UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Diseño y desarrollo de una fuente de datos sobre hotspots asociados al criterio PM1 de las guías ACMG-AMP 2015 aplicado a cardiopatías familiares

Trabajo Fin de Grado

Grado en Ingeniería Biomédica

AUTOR/A: García Zarzoso, Alba

Tutor/a: Costa Sánchez, Mireia

Cotutor/a: Pastor López, Oscar

CURSO ACADÉMICO: 2022/2023

AGRADECIMIENTOS

En primer lugar, me gustaría dar las gracias a mi tutora Mireia Costa Sánchez y a mi cotutor Óscar Pastor López por apostar por mí para llevar a cabo este Trabajo de Final de Grado. Estoy muy agradecida por todo el apoyo y la ayuda que he recibido a lo largo de mi estancia en VRAIN durante el periodo de beca de colaboración, así como durante la realización del Trabajo de Final de Grado.

También me gustaría dar las gracias a todos mis amigos de la carrera, hemos pasado muchas horas en la biblioteca estudiando para conseguir convertirnos en ingenieros biomédicos, sin vosotros esto no habría sido posible. Gracias por ser mi apoyo en mis días malos y mis mejores admiradores en mis días buenos. Siempre llevaré conmigo todo el aprendizaje y los buenos momentos que hemos pasado estos cuatro años juntos. Asimismo, me gustaría hacer una mención especial a Marina, mi compañera de trabajo a lo largo de las prácticas, muchas gracias por confiar en mí y ofrecer siempre tu ayuda cuando lo necesitaba.

Finalmente, me gustaría dar las gracias a mi familia, mi mayor apoyo. Ellos han sido los que siempre han confiado en mí, los que me han escuchado y aguantado a lo largo de estos cuatro años. Sin su apoyo incondicional no lo habría logrado.

RESUMEN

Los hotspots son regiones del ADN muy susceptibles de ser mutadas por una inestabilidad inherente, una tendencia al entrecruzamiento desigual o una predisposición química a sustituciones de nucleótidos simples. Las guías ACMG-AMP del 2015 son las más utilizadas a nivel clínico para determinar la relevancia clínica de una variación, de acuerdo con estas, el determinar si una variación se encuentra en un hotspot es un criterio – criterio PM1 de las guías – que indica una evidencia moderada de patogenicidad. La información relacionada con la localización de los hotspots se encuentra principalmente en la literatura, cuestión que dificulta la localización de las variaciones que se encuentran en una de esas regiones, y la posible automatización de dicho criterio. Por este motivo, el objetivo de este Trabajo Final de Grado es diseñar y desarrollar una fuente de datos de hotspots utilizando como caso de uso los fenotipos asociados a las cardiopatías familiares. La fase de diseño de esta fuente de datos comenzó con la caracterización de un hotspot y la definición de la estructura de esta fuente de datos, seleccionando la información de interés de la literatura. Tras esto, se colaboró en el desarrollo de una página web que recogiese toda la información y que resultase amigable para los expertos. Por último, se realizó una evaluación de los resultados basada en la comparación de estos con los extraídos en un caso real y su posterior análisis.

Palabras Clave: hotspots; fuente de datos; guías; variaciones genómicas; automatización.

RESUM

Els hotspots son regions del ADN molt susceptibles de ser mutades per una inestabilitat inherent, una tendència al entrecreuament desigual o una predisposició química a substitucions de nucleòtids simples. Les guies ACMG-AMP del 2015 son les més emprades a nivell clínic per a determinar la rellevància clínica d'una variació, d'acord amb estes, el determinar si una variació es troba en un hotspot es un criteri – criteri PM1 de les guies – que indica una evidencia moderada de patogenicitat. La informació relacionada amb la localització dels hotspots es troba principalment en la literatura, qüestió que dificulta la localització de les variacions que es troben en una de eixes regions, i la possible automatització del criteri. Per aquest motiu, el objectiu de este Treball Final de Grau es dissenyar i desenvolupar una font de dades de hotspots utilitzant com cas d'us els fenotips associats a cardiopaties familiars. La fase de disseny de esta font de dades començà amb la caracterització de un hotspot i la definició de la estructura de aquesta font de dades, seleccionant la informació d'interès de la literatura. A més, es col·laborà en el desenvolupament d'una pàgina web que reflexa tota la informació replegada i que resultés amigable per als experts. Per últim, es realitzà una avaluació dels resultats basada en la comparació de estos amb els extrets de casos reals i el seu posterior anàlisi.

Paraules clau: hotspots; font de dades; guies; variacions genòmiques; automatització.

ABSTRACT

Hotspots are regions of DNA that are highly susceptible to mutation due to inherent instability, a tendency for unequal crosslinking, or a chemical predisposition to single nucleotide substitutions. The 2015 ACMG-AMP guidelines are the most widely used at the clinical level to determine the clinical relevance of a variation, according to these, determining whether a variation is in a hotspot is a criterion - criterion PM1 of the guidelines - that indicates moderate evidence of pathogenicity. The information related to the location of hotspots is mainly found in the literature, which makes it difficult to locate the variations found in one of these regions, and the possible automation of this criterion. For this reason, the aim of this Final Degree Project is to design and develop a hotspot data source using the phenotypes associated with familial heart disease as a use case. The design phase of this data source started with the characterization of a hotspot and the definition of the structure of this data source, selecting the information of interest from the literature. After this, we collaborated in the development of a web page that would gather all the information and would be friendly for the experts. Finally, a evaluation of the results was carried out based on the comparison of the results with those extracted in a real case and their further analysis.

Keywords: hotspot; data source; guidelines; genomic variation; automation.

ÍNDICE

DOCUMENTOS CONTENIDOS EN EL TFG

- Memoria
- Presupuestos
- Anexos

ÍNDICE DE LA MEMORIA

ÍNDICE DEL PRESUPUESTO

ÍNDICE DE LOS ANEXOS

ÍNDICE DE LAS FIGURAS

ÍNDICE DE LAS TABLAS

MEMORIA:

Diseño y desarrollo de una fuente de datos sobre hotspots asociados al criterio PM1 de las guías ACMG – AMP del 2015 aplicado a cardiopatías familiares

Documento I

Alumna: Alba García Zarzoso

Tutora: Mireia Costa Sánchez

Cotutor: Óscar Pastor López

Grado en Ingeniería Biomédica

Curso 2022 - 2023

CAPÍTULO 1: INTRODUCCIÓN

En este Trabajo Final de Grado se va a diseñar, desarrollar y evaluar una fuente de datos de hotspots aplicada a cardiopatías familiares acorde con el criterio PM1 de las guías ACMG – AMP del 2015 para que esta pueda ser aplicada en el ámbito del diagnóstico clínico-genómico. Es por ello, por lo que este trabajo se encuentra enmarcado en el contexto de la bioinformática, uno de los más relevantes dentro del campo de la ingeniería biomédica.

En este capítulo se va a definir el marco contextual del trabajo, la problemática que se plantea, la motivación para realización del trabajo y la estructura de este.

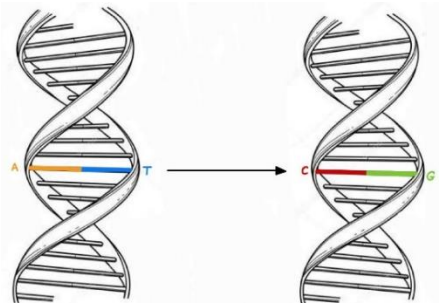
1.1. MEDICINA DE PRECISIÓN

Según el *National Human Genome Research Institute* "la medicina de precisión puede definirse como un enfoque innovador que utiliza información sobre características genómicas, ambientales y del estilo de vida de una persona para guiar la toma de decisiones relacionadas a su atención médica*"*(*Medicina de Precisión*, n.d.).

De la definición anterior se puede extraer que la medicina de precisión engloba muchas especialidades, siendo una de las más importantes la genética. Esta especialidad se basa en el conocimiento del ADN de cada paciente para poder ofrecer los mejores diagnósticos y tratamientos posibles. Por este motivo, la evaluación de la variabilidad genética entre individuos es uno de los pilares fundamentales de la medicina de precisión.

