATING/inputaccelerating.maf",

sep = "\t", quote = FALSE, row.names = FALSE)
```

Este es el código empleado para pasos intermedios.

```
geneinfo<- read.delim(
```

```
"D:/TFM_ORDENADO/2-IMPLEMENTACIONDEMETODOS/3 - ACCELERATING/Homo_sapiens.gene_info")

genes_interes <- c(unique(limpio$GENEINFO))

geneinfo <- geneinfo[geneinfo$Symbol %in% genes_interes, ]

library(tidyr)

df_separado <- geneinfo %>%
  separate(dbXrefs, into = paste0("ref", 1:100), sep = "\\|",

  fill = "right", extra = "merge")

df_separado <- df_separado[, colSums(is.na(df_separado)) < nrow(df_separado)]

df_separado <- df_separado %>%
  rename(HGNC_ID = ref2, ENSG = ref3,GENEINFO=Symbol)

limpio <- limpio %>%
  left_join(df_separado %>% select(GENEINFO, HGNC_ID = HGNC_ID), by = "GENEINFO")


limpio <- limpio %>%
  left_join(df_separado %>% select(GENEINFO, ENSG = ENSG), by = "GENEINFO")

limpio <- limpio %>%
  left_join(df_separado %>% select(GENEINFO, Entrez_Gene_ID = GeneID), by = "GENEINFO")

limpio$HGNC_ID <- sub(".*:", "", limpio$HGNC_ID)
limpio$ENSG <- sub(".*:", "", limpio$ENSG)

################################################################################

path<-"D:/TFM_ORDENADO/2-IMPLEMENTACIONDEMETODOS/1 - HOTSPOTS/HotDriver/gene_length.tsv"

length<-read.delim(path,header = FALSE)

length <- length %>%
  rename(GENEINFO = V1, aalength = V2)
```

```
length <- length[length$GENEINFO %in% genes_interes, ]

limpio <- limpio %>%
  left_join(length %>% select(GENEINFO, aalength = aalength), by = "GENEINFO")
```

1.9. Anexo 9: Código de preprocesado de los datos de entrada del método 3.

```
# 0: Carga de librerias

```{r}
library(readxl)
library(dplyr)
library(readr)
library(tidyr)
library(httr)
library(jsonlite)
library(openxlsx)
library(tidyverse)
library(stringr)
```


1: Sacar la información de Protein_variant

1.1: Sacar los GeneID correspondientes a los AlleleID únicos

```{r}
clinvarlimpio <- read_excel("D:/TFM/BIBLIOGRAFIA/FASE4/clinvarlimpio.xlsx")
clinvarlimpio <- clinvarlimpio %>% rename(AlleleID = ALLELEID)
```

Sacamos los alleleID únicos

```{r}
alleleid_unique <- as.data.frame(unique(clinvarlimpio$AlleleID))
alleleid_unique <- alleleid_unique %>% rename(AlleleID = 'unique(clinvarlimpio$AlleleID)')
```
```

1.2: Sacar de hgsv4variation.txt la información que necesitamos

Cargamos hgvs4variation.txt (ftp de clinvar) dividido en chunks (divisionchunks.R),

creamos una lista, nos guardamos el encabezado y aseguramos los tipos de columna.

```{r}
file_hgsv4variation <- "D:/TFM/BIBLIOGRAFIA/FASE4/HOTSPOTS/hgvs4variation.txt/chunks/"
chunks <- list.files(path = file_hgsv4variation, pattern = "chunk_\\d+\\.csv",

full.names = TRUE)

# Leer el encabezado de uno de los archivos chunk para obtener la estructura

de las columnas
header_chunk <- read_csv(chunks[1], n_max = 0, show_col_types = FALSE)
# Asegurar que los tipos de datos sean consistentes
col_types <- cols(
  VariationID = col_double(),
  AlleleID = col_double(),
  .default = col_character()
)
```

Se genera resultado y las columnas de interés