En consecuencia, resulta fundamental definir algunos conceptos para contextualizar el trabajo. El primero de ellos es el de variación genómica, que se define como un cambio en la secuencia de ADN de un individuo con respecto a una secuencia de referencia que representa a un individuo ideal completamente sano. Tal y como se observa en la Figura 1, que representa un ejemplo de variación genómica, pasamos de tener en la secuencia de referencia los nucleótidos[1] A-T, a tener los nucleótidos C-G. El segundo concepto es el de secuencia de referencia, el cual se define como una representación aceptada de la secuencia del genoma humano que usan los investigadores como estándar para la comparación de secuencias de ADN generadas en sus estudios según el *National Human Genome Research Institute* (*Secuencia de Referencia Del Genoma Humano*, n.d.).

[1] **Nucleótido**: Estructura fundamental básica de los ácidos nucleicos (ARN y ADN). Esta estructura consta de una molécula de azúcar unida a un grupo fosfato y a una base nitrogenada (*Nucleótido*, n.d.).



Secuencia de referencia Secuencia alterada

Figura 1: Representación gráfica de un tipo de variación genómica conocida como *Polimorfismo de nucleótido simple*[2]. [Elaboración propia]

Todas estas variaciones nos diferencian entre individuos y, algunas de ellas, son las responsables del desarrollo de distintas enfermedades. Por este motivo resulta crucial la identificación de estas variaciones genómicas ya que suponen puntos clave a la hora del diagnóstico de enfermedades. Dado que distinguir estas variaciones no es un trabajo sencillo y que la enorme cantidad de enfermedades genéticas que existen complican más esa distinción, este trabajo se centra en el estudio de un tipo de ellas, las cardiopatías familiares, para así contribuir a la solución de la problemática del trabajo.

1.2. CARDIOPATÍAS FAMILIARES

Dado que el objetivo principal de este trabajo es el diseño y desarrollo de una fuente de datos asociados a hotspots de cardiopatías familiares, resulta evidente la necesidad de contextualizarlas. Por este motivo esta subsección pretende definir y clasificar los diferentes tipos de cardiopatías familiares para así, tener un marco teórico al que recurrir.

Según la Fundación del Corazón, las cardiopatías familiares se definen como "enfermedades que afectan al corazón y a los grandes vasos arteriales debido a alteraciones genéticas"(Cardiopatías Familiares y Genética - Fundación Española Del Corazón, n.d.). Estas enfermedades se clasifican en tres grandes grupos: miocardiopatías, canalopatías y aortopatías.

Las miocardiopatías son enfermedades que afectan al músculo cardíaco sin motivo aparente y que pueden provocar insuficiencia cardíaca entre otros síntomas. Las miocardiopatías familiares a su vez se dividen en cinco tipos: (1) miocardiopatía hipertrófica o MCH, (2) miocardiopatía dilatada o MCD, (3) miocardiopatía restrictiva, (4) miocardiopatía arritmogénica o MCA y (5) miocardiopatía no compactada.

El primer tipo es el de la miocardiopatía hipertrófica, la cual se caracteriza por un aumento anormal en el espesor del músculo cardíaco a consecuencia de un entrenamiento físico prolongado o por hipertensión arterial. El patrón hereditario de esta enfermedad es generacional, sin saltos entre generaciones y cada hijo o hija de una persona afectada tiene el 50% de posibilidades de heredarla (*Miocardiopatía Hipertrófica - Fundación Española Del Corazón*, n.d.).

El segundo tipo es el de la miocardiopatía dilatada, la cual se caracteriza por un aumento del tamaño de los ventrículos, dificultando así el bombeo fisiológico de sangre. Esta enfermedad se manifiesta mediante la insuficiencia cardíaca, siendo este término un conjunto de síntomas relacionados entre sí.

[2] **Polimorfismo de Nucleótidos simple o SNP**: variante genómica en la posición de una base única en el ADN.

Estos síntomas son la dificultad de respirar, la hinchazón en tobillos y abdomen y el cansancio general.(*Miocardiopatía Dilatada - Fundación Española Del Corazón*, n.d.).

El tercer tipo es el de la miocardiopatía restrictiva, las cuales presentan alteraciones en la relajación del corazón - función diastólica - es decir, el corazón no es capaz de relajarse correctamente y, por tanto, no llena sus cavidades de forma correcta. Estas enfermedades suelen ser de origen genético o provocadas por enfermedades que infiltran el miocardio como puede ser la sarcoidosis o la amiloidosis(*Miocardiopatía Restrictiva - Fundación Española Del Corazón*, n.d.) .

El cuarto tipo de miocardiopatía es la arritmogénica o MCA. Esta miocardiopatía es provocada debido a un mal desarrollo de proteínas que mantienen el músculo del corazón unido. Esta enfermedad se muestra mediante palpitaciones, mareos, pérdidas de conocimiento y dificultad al respirar(*Miocardiopatía Arritmogénica - Fundación Española Del Corazón*, n.d.).

El quinto, y último tipo de miocardiopatía es la no compactada. Esta clase se caracterizan por presentar el corazón dividido por trabéculas que pueden provocar una pérdida de fuerza del corazón (*Miocardiopatía No Compactada - Fundación Española Del Corazón*, n.d.). Una vez introducidos los diferentes tipos de miocardiopatías se pasa a la definición de canalopatías.

Las canalopatías son enfermedades que afectan a la actividad eléctrica del corazón y, por tanto, la estructura del corazón no se ve afectada. Una de las características de este grupo es que no se puede diagnosticar mediante técnicas como son el ecocardiograma o el estudio autópsico del corazón, debido a que la estructura del corazón no se ve modificada (*Síndrome de Brugada – Cardiopatias Familiares*, n.d.). Algunas de las más comunes son el síndrome de QT largo, el síndrome de QT corto, el síndrome de Brugada, la taquicardia ventricular catecolaminérgica, el síndrome de repolarización precoz, la fibrilación ventricular idiopática, la fibrilación auricular familiar, la amiloidosis y la muerte súbita familiar entre otras. De todas estas enfermedades las más importantes son el síndrome de QT largo y corto, el síndrome de Brugada y la taquicardia ventricular catecolaminérgica y, por tanto, a continuación, se va a explicar un poco de cada una de ellas.

El síndrome de QT largo es una canalopatía que afecta al proceso de repolarización dentro de la contracción cardíaca y, por tanto, se produce debido a una alteración en los canales de sodio y potasio(Escobar Cervantes et al., 2005). Los síntomas más habituales de este síndrome son los síncopes o la muerte súbita por paro cardíaco (Medeiros-Domingo et al., 2007). El diagnóstico de esta enfermedad es a través del electrocardiograma o ECG buscando un alargamiento del intervalo QT (*Síndrome de QT Largo | The Texas Heart Institute*, n.d.). Este tipo de canalopatía facilita la aparición de arritmias mortales tales como la taquicardia ventricular polimórfica tipo torsades des pointes o fibrilación ventricular (“Síndrome de QT Largo Congénito: Revisión de La Literatura,” n.d.).

El síndrome de QT corto es un tipo de canalopatía provocada por alteraciones de flujo de corriente en los canales de potasio, incrementándolas, o en los canales de calcio, disminuyéndolas y se manifiesta presentando un acortamiento heterogéneo de la repolarización ventricular. Esta enfermedad se diagnostica gracias al ECG, fijando la atención en detectar acortamientos del intervalo QT, tal y como su propio nombre indica (Ginesi et al., n.d.). En cuanto a los síntomas se puede destacar que esta enfermedad presenta alta variabilidad de síntomas, teniendo entre ellos la fibrilación auricular paroxística, el síncope, la muerte súbita o las arritmias ventriculares (Cardentey et al., 2009).

El síndrome de Brugada es un tipo de canalopatía originada por alteraciones en el flujo de corrientes de iones involucrados en la contracción muscular del corazón (Benito et al., 2009). Estas alteraciones se manifiestan con la disminución de la corriente de entrada a la célula de sodio o calcio por un lado o, por el otro lado, con un aumento de las corrientes de salida de potasio de forma temprana . Esta

enfermedad se manifiesta con la posible aparición de arritmias peligrosas y algunos de los pacientes pueden padecer síncopes o muerte súbita causada por las arritmias. El diagnóstico de esta enfermedad es a través del electrocardiograma o ECG buscando una alteración de onda producida en el segmento ST, ya que se presentan alteraciones eléctricas de un área especial del ventrículo derecho (*Síndrome de Brugada – Cardiopatias Familiares*, n.d.).

La taquicardia ventricular catecolaminérgica es una canalopatía que se caracteriza por presentar alteraciones en la regulación del ion calcio intracelular, favoreciendo las arritmias ventriculares con riesgo de muerte súbita. El diagnóstico de esta enfermedad reside en la monitorización electrocardiográfica durante 24h, incluyéndose una prueba de esfuerzo o un test de epinefrina o isoproterenol (Argelia Medeiros-Domingo & Medeiros-Domingo, 2009). Esto es debido a que el electrocardiograma de una persona que padece esta enfermedad es fisiológico cuando no se somete este a un esfuerzo.

Las aortopatías son enfermedades que afectan principalmente a la pared de la aorta y las más comunes son el síndrome de Marfan, el síndrome de Loeys – Dietz y la afectación vascular del síndrome de Ehlers Danlos. A continuación, se van a explicar los dos primeros síndromes enumerados debido a su importancia ya que, ambos dos, presentan síntomas específicos en el órgano cardíaco mientras que, el síndrome de Ehlers Danlos no tiene el corazón como órgano diana, aunque también le afecte.

El síndrome de Marfan es una enfermedad provocada por una alteración genética en el cromosoma 15 que se manifiesta afectando al tejido conectivo en general (Barriales-Villa et al., 2011). Esta enfermedad afecta tanto al corazón y los vasos sanguíneos como al esqueleto y los ojos. Concretamente, uno de los principales síntomas relacionados con el corazón y los vasos sanguíneos es la dilatación de la aorta debido a la debilidad de la pared, pudiendo provocar un desgarro o una rotura de esta (Oliva et al., 2006). El diagnóstico de esta enfermedad pasa por un ecocardiograma y una revisión ocular, a parte de la revisión de la historia clínica familiar (*Síndrome de Marfan - Fundación Española Del Corazón*, n.d.). Asimismo, existen unos criterios diagnósticos llamados los criterios nosológicos de Ghent de 1996, a través de los cuales se determina de forma diferencial si una persona tiene síndrome de Marfan o no (Primaria et al., n.d.).

El síndrome de Loeys – Dietz es un tipo de aortopatía genética con un patrón de herencia autosómica dominante y que afecta principalmente al tejido conectivo. Esta enfermedad se presenta mediante la presencia de aneurismas o disecciones arteriales, cerebrales, torácicas, abdominal y con manifestaciones esqueléticas craneofaciales (Ayerza Casas et al., 2017). Dentro de este cuadro sintomatológico, y más concretamente en el ámbito de interés de este Trabajo de Final de Grado, los síntomas más comunes son el agrandamiento de la aorta y la debilidad de los vasos sanguíneos, así como la presencia de protuberancias o el alargamiento de las arterias (*Síndrome de Loeys-Dietz - Stanford Medicine Children's Health*, n.d.). Esta enfermedad aparece como resultado de la mutaciones de los cinco genes encargados de la vía de señalización celular del factor de crecimiento transformante beta (TGF -β) (*Síndrome de Loeys-Dietz - Stanford Medicine Children's Health*, n.d.).

En la Figura 2 se puede observar un esquema representativo de cardiopatías familiares existentes divididas por tipos. Es importante recalcar que en este esquema no quedan representados todos los tipos y subtipos de cardiopatías, pero sí que muestra a rasgos generales los tipos más importantes y destacados. Asimismo, este apartado no explica de forma detallada todas las cardiopatías familiares dado que este estudio se centra únicamente en aquellas presentes en el marco del proyecto OGMIOS.

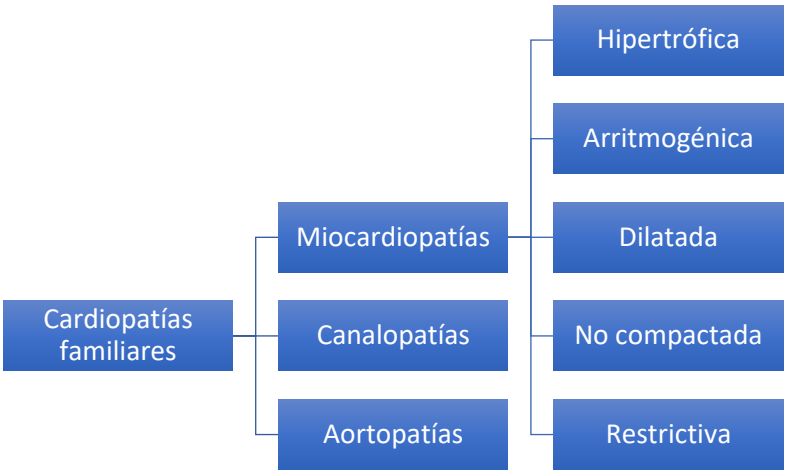


Figura 2: Cardiopatías familiares y sus tipos [Elaboración propia]

Una vez se conoce que son las cardiopatías familiares y cuáles son sus tipos, ya se pueden estudiar estas enfermedades. Debido a que estas enfermedades tienen una herencia genética bastante fuerte, la necesidad de una correcta clasificación de las variaciones genómicas de la misma resulta evidente. Sin embargo, esta necesidad no se soluciona de forma sencilla dado que tiene diferentes enfoques, es por ello por lo que en la sección 1.3 se presentan todas estas problemáticas para poder centrar el foco del trabajo en la resolución de estas.

1.3. PROBLEMÁTICA

Este Trabajo de Final de Grado tiene como cometido el diseño y desarrollo de una fuente de datos de hotspots asociados al criterio PM1 de las guías ACMG – AMP del 2015 aplicado a cardiopatías familiares y, por tanto, es importante no perder el foco a la hora de estudiar los problemas a los que se enfrenta este proyecto.

La problemática que se aborda en este Trabajo de Final de Grado puede dividirse en dos grandes problemas principales. En primer lugar, la comunidad científica difiere actualmente sobre los límites de definición de hotspot. El hecho de no tener una definición clara de hotspot conlleva una incertidumbre y la aparición de diferencias acerca de lo que es considerado un hotspot mutacional entre los diferentes expertos.

En segundo lugar, en este campo de estudio faltan bases de datos y repositorios que proporcionen información precisa sobre hotspots debido a la imprecisión de su definición. A pesar de que existen algunos repositorios únicamente dedicados a los hotspots, todos ellos son dedicados a la oncología, existiendo una falta de información importante para el resto de las disciplinas médicas.

El grupo PROS de la Universitat Politécnica de València (UPV) está trabajando dentro del ámbito clínico de las cardiopatías familiares, colaborando con los hospitales de la Fe y el hospital de Alicante en los proyectos OGMIOS (INNEST/2021/57) y CARDIOVAL (INBIO2021/AP2021-05). Estos proyectos pretenden mejorar y facilitar el diagnóstico de enfermedades a través de la interpretación de variaciones genómicas mediante el modelo conceptual del genoma humano. Los hotspots son un concepto íntimamente relacionado con la patogenicidad de ciertas enfermedades y, es por ello, que identificarlos resulta crucial. Dentro de ámbitos como la oncología sí existen bases de datos acerca de hotspots, pero en el ámbito cardiológico no es así y, por tanto, el estudio y la evaluación de ciertas variaciones que se encuentran en estas regiones se ve limitada. Por este motivo, el diseño y desarrollo

de la fuente de datos de hotspots para cardiología, objetivo de este Trabajo de Fin de Grado, supone una ventaja, no solo para la comunidad científica en general, sino que también para los investigadores del grupo PROS – UPV y para los hospitales que participan en estos proyectos específicos.

Las cardiopatías hereditarias son un factor de riesgo muy relacionado con la muerte súbita. Según la Fundación Española del Corazón, se estima que este tipo de cardiopatías afectan a 1 de cada 400 personas y es por ello por lo que diagnosticarlas a tiempo puede suponer una disminución del riesgo de padecer muerte súbita u otras complicaciones relacionadas (Cardiopatías Familiares y Genética - Fundación Española Del Corazón, n.d.). Para realizar esta identificación de forma apropiada, abordar la caracterización de hotspots mutacionales relacionados con este tipo de enfermedades es una tarea crucial. El censo de España de 2013, que analiza la prevalencia de cada una de las cardiopatías hereditarias, estima que hay 200.000 habitantes españoles con alguna cardiopatía hereditaria y, más concretamente dentro de la Comunidad Valenciana, se encuentran alrededor de 20.000 de ellos puesto que esta comunidad supone el 10% de la población española (Universitat & València, 2021).