```{r}
resultado <- header_chunk[FALSE, ]

resultado <- resultado %>% mutate (
  GeneID = as.character(GeneID),
  VariationID =as.double(VariationID),
  AlleleID =as.double(AlleleID)
)

# Iterar sobre cada archivo CSV y combinar los datos
for (file_path in chunks) {
  chunk_df <- read_csv(file_path,col_types = col_types, col_names = names(header_chunk),
```

```
  skip = 1)
  # Filtrar por AlleleID que están en alleleid_unique
  filtered_chunk_df <- chunk_df %>% filter(AlleleID %in% alleleid_unique$AlleleID)
  # Almacenar en el DataFrame resultante
  resultado <- bind_rows(resultado, filtered_chunk_df)
  gc() # Garbage collection para liberar memoria
}

```

Diferentes tipos de filtrado

```{r}
resultado3<- resultado %>% filter (Type=="coding")
resultado31 <- resultado3 %>% filter(ProteinChange!="-")
```

1.3: Encontrar la secuencia canónica a través de una consulta a UniProtKB

Primero definimos las funciones que vamos a usar par consultar a uniprotkb

```{r}
get_uniprot_info <- function(gene_name) {
  base_url <- "https://www.ebi.ac.uk/proteins/api/proteins"
  query <- paste0("?offset=0&size=1&gene=", gene_name, "&reviewed=true&taxid=9606")
  url <- paste0(base_url, query)

  response <- GET(url)

  if (status_code(response) == 200) {
    content <- content(response, as = "text")
    json_content <- fromJSON(content, simplifyVector = FALSE)
    if (length(json_content) > 0) {
      return(json_content[[1]])
    } else {
      return(NULL)
    }
  } else {
    stop("Error al acceder a UniProtKB")
  }
}
```

```
# Función para extraer el identificador de la isoforma canónica
extract_canonical_isoform <- function(protein_info) {
  if (is.null(protein_info) || !("comments" %in% names(protein_info))) {
    return(NULL)
  }

  canonical_isoform <- NULL

  for (comment in protein_info$comments) {
    if (is.list(comment) && "type" %in% names(comment) &&

    comment$type == "ALTERNATIVE_PRODUCTS") {
      for (isoform in comment$isoforms) {
        if ("sequenceStatus" %in% names(isoform) &&

        isoform$sequenceStatus == "displayed") {
          canonical_isoform <- isoform$ids[[1]]
        }
      }
    }
  }

  return(canonical_isoform)
}

# Función para extraer el identificador RefSeq correspondiente a la isoforma canónica
extract_refseq_id <- function(protein_info, isoform_id) {
  if (is.null(protein_info) || !("dbReferences" %in% names(protein_info))) {
    return(NULL)
  }

  refseq_id <- NULL

  for (dbReference in protein_info$dbReferences) {
    if (dbReference$type == "RefSeq" && "isoform" %in% names(dbReference) &&

    dbReference$isoform == isoform_id) {
      refseq_id <- dbReference$id
      break
    }
```

```
  }

  return(refseq_id)
}

refseqid_nocanonical <- function (input){
  if (is.null(protein_info2) || !("dbReferences" %in% names(protein_info2))) {
    return(NULL)
  }

  refseq_id2 <- NULL

  for (db in protein_info2$dbReferences) {
    if (db$type == "RefSeq") {
      refseq_id2 <- db$id
      break
    }
  }

  return(refseq_id2)
}
```

1.3.1: Búsqueda de la secuencia de canónicos

```{r}

gene_names <- c("RBM20","NEXN","NF1","TMEM43","YARS2","PDSS2","COQ2","ANO5","SCN4B","PUS1",