Afortunadamente, hoy en día se sabe el fundamento genético de algunas cardiopatías hereditarias como son por ejemplo la miocardiopatía hipertrófica (MCH) o el síndrome de Brugada (SB) (Ackerman et al., 2013). Sin embargo, la heterogeneidad clínica y genética es una de las características clave de estos trastornos. Por tanto, la tarea de identificar variaciones genómicas relevantes no es sencilla, factor que se agrava por la dificultad de encontrar información sobre hotspots en este dominio. Por todo lo expuesto anteriormente, la necesidad de la caracterización de los hotspots relacionados con cardiopatías familiares y el diseño y desarrollo de una fuente de datos de hotspots recae sobre su propio peso.

1.4. MOTIVACIÓN

Tal y como se ha expuesto en el apartado 1.2 de este capítulo, la generación y uso de la información genética se encuentra en un momento de auge. Sin embargo, la heterogeneidad y la imprecisión de las definiciones está dificultando el uso de esta información en ámbitos clínicos reales. Por este motivo, se precisa de herramientas que permitan acceder fácilmente a una información bien estructurada y fiable.

La caracterización de los hotspots resulta fundamental para la clasificación de las variaciones genómicas. Sin embargo, existe un número limitado de fuentes de información sobre hotspots, y todas se centran en el ámbito de la oncología. Esto hace que los expertos de otros dominios se vean obligados a navegar entre las múltiples fuentes de información y referencias bibliográficas para encontrar la información necesaria. Esta ausencia de fuentes de información supone una inversión de tiempo excesiva por parte de los expertos, así como una amplia frustración a la hora de investigar sobre un ámbito del cual no tienen acceso a la información que necesitan de forma rápida y sencilla.

Para la clasificación de las variaciones genómicas se recurre a las guías ACMG – AMP del 2015 (Richards et al., 2015), tal y como se explica en el apartado 5.4 del capítulo 5. El motivo fundamental de la utilización de estas guías es el hecho de que estas son las más empleadas en el ámbito clínico. Es importante recordar que uno de los principales cometidos de este trabajo es la homogeneización y unificación de la información de hotspots a través de la creación de una fuente de datos. Por este motivo, la utilización de las guías más empleadas en la clasificación de variaciones resulta evidente para poder conseguir nuestro cometido.

El grupo PROS del instituto VRAIN ha participado activamente en los proyectos OGMIOS y CARDIOVAL, los cuales buscan facilitar la clasificación de las variaciones en el ámbito de la cardiología. Por un lado, el proyecto OGMIOS tiene como principal cometido desarrollar una plataforma inteligente para facilitar y promover el acceso a los datos clínicos y genómicos a los investigadores y expertos dentro de los ámbitos de la oncología pediátrica y la cardiología. Por el otro lado, el proyecto CARDIOVAL tiene como objetivo principal diseñar y desarrollar un prototipo basado en la inteligencia artificial explicable para la gestión de la información genética relacionada con el riesgo de sufrir muerte súbita de origen cardíaco.

El punto de unión de ambos proyectos reside en la clasificación de las variaciones genómicas relacionadas con cardiopatías familiares hereditarias según los criterios de las guías ACMG – AMP del 2015. Más concretamente, la identificación de un hotspot supone la evaluación del criterio PM1 de estas guías (Richards et al., 2015). Este criterio dice que ciertos dominios proteicos son críticos para definir la función de la proteína asociada a ese dominio y, que las variantes sin sentido de esos dominios suelen ser patogénicas. Además, es importante recalcar que, en estas zonas no hay muchas variantes benignas (Richards et al., 2015).

Asimismo, los hotspots mutacionales suelen encontrarse en regiones peor caracterizadas donde las variantes genómicas patogénicas se han detectado con una mayor frecuencia. Cualquiera de los dos motivos – estar un dominio proteico crítico o presentar alta frecuencia de mutabilidad – son evidencias moderadas de patogenicidad y, por tanto, deben clasificarse estas variantes con el criterio PM1 (Richards et al., 2015).

1.5. ESTRUCTURA DEL TRABAJO

Este Trabajo de Final de Grado se organiza en 9 capítulos:

- **Capítulo 1. Introducción:** En este se introduce el marco contextual del Trabajo de Final de Grado, así como la problemática y motivación para llevarlo a cabo.
- **Capítulo 2. Objetivos:** En este se detallan los objetivos del trabajo, así como las preguntas de investigación asociadas a cada uno de ellos.
- **Capítulo 3. Metodología**: En este se expone la metodología del *Design Science* y se explica por qué se ha empleado esta para la realización del Trabajo de Final de Grado.
- **Capítulo 4. Estado del arte:** Se presentan los pocos recursos disponibles acerca de hotspots y se expone la heterogeneidad bibliográfica en cuanto a la definición de hotspot .
- **Capítulo 5. Investigación del problema:** Se presentan las definiciones finales consideradas de hotspot, así como el concepto del modelado conceptual.
- **Capítulo 6. Diseño de la solución:** Se presenta la fuente de datos de hotspots relacionados con cardiología realizada, así como la web desarrollada.
- **Capítulo 7. Resultados:** Se muestran los resultados de la validación de los resultados con datos reales.
- **Capítulo 8. Conclusiones y trabajo futuros:** Se presentan las conclusiones del Trabajo de Final de Grado respondiendo a las preguntas presentadas en el capítulo 2 y se presentan futuras líneas de investigación.
- **Capítulo 9. Referencias bibliográficas:** Se presenta la bibliografía empleada para este trabajo.

CAPÍTULO 2: OBJETIVOS

El objetivo de este Trabajo de Final de Grado es diseñar y desarrollar una fuente de datos sobre hotspots asociados al criterio PM1 de las guías ACMG – AMP del 2015 en el caso concreto de las cardiopatías familiares. Para poder lograr este cometido se plantean una serie de objetivos más específicos, cada uno de ellos con una serie de preguntas de investigación asociadas:

Objetivo 1. Investigación del problema

PI1. ¿Cuáles son los usuarios objetivo?

PI2. ¿Qué necesidades de estos usuarios se espera cubrir con la fuente de datos?

PI3. ¿Qué definición se adopta como hotspot?

PI4. ¿Cuál es la base conceptual del trabajo?

PI5. ¿Cómo interpretamos una variación genómica dentro de un hotspot?

PI6. ¿Cuál es la base metodológica que debería utilizarse para desarrollar la fuente de datos?

Objetivo 2. Diseño y desarrollo de la fuente de datos

PI7. ¿Qué estructura tiene la fuente de datos?

PI8. ¿Cuál ha sido el criterio de selección de los documentos?

PI9. ¿Cuál es la información de interés de los artículos?

PI10. ¿Cómo ha sido la extracción de datos?

PI11. ¿Cómo ha sido el proceso de estandarización de la información?

PI12. ¿Cómo se presenta esta fuente de datos hacia el usuario?

Objetivo 3: Evaluación de la fuente de datos.

PI13. ¿Cuáles han sido los resultados obtenidos?

PI14. ¿La fuente de datos cumple los objetivos expuestos?

CAPÍTULO 3: METODOLOGÍA

El desarrollo de este trabajo se basa en una metodología de ciencia de diseño o *Design Science* propuesta por Wieringa (Wieringa, 2014), y que consiste en diseñar e investigar artefactos en un contexto para dar respuesta a un problema específico.

El artefacto de este Trabajo de Final de Grado consiste en diseñar y desarrollar una fuente de datos sobre hotspots asociados al criterio PM1 de las guías ACMG – AMP del 2015 en el contexto de la genética de las cardiopatías familiares.

En la metodología del *Design Science* se clasifican los problemas de dos maneras; por un lado, tenemos los problemas prácticos, los cuales se encuentran asociados a un ciclo de diseño o ingeniería y, por el otro lado, tenemos los problemas de conocimiento, los cuales están relacionados con un ciclo empírico o experimental.

En el primer tipo de problema, se busca aplicar esta metodología 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 el segundo tipo de problema, la metodología presenta un marco contextual del problema con el objetivo de tener más conocimiento acerca del mundo, pero sin la necesidad de crear una solución.

Para adaptar esta metodología al caso de estudio primero hay que identificar qué tipo de problema se aborda. Por todo lo expuesto anteriormente se deduce que este Trabajo de Fin de Grado se trata de un problema práctico, sobre el que se debe aplicar un ciclo de ingeniería. El ciclo de la ingeniería es un proceso racional de resolución de problemas que consta de cinco etapas:

1. **Investigación del problema.** Esta primera etapa consiste en la caracterización del problema resolviendo a las preguntas que se plantean a continuación: ¿cuáles son las necesidades de los usuarios de diseño?, ¿cuál es el marco conceptual del problema?, ¿qué se pretende conseguir con el diseño?
2. **Diseño de solución.** Esta segunda etapa consiste en la especificación de todos los requerimientos de la solución propuesta para el problema práctico que se propone.
3. **Validación de la solución.** Esta tercera etapa consiste en el estudio de si la solución diseñada puede cubrir todas las dimensiones del problema o no .
4. **Implementación de la solución.** Esta cuarta etapa consiste en la implementación de la solución diseñada al problema de ingeniería.
5. **Evaluación de la implementación.** Esta última etapa consiste en analizar el éxito o no de la implementación de la solución.

Debido a que los últimos dos puntos del ciclo se corresponden con la transferencia a entornos industriales, este Trabajo de Final de Grado únicamente se centra en la ejecución de las dos primeras etapas del ciclo regulativo del diseño y una tercera en la que se realiza una evaluación de la fuente de datos obtenida. En la Figura 3 podemos observar un esquema de la implementación de este ciclo para este caso en concreto.

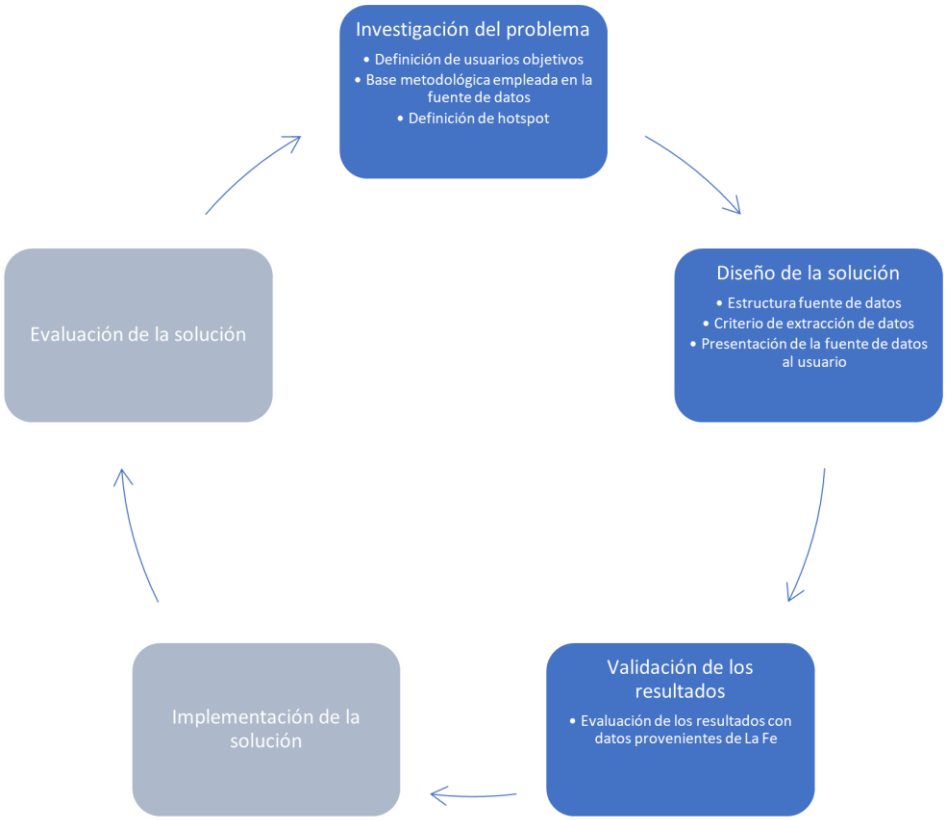


Figura 3: Ciclo regulativo de diseño [Elaboración propia]

El primer paso que realizar es la investigación del problema. En este paso se especifica la definición de hotspot empleada para nuestro caso en concreto y se define la base conceptual del problema, para así estructurar la información bajo las mismas premisas, la cual va a ser el modelado conceptual. Asimismo, en esta etapa se escoge la base metodológica que se va a emplear, en este caso, el método SILE.

El segundo paso para realizar es el diseño y desarrollo de la solución. En este paso se especifican los requerimientos de la propia fuente de datos y la aplicación de la base metodológica escogida para la resolución del problema, así como la estandarización de la información extraída.

El tercer paso del ciclo regulativo de diseño es el de la validación de resultados. Para llevar a cabo una correcta validación externa de los resultados se precisa la supervisión de expertos clínicos para ratificar la veracidad de distintos hotspots. En este paso se comprueba si la fuente de datos generada como propuesta de solución al artefacto realmente describe hotspots.

Sin embargo, debido a que el proyecto OGMIOS - marco en el cual se sitúa este Trabajo de Final de Grado - se encuentra en proceso de evaluación, no se ha pasado a la etapa de validación todavía y, por tanto, no se tiene acceso a expertos que ratifiquen los resultados. Es por ello por lo que esta etapa se ha sustituido por una evaluación de la fuente de datos usando como caso de uso datos de pacientes reales provenientes del Hospital La Fe de Valencia. Esta evaluación pretende medir la precisión y exactitud de la fuente de datos, así como la capacidad de detectar variaciones genómicas presentes en pacientes dentro de hotspots definidos por la fuente de datos desarrollada en este proyecto.

CAPÍTULO 4: ESTADO DEL ARTE

En la actualidad, existe un número limitado de fuentes de datos que proporcionen información sobre hotspots. En este capítulo se describen las fuentes más relevantes en este dominio, y se justifica la necesidad de este Trabajo de Fin de Grado en relación con las fuentes existentes.

4.1 FUENTES DE DATOS DE HOTSPOTS

En este apartado del capítulo se pretende dejar patente la poca presencia de bases de datos o repositorios de hotspots encontrados en la literatura. Tras una búsqueda exhaustiva acerca de bases de datos o repositorios de interés se encontraron dos fuentes en concreto que, a priori, son candidatas a cumplir con las características requeridas.

CancerHotspots es la primera de estas fuentes y esta es una web abierta en la que se puede consultar información relevante de hotspots relacionados con la oncología. Esta página consta de una gran tabla dividida en 6 columnas: (1) gen, (2) residuo, (3) tipo, (4) variaciones, (5) valor Q[3] y (6) número de muestras. Este recurso resulta muy interesante ya que recoge de forma simple toda la información extraída de dos artículos científicos concretos (Chang et al., 2018) y (Chang et al., 2015).

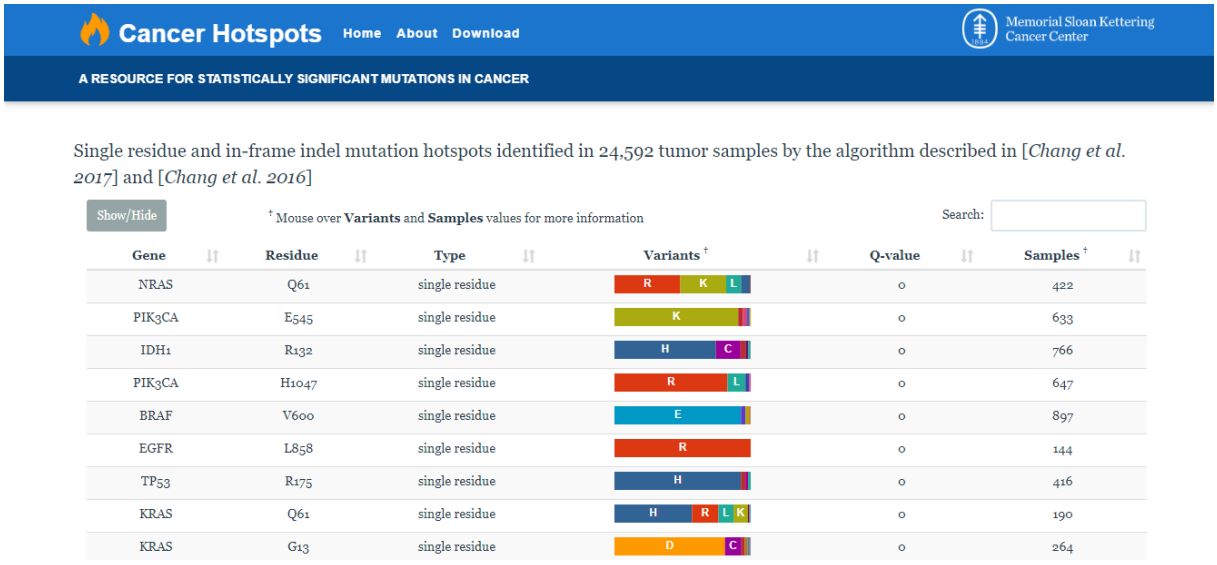


Figura 4: Vista principal de CancerHotspots [Fuente: CancerHotspots]