"KCNQ1","FKTN","TRPM4","ALMS1","FKRP","TBX20","HCN4","TCAP","SCO2","MYOT",

"BAG3","GNE","KCNE2","COX15","ABCC6","PKP2","SHOC2","PRKAG2","TBX5",

"ABCC9","SGCD","MAP2K2","CAV3","MYBPC3","HADHA","SDHA","KCNJ2","CPT2","NKX2-5",

"GATA4","SLC25A3","SCN1B","SCN5A","LAMP2","GLA","EMD","DMD","AIFM1","FHL1",

"TNNT2","TNNI3","TNNC1","TPM1","KRAS","HRAS","TTN","SOS1","RYR2","RYR1",

"PTPN11","MAP2K1","TTR","KCNA5","KCNE1","JUP","PLN","CBL","NRAS","RAF1",


```

```
"BRAF","MYL3","MYL2","MYH7","MYH6","KCNH2","LMNA","DSG2","DES","DSP",

"DSC2","CRYAB","GJA5","CASQ2","CACNA1C","APOA1","PSEN1","SLC25A4","ACTA1",

"FLNC","ACTC1","KCNJ5","MAN1B1","JPH2","MYOZ2","SRD5A3","COA5","GATAD1",

"MYPN","MTO1","TK2","CACNA1D","CALM1","FARS2","MFF","RIT1","PRDM16","MRPL44",

"FBXL4","SGCB","CALM2","DOLK","LZTR1","SCN3B","TSFM","CACNB2","TNNI3K",

"GYG1","ACTN2","VCL","AKAP9","SDHD","COQ4","CSRP3","HADHB","NDUFS1","SYNE1",

"TGFB3","ACADVL","ANK2","SOS2","PMPCA","OPA1","ECHS1","DPM3","PKD2","IDH2",

"KCND3","CALM3","POLG","GAA","GATA5","EYA4","SCN10A","LDB3","NEBL","LARS2",

"CDH2","NDUFA13","LAMA2","SYNE2","MIB1","FHOD3","CASZ1","DTNA","EARS2",

"DHDDS")  # Lista de genes a buscar


genes <- list()  # Lista para almacenar los resultados únicos


results <- data.frame(Gene = character(), IsoformID = character(), RefSeqID = character(), s

# Iterar sobre los nombres de genes y obtener la información
for (gene_name in gene_names) {
  protein_info <- get_uniprot_info(gene_name)

  if (!is.null(protein_info)) {
    genes[[gene_name]] <- protein_info
  }
}

# Mostrar la isoforma canónica y el identificador RefSeq encontrados
for (gene_name in names(genes)) {
  protein_info <- genes[[gene_name]]
  canonical_isoform <- extract_canonical_isoform(protein_info)

  if (!is.null(canonical_isoform)) {
```

```
      refseq_id <- extract_refseq_id(protein_info, canonical_isoform)

      if (!is.null(refseq_id) && length(refseq_id) > 0) {
        results <- rbind(results, data.frame(Gene = gene_name,

        IsoformID = canonical_isoform, RefSeqID = refseq_id, stringsAsFactors = FALSE))
      } else {
        cat("No se encontró un ID RefSeq para la isoforma canónica del gen",

        gene_name, "\n\n")
      }
    } else {
      cat("No se encontró una isoforma canónica para el gen", gene_name, "\n\n")
    }
}

print(results)
```

Hay genes para los que no se encuentra isoforma canónica debido a que no tienen,

solo tienen una isoforma y esa es la canónica y también hay que tratar eso

```{r}
tmp <-results
```

1.3.2: Tratamiento de los genes que no tienen isoforma canónica

```{r}

genes_can <- tmp$Gene
genes_can <- as.data.frame(genes_can)

colnames(genes_can)[colnames(genes_can) == "genes_can"] <- "gen"

todos <- as.data.frame(gene_names)

colnames(todos)[colnames(todos) == "gene_names"] <- "gen"

genes_sin_can <-todos %>% anti_join(genes_can)
```

```
genes_sin_can<- c("RBM20","TMEM43","YARS2","COQ2","ANO5","PUS1","ALMS1","FKRP","TBX20",
                  "HCN4","TCAP","SCO2","MYOT","BAG3","KCNE2","MAP2K2","CAV3","SDHA",
                  "KCNJ2","CPT2","GLA","EMD","DMD","TNNI3","TNNC1","HRAS","KCNE1","JUP",
                  "PLN","CBL","NRAS","BRAF","MYL3","MYL2","MYH7","MYH6" ,"DSG2","DES",
                  "CRYAB","GJA5","CACNA1C","APOA1","SLC25A4","ACTA1","ACTC1","KCNJ5",
                  "MAN1B1","MYOZ2","SRD5A3","COA5","GATAD1","CALM1","FARS2","MRPL44",
                  "FBXL4","CALM2","DOLK","LZTR1","SCN3B","AKAP9","ECHS1","CALM3","POLG",
                  "GATA5","SCN10A","LARS2","LAMA2","MIB1")