Anteriormente, también existía una base de datos llamada HotSpotAnnotations, la cual centraba su campo de estudio en la oncología. La información que esta base manejaba era de un repositorio público de cáncer llamado TCGA (*FireBrowse*, n.d.). Esta base de datos obtiene una estimación de dónde se pueden encontrar hotspots a través del empleo de un modelo beta – binomial para estimar alteraciones recurrentes (Trevino, 2000). La vista de usuario de este recurso se puede observar en la Figura 5 y Figura 6. Desafortunadamente, este recurso ya no está disponible y, por tanto, no se puede obtener información de este ni consultarse.

[3] **Valor Q:** Es un tipo de valor p específico para la tasa de descubrimiento falso, siendo esta la proporción de falsos positivos esperada. La diferencia principal entre el valor p y el q, es que el valor p sirve para la probabilidad de un falso positivo en una prueba mientras que el valor q es en múltiples pruebas (▷ *Valor Q: Definición y Ejemplos En 2023 → STATOLOGOS®*, n.d.).

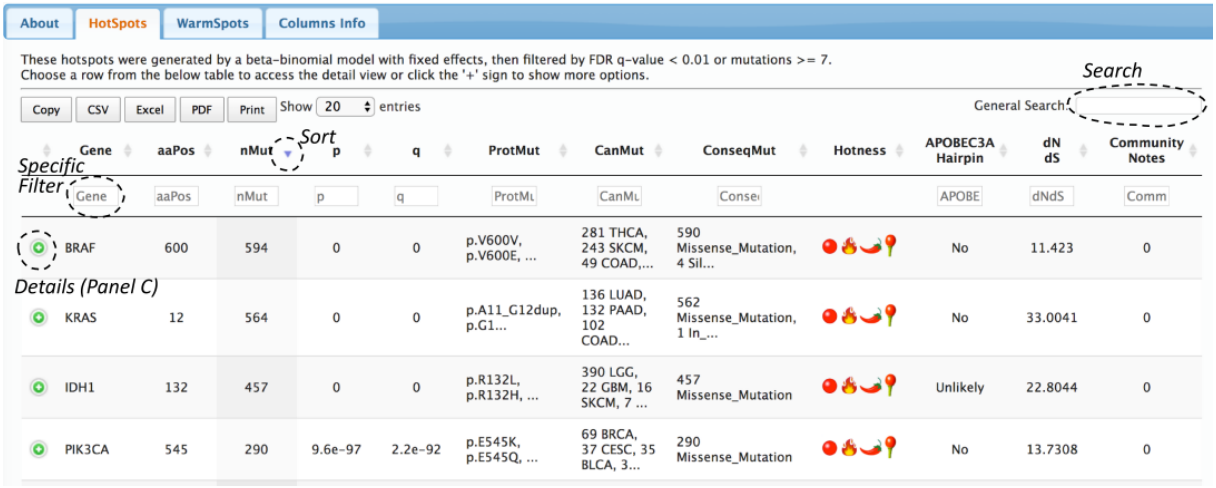


Figura 5: Vista de usuario de HotspotsAnnotation (I). [Fuente: HotspotsAnnotation]

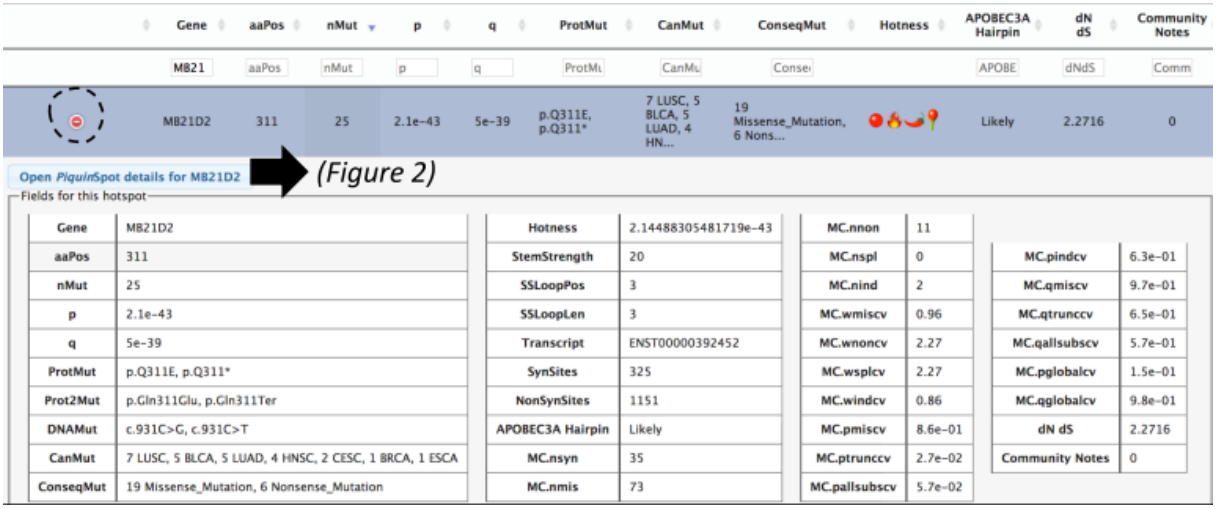


Figura 6: Vista de usuario de HotspotsAnnotation (II) [Fuente: HotspotsAnnotation]

Dado que no existen muchos repositorios específicos sobre hotspots, se realizó un cambio de estrategia, siendo la nueva buscar bases de datos centradas en variaciones genómicas en las que se pudiera encontrar información relacionada con los hotspots para así poderla incluir en la fuente de datos. Para focalizar la estrategia se realizó el análisis de las bases de datos centrándonos en las más importantes dentro de una lista de bases de datos genéticas proporcionada por Oxford Academic Journals (*Database Summary Paper Categories*, n.d.).

Las bases de datos y repositorios candidatos para tener información relacionada con hotspots son:

- **Almena.** Este recurso es una base de datos que recoge de forma exhaustiva variantes genómicas de Oriente Medio y el Norte de África. En Almena se proporciona información sobre más de 26 millones de variantes derivadas de la integración de múltiples estudios de genoma completo y exoma de la región. Asimismo, se proporciona las frecuencias alélicas de las variantes genómicas y aportaciones pertinentes para variantes genómicas clínicamente relevantes(Koshy et al., 2017).
- **CSVS.** ***The Collaborative Spanish Variability Server.*** Este recurso se basa en el *crowsourcing* para proporcionar información sobre variabilidad genómica de la población española a la comunidad científica y médica. Este se utiliza para la filtración de polimorfismos y variaciones

locales con el objetivo de priorizar genes candidatos a enfermedades, almacenando información de 2105 individuos españoles no emparentados (López-López et al., 2023).