genes2<-list()
# Iterar sobre los nombres de genes y obtener la información
for (gene_name in genes_sin_can) {
  protein_info2 <- get_uniprot_info(gene_name)

  if (!is.null(protein_info)) {
    genes2[[gene_name]] <- protein_info2
  }
}


results2 <- data.frame(Gene = character(), RefSeqID = character(), stringsAsFactors = FALSE)

# Mostrar la isoforma canónica y el identificador RefSeq encontrados
for (gene_name in names(genes2)) {
  protein_info2 <- genes2[[gene_name]]

  refseq_id2 <- refseqid_nocanonical(protein_info2)
  if (!is.null(refseq_id2) && length(refseq_id2) > 0) {
    results2 <- rbind(results2, data.frame(Gene = gene_name, RefSeqID = refseq_id2,

    stringsAsFactors = FALSE))
  } else {
    cat("No se encontró un ID RefSeq para la isoforma canónica del gen", gene_name, "\n\n")
  }

}

print(results2)
```

```
results2$IsoformID <-"NA"
results3 <-rbind(results,results2)

‘‘‘

## 1.4: Hacer el match entre el resultado de la tarea 1.2 y 1.3

‘‘‘{r}

####### hacer el match con los proteinchange

resultado31_sep <- resultado31 %>% separate (ProteinExpression, into = c("NP","ProtCh"),

sep = ":")

colnames(resultado31_sep)[which(names(resultado31_sep) == '#Symbol')] <- 'Gene'


tmp <-resultado31_sep %>% filter(resultado31_sep$NP %in% results3$RefSeqID)

####### hacer el match de clinvarlimpio con tmp por el AlleleID

tmp2 <-tmp %>% filter(tmp$AlleleID %in% clinvarlimpio$AlleleID)


#### hacer el match en clinvar limpio del proteinchange añadiendo una columna
clinvarlimpio <- clinvarlimpio[!duplicated(clinvarlimpio$AlleleID), ]
tmp2 <- tmp2[!duplicated(tmp2$AlleleID), ]


tmp4 <- merge(clinvarlimpio, tmp2[, c("AlleleID", "ProtCh")], by = "AlleleID", a

ll.x = TRUE)

write.xlsx(tmp4,"protchange.xlsx")
‘‘‘


# 2: Sacar la información de Start - End
```

```{r}

filepath =

"D:/TFM/BIBLIOGRAFIA/FASE4/archivos_auxiliares/variant_summarydescomprimido.txt"

filepath2 =

"D:/TFM/BIBLIOGRAFIA/FASE4/HOTSPOTS/hgvs4variation.txt/chunks/protchange.xlsx"

#leer los archivos

clinvarlimpio = read_excel(filepath2)
vs = read.csv(filepath,sep='\t')

# filtramos los alleleid  Filter rows in df where AlleleID matches in clinvar
df_filt <- vs[vs$`X.AlleleID` %in% clinvarlimpio$AlleleID, ]

df_filt2 <- df_filt %>% filter(df_filt$Assembly=="GRCh37")


# Rename columns and merge the DataFrames
df_filt3 <- df_filt2 %>% rename(AlleleID = `X.AlleleID`)


clinvarlimpio <- clinvarlimpio %>% left_join(df_filt3[, c("AlleleID", "Start", "Stop")],

by = c("AlleleID"))

# Write the final DataFrame to an Excel file
write.xlsx(clinvarlimpio, "startendclinvar.xlsx")



```


3: Sacar la información de Mutation_subtype