- **DbCRID.** Este recurso es una base de datos de reordenamientos cromosómicos en las enfermedades. En él se documenta el tipo de evento, la enfermedad o síntomas asociados a ese evento e información detallada como las posiciones precisas de los puntos de rotura en el genoma, las secuencias de unión, los genes y las regiones alteradas (Kong et al., n.d.).
- **DG – CST.** Esta base de datos es una colección de elementos de secuencia conservados de secuencias genómicas patogénicas humanas comparadas con los murinos (Boccia et al., 2005). En este recurso podemos buscar nuestra región de interés para que nos haga un análisis de este, mostrándose el fenotipo asociado entre otras características.
- **OLIDA.** Este recurso es una base de datos de enfermedades oligogénicas y de las variantes genómicas causantes de esas enfermedades (Nachtegael et al., 2022).
- **GnomAD.** Este recurso está desarrollado por una coalición internacional de investigadores con el objetivo de agregar y armonizar datos de la secuenciación del exoma y del genoma de muchos proyectos(Chen et al., 2022). En este recurso se puede observar la clasificación de las variaciones recogidas en función de la clasificación de las guías ACMG – AMP 2015.
- **GeneCards.** Esta base de datos permite la realización de búsquedas ofreciendo información sobre todos los genes humanos anotados. En esta se integran automáticamente los conocimientos procedentes de 150 fuentes web, incluyendo su información genómica, proteómica, genética, clínica y funcional (Stelzer et al., 2016).
- **Clinvar.** Este recurso es un archivo público de libre acceso de informes sobre las relaciones entre las variaciones humanas y los fenotipos basados en pruebas de apoyo (Landrum et al., 2018). Esta base de datos incluye aspectos como la localización genómica, el tipo de variación o las líneas germinales.
- **Ensembl.** Es un buscador de genomas de vertebrados. Uno de los aspectos interesantes de esta base de datos es que contiene información de genes con posibles alineaciones prediciendo funciones reguladoras e información sobre enfermedades de distintas especies (Cunningham et al., 2022).
- **GWAS Catalog.** Esta base de datos recopila información de estudios de asociación de genoma. Esta herramienta sirve, entre otros aspectos, para identificar loci genéticos asociados con enfermedades a través del análisis de variantes categorizadas en todo el genoma (Sollis et al., 2023) .

Una vez estudiadas las posibles fuentes de información sobre hotspots, se concluye que únicamente GeneCards contiene información sobre información asociada a los hotspots. En este recurso, se menciona la existencia de un hotspot en un gen en concreto únicamente si en alguna de las descripciones del gen se menciona dicho término, tal y como puede observarse en la Figura 7 y en la Figura 8.

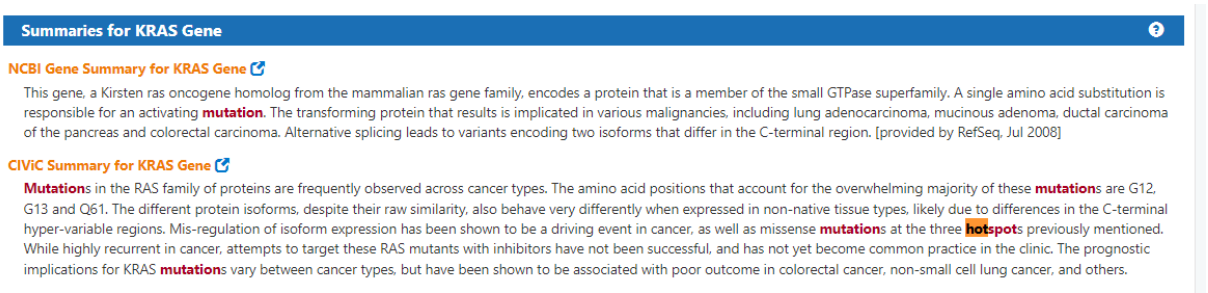


Figura 7: Cómo encontrar información sobre hotspots en el gen KRAS en GeneCards [Fuente: GeneCards]

	Title	Association	Year	# Cit
⊕	TP53 **mutation**s predict disease control in metastatic colorectal cancer treated with cetuximab-based chemotherapy. *British journal of cancer* Oden-Gangloff A ... Frebourg T (PMID:19367287)	Gene [3 39 21], GeneVariant [108], Disorder [69]	2009	44
⊕	Germline TP53 **mutation**s in BRCA1 and BRCA2 **mutation**-negative French Canadian breast cancer families. *Breast cancer research and treatment* Arcand SL ... Tonin PN (PMID:17541742)	Gene [3 39 21], GeneVariant [108], Disorder [69]	2008	16
⊕	BRCA1, BRCA2 and TP53 **mutation**s in very early-onset breast cancer with associated risks to relatives. *European journal of cancer (Oxford, England : 1990)* Lalloo F ... Evans DG (PMID:16644204)	Gene [3 39 21], Disorder [69], GeneVariant [108]	2006	51
⊕	Novel germline **mutation**s in breast cancer susceptibility genes BRCA1, BRCA2 and p53 gene in breast cancer patients from India. *Breast cancer research and treatment* Hedau S ... Das BC (PMID:15564800)	Gene [3 39 21], GeneVariant [108], Disorder [69]	2004	26
⊕	TP53, BRCA1, and BRCA2 tumor suppressor genes are not commonly mutated in survivors of Hodgkin's disease with second primary neoplasms. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* Nichols KE ... Diller L (PMID:14673037)	Gene [3 39 21], GeneVariant [108], Disorder [69]	2003	10
⊕	Hereditary TP53 codon 292 and somatic P16INK4A codon 94 **mutation**s in a Li-Fraumeni syndrome family. *Cancer genetics and cytogenetics* Güran S ... Imirzalioğlu N (PMID:10484981)	Gene [3 21], Protein [4], GeneVariant [108], Disorder [69]	1999	1
⊕	Hereditary and acquired p53 gene **mutation**s in childhood acute lymphoblastic leukemia. *The Journal of clinical investigation* Felix CA ... Whang-Peng J (PMID:1737852)	Gene [3 21], Protein [4], GeneVariant [108], Disorder [69]	1992	12
⊕	**Mutation**al **hotspot** in the p53 gene in human hepatocellular carcinomas. *Nature* Hsu IC ... Harris CC (PMID:1849234)	Gene [3 21], Protein [4], Disorder [69], GeneVariant [108]	1991	300

Figura 8: Cómo encontrar información acerca de hotspots en el gen TP53 en GeneCards. [Fuente: GeneCards]

Tal y como puede observarse en las figuras anteriores, la información acerca de hotspots que ofrece GeneCards se basa exclusivamente en la existencia de documentación científica acerca de ese hotspots o en la descripción que recursos externos como ClVic hacen sobre el gen que se esté estudiando en cada caso.

Si se analizan los resultados del estado del arte, se puede observar cómo la comunidad científica no ha creado casi bases de datos o fuentes de datos sobre hotspots, y menos aún sobre hotspots de cardiopatías. Ese es el motivo por el cual nace esta iniciativa, cuyo proceso de diseño y desarrollo va a ser descrito detalladamente en los capítulos posteriores.

CAPÍTULO 5. INVESTIGACIÓN DEL PROBLEMA

En este capítulo se lleva a cabo la primera etapa de la metodología empleada, que consiste en la caracterización del problema práctico. De esta manera, el capítulo se va a dividir en diferentes secciones con la intención de ir definiendo las cuestiones concretas abordadas en esta etapa. En la sección 5.1 se van a definir los usuarios objetivos de esta fuente de datos, para así poder acotar los principales beneficiarios y poder ofrecer una fuente de datos que tenga en cuenta sus necesidades. A continuación, en la sección 5.2, se explican las necesidades de estos usuarios que se ha buscado cubrir con esta fuente de datos. Una vez definidos los usuarios objetivos y las necesidades que cubre, se definen de forma precisa que es un hotspot y a su rol en la interpretación de las variaciones genómicas en las secciones 5.3 y 5.4, respectivamente.

5.1. USUARIOS OBJETIVO

Este proyecto pretende obtener e integrar información sobre hotspots asociados a cardiopatías familiares para generar una fuente de datos que permita caracterizarlos de una forma clara. Es por ello por lo que, teniendo en cuenta el propósito de este trabajo, la fuente de datos está dirigida a dos tipos de usuarios con el mismo objetivo: encontrar información útil dentro del caos de datos genómico acerca de los hotspot.

El primer tipo de usuario son los expertos clínicos, pues estos profesionales son usuarios de la información genética en su día a día. El interés de estos profesionales por emplear la información genómica para el diagnóstico en función de las variaciones genómicas presentes en un paciente les convierte en un claro usuario objetivo. La posibilidad de acceder a información sobre hotspots de manera rápida y sencilla les facilitará el diagnóstico preventivo y precoz de enfermedades.