```{r}
```

```
combined.df.final = clinvarlimpio

combined.df.final <- separate(combined.df.final, MC, into = c("MC_ID",

"MC_mutation_subtype"), sep = "\\|")

combined.df.final$MC_mutation_subtype <- sub(",.*", "",

combined.df.final$MC_mutation_subtype)

```

3.1: Creación de un filtro de mutation_subtype

```{r}

filtro <-c("Deletion", "Insertion", "nonsense","inframe_deletion","inframe_insertion",
           "missense_variant")

library(dplyr)
library(stringr)

filtered_df <- combined.df.final %>%
  filter(str_detect(CLNVC, paste(filtro, collapse = "|")) |
           str_detect(MC_mutation_subtype, paste(filtro, collapse = "|")))

filtered_df <- filtered_df %>%
  mutate(MC_mutation_subtype = case_when(
    MC_mutation_subtype == "missense_variant" ~ "Missense",
    MC_mutation_subtype == "nonsense" ~ "Nonsense",
    MC_mutation_subtype == "inframe_deletion" ~ "Deletion",
    MC_mutation_subtype == "inframe_insertion" ~ "Insertion",
    TRUE ~ MC_mutation_subtype
  ))


# Define los valores de MC_mutation_subtype que deseas reemplazar
replace_values <- c("frameshift_variant", "splice_acceptor_variant",
```

```
"genic_upstream_transcript_variant",
                    "inframe_deletion", "inframe_indel", "inframe_insertion",

                    "initiator_codon_variant",
                    "intron_variant", "non-coding_transcript_variant", "nonsense",
                    "splice_acceptor_variant", "splice_donor_variant", "")

# Usar mutate y case_when para reemplazar los valores en MC_mutation_subtype por los

valores correspondientes en CLNVC
filtered_df2 <- filtered_df %>%
  mutate(MC_mutation_subtype = case_when(
    MC_mutation_subtype %in% replace_values ~ CLNVC,
    TRUE ~ MC_mutation_subtype))

```

4: Adecuar a la forma especificada

```{r}

input <- filtered_df2[,c("ID","GENEINFO","CHROM","Start","REF","ALT","CLNVC",

"MC_mutation_subtype","Stop","ProtCh")]

names(input)[names(input) == "CHROM"] <- "Chromosome"
names(input)[names(input) == "ID"] <- "Sample_ID"
names(input)[names(input) == "GENEINFO"] <- "Gene_ID"
names(input)[names(input) == "Start"] <- "Start"
names(input)[names(input) == "Stop"] <- "End"
names(input)[names(input) == "REF"] <- "Ref_allele"
names(input)[names(input) == "ALT"] <- "Alt_allele"
names(input)[names(input) == "MC_mutation_subtype"] <- "Mutation_subtype"
names(input)[names(input) == "ProtCh"] <- "Protein_variant"

write.xlsx (input, "input.xlsx")


input2 <-input[!is.na(input$Protein_variant),]
```

```
```

4.1: Arreglo del protein_variant a la forma que toca

```{r}

dic_aa <- data.frame(Acronimo = c("Val","Leu","Thr","Lys","Trp", "His", "Phe",
                                  "Ile", "Arg", "Met", "Ala", "Pro", "Gly",
                                  "Ser", "Cys", "Asn", "Gln", "Tyr", "Asp",
                                  "Glu"),
                     Letra=c("V", "L", "T", "K","W", "H", "F", "I","R","M","A",
                             "P", "G","S", "C","N", "Q","Y", "D","E"))
list_dic_aa <- as.list(dic_aa)

# Quitar todo lo que hay antes del punto en Protein_variant
input2 <- input2 %>%
  mutate(Protein_variant = str_replace(Protein_variant, ".*\\.", ""))

# Reemplazar los acrónimos por las letras correspondientes
for (i in 1:nrow(dic_aa)) {
  input2$Protein_variant <- str_replace_all(input2$Protein_variant,

  dic_aa$Acronimo[i], dic_aa$Letra[i])
}


# EXPRESIÓN REGULAR PARA Quitar todo lo que hay detrás del número
input2 <- input2 %>%
  mutate(Protein_variant = str_replace(Protein_variant, "(\\d+).*", "\\1"))

library(openxlsx)
write.xlsx(input2, "inputhotdriver.xlsx")


```


4.2: Reordenación de las columnas para que se adecuen a las necesidades

```
```{r}
prueba =subset(input2,select = -c(CLNVC))
prueba = subset (prueba, select=c(1,2,3,4,8,5,6,9,7))
genes = subset (prueba, select=c(2))
genes= unique(genes)

prueba$Chromosome[prueba$Chromosome == "X"] <- 23
prueba$Chromosome[prueba$Chromosome == "Y"] <- 24
prueba$Chromosome <- as.numeric(prueba$Chromosome)

write_tsv(genes,"gene_list.tsv")
write_tsv(prueba,"mutation_data3.tsv")
```
```

1.10. Anexo 10: Código de preprocesado de los datos de entrada del método 4.

```
library(openxlsx)
library(readxl)

path <-"D:/TFM/BIBLIOGRAFIA/FASE4/HOTSPOTS/mutation_data.tsv"
data <- read.delim(path)

missense <- data[data$Mutation_subtype == "Missense", ]

nonsense <- data[data$Mutation_subtype == "Nonsense", ]

insertion <- data[data$Mutation_subtype == "Insertion", ]

deletion <- data[data$Mutation_subtype == "Deletion", ]

na <- data[data$Mutation_subtype == "NA", ]


path2 <-"D:/TFM/BIBLIOGRAFIA/FASE4/HOTSPOTS/gene_length.tsv"
library(readr)
gene.lengths <- read_tsv(path2)
```

```
# Divide la columna en dos basándote en el tabulador
gene.lengths <- strsplit(gene.lengths$`GEN LENGTH`, "\t", fixed = TRUE)
# Convertir la lista en un data frame
gene.lengths <- as.data.frame(do.call(rbind, gene.lengths))

# Asigna nombres de columna
colnames(gene.lengths) <- c("GENE", "LENGTH")


mutation.missense <- missense
mutation.missense <- mutation.missense[c("Gene_ID", "Protein_variant")]

nonsense <- nonsense[complete.cases(nonsense), ]

mutation.nonsense <- nonsense
mutation.nonsense <- mutation.nonsense[c("Gene_ID", "Protein_variant")]


# QUITAR LA LETRA DEL AA

mutation.missense$Protein_variant <- substr(mutation.missense$Protein_variant, 2, nchar(muta
mutation.nonsense$Protein_variant <- substr(mutation.nonsense$Protein_variant, 2, nchar(muta


colnames(mutation.missense) <- c("symbol", "position")
colnames(mutation.nonsense) <- c("symbol", "position")
colnames(gene.lengths) <- c("symbol", "length")
gene.lengths <- gene.lengths[1:(nrow(gene.lengths)-1), ]

gene.lengths$length <- as.numeric(gene.lengths$length)
mutation.missense$position <- as.numeric(mutation.missense$position)
mutation.nonsense$position <- as.numeric(mutation.nonsense$position)

##############################################################################

results_missense = read.table("mutClustSW_results_MISSENSE.txt", header=TRUE, sep="\t")

hotspots_missense_0.10 <- results_missense[results_missense$Pvalue < 0.1, ]
##############################################################################

genes<- as.data.frame(unique(mutation.nonsense$symbol))
```

```
colnames(genes) <- c("symbol_nonsense")


# Extraer los símbolos de genes y gene.lengths
symbols_genes <- unique(genes$symbol_nonsense)
symbols_gene_lengths <- unique(gene.lengths$GENE)

# Comparar coincidencias
common_symbols <- intersect(symbols_genes, symbols_gene_lengths)
################################################################################
results_nonsense = read.table("mutClustSW_results_NONSENSE.txt", header=TRUE, sep="\t")

hotspots_nonsense_0.10 <- results_nonsense[results_nonsense$Pvalue < 0.1, ]
```