El segundo tipo de usuario son los analistas de datos genómicos, responsables de la gestión de la información genómica para generar conocimiento acerca de fenotipos, genes o variaciones concretas. Los analistas de datos usan toda esta información en un ámbito de investigación no en un ámbito clínico, pues su cometido es generar información de interés para los expertos clínicos. Es por ello por lo que esta fuente de datos que se plantea en este Trabajo de Fin de Grado es una herramienta potente para estos expertos.

Tras la definición de los usuarios objetivos, en la siguiente sección se pasa a definir las necesidades de los usuarios a cubrir por la fuente de datos que se ha diseñado en este Trabajo de Fin de Grado.

5.2. NECESIDADES A CUBRIR

Tal y como queda expuesto en el apartado 1.2, la información genética se caracteriza por presentarse de una forma heterogénea y compleja. Ese es el motivo por el cual la sociedad actual se ve envuelta en una situación de *caos de datos genómicos*. Este concepto es ampliamente utilizado en el ámbito de la genómica para describir el aumento exponencial de datos genómicos disponibles y la falta de estructuración, homogeneización y aplicabilidad de estos. Este hecho supone la pérdida de mucha información relevante dentro del ámbito clínico, dificultando el uso de la genómica como método diagnóstico.

El fenómeno de caos de datos genómicos expuesto anteriormente genera varios problemas para los investigadores y para los expertos clínicos. Por una parte, los investigadores se enfrentan a una

heterogeneidad de los datos, lo que les dificulta el entendimiento, automatización y generalización de la información. Por otra parte, los clínicos ven complicada la toma de decisiones, afectando a la calidad de los diagnósticos y tratamientos que son capaces de ofrecer.

Por consiguiente, una de las necesidades primordiales a cubrir por esta fuente de datos es la estandarización de la información. El concepto de estandarizar en este contexto se entiende como la necesidad de presentar toda la información estructurada de la misma forma y bajo los mismos parámetros y estándares, con el propósito de una correcta presentación de la información relacionada con los hotspots.

Asimismo, también es importante la integración de toda esta información para que los expertos no tengan que emplear su tiempo en la búsqueda exhaustiva de información que se encuentra dispersa en múltiples fuentes. La fuente de datos que se plantea en este Trabajo de Fin de Grado pretende ser una vía para facilitar el acceso a toda la información disponible acerca de este tema de estudio, de forma que los expertos reduzcan considerablemente el tiempo que necesitan para obtener la información de interés. Esta necesidad está estrechamente relacionada con la de proporcionar un fácil acceso a la información ya que, con la integración de toda la información disponible se consigue acotar la búsqueda y se garantiza un fácil acceso a la información.

Finalmente, la última necesidad a cubrir de este Trabajo de Final de Grado es el de la caracterización de un hotspot. El motivo de este reside en la poca precisión existente en la literatura acerca de cómo identificar y caracterizar un hotspot. El hecho de que la comunidad científica no haya sido capaz de definir de forma unánime las características relevantes de un hotspot, supone que la información extraída sobre estos se represente de forma muy heterogénea. Por ejemplo, algunos documentos reportan la posición de un hotspot en un determinado transcrito, otras mediante la posición proteica, y otras mediante la posición en el cromosoma. Esta diferencia de representación se traduce en un esfuerzo extra para los investigadores, pues estos se convierten en los responsables de estandarizar y caracterizar la información que tienen a su alcance para poder estudiarla de forma conjunta. Por este motivo, una de las necesidades que pretende cubrir este trabajo es el de la caracterización del concepto de hotspots. En la Figura 9, se presenta un esquema de todas las necesidades que esta fuente de datos pretende cubrir.

Figura 9: Esquema de las principales necesidades cubiertas por la fuente de datos. [Fuente: Elaboración propia]

5.3. HOTSPOTS

El término hotspot se utiliza en diferentes disciplinas científicas como la geofísica, las ciencias ambientales y la genética. En estas disciplinas, el término se define como un área de gran actividad o importancia (Rodrigues, 2013). En el ámbito de estudio de este Trabajo de Fin de Grado, la genética, el término hotspot tiene una definición imprecisa dentro de la comunidad científica.

Resulta fundamental no confundir dos términos que, a priori, pueden parecer similares, siendo estos el de hotspot de recombinación y el de hotspot mutacional. Un hotspot de recombinación se puede definir como regiones del genoma donde la frecuencia de recombinación varía entre 1 y 2 kb (Paul et al., 2016). Estas regiones no están distribuidas de forma uniforme en el genoma humano y son las causantes de distorsiones en el mapa genético. Otra definición de hotspot de recombinación los define como regiones locales dentro de cromosomas en las que la recombinación [4] suele estar concentrada y normalmente se encuentra entre coldspots, que son regiones de poca frecuencia recombinatoria (Choi & Henderson, 2015)(Lichten & Goldman, 2003). Según este último estudio, se comienza a considerar hotspot cuando la frecuencia de recombinación supera el 1cM por mega base de ADN (Lichten & Goldman, 2003).

El término de hotspot mutacional tiene una definición más amplia y difusa que el anterior. Revisando literatura científica sobre este concepto se puede observar como todos ellos lo definen de una forma muy general, sin entrar en detalles concretos como la determinación de cuantas variaciones genómicas patogénicas son necesarias para clasificar la región como un hotspot. Sin embargo, este concepto resulta fundamental para el diagnóstico de enfermedades hereditarias y, por tanto, es este el término que buscamos caracterizar en este Trabajo de Final de Grado.

Una de las primeras definiciones de hotspot es la que propone la Fundación Instituto Roche, en la que los hotspot son definidos 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 se observan mutaciones con más frecuencia de lo habitual"(*Fundación Instituto Roche - Glosario de Genética - Puntos Calientes de Mutación*, n.d.).

Otra posible definición es la que propone la Guía de aplicación clínica de la secuenciación masiva en síndromes mielodisplásicos y leucemia mielomonocítica crónica, la cual define un hotspot como una "zona dentro del genoma propensa a ser alterada y en la cual detectan variantes más frecuentes. Dicha región puede comprender un solo nucleótido, un codón o un exón" (*Guía de Aplicación Clínica de La Secuenciación Masiva En Síndromes Mielodisplásicos y Leucemia Mielomonocítica Crónica – GCECGH*, n.d.) .

Del mismo modo, este término se encuentra definido en la revisión de Rogozin et al (2003) como "posiciones de nucleótidos con una frecuencia de mutación excepcionalmente alta"(Rogozin & 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[5] o como zonas con mecanismos específicos de mutación.

En definitiva, podemos observar como este término presenta definiciones muy similares a la par que abiertas. Esto se debe a que, a diferencia de los hotspot recombinantes, no existe ningún parámetro

[4] **Recombinación homóloga**: Tipo de recombinación genética en el que se intercambian secuencias de nucleótidos entre dos moléculas parecidas o idénticas de ADN (*Recombinación Homóloga*, n.d.).

[5] **Mutágeno**: Sustancia química o agente físico capaz de incidir en cambios del ADN llamados mutaciones.(*Mutágeno*, n.d.).

cuantitativo que ayude a la identificación y caracterización de los hotspot mutacionales. El hecho de no tener ninguna fuente que determine y caracterice un hotspot de forma clara y precisa supone uno de los retos principales de este proyecto.

5.4. HOTSPOT EN LA INTERPRETACIÓN DE VARIACIONES

El *American College of Medical Genetics and Genomics* o ACMG, junto con *Association for Molecular Pathology* o AMP han desarrollado unas guías para interpretar el rol de variaciones genómicas en el desarrollo de enfermedad(Richards et al., 2015) . Estas guías han sido diseñadas por directores de laboratorios y expertos clínicos en general, y se proponen la clasificación de las variaciones en cinco tipos: (1) Patogénica*,* (2) Probablemente patogénica*,* (3) Significado incierto, (4) Probablemente benigna*,* y (5) Benigna (Richards et al., 2015)*.* Esta clasificación sirve para reflejar como actúa esa variación en una enfermedad en concreto, ya que no siempre una variación genómica implica patogenicidad.

Para determinar la clasificación más adecuada para cada variación, estas guías definen los criterios de clasificación, que se dividen en cuatro categorías: (1) Muy fuerte, (2) Fuerte, (3) Moderado y (4) De soporte, según el nivel de evidencia que proporcionen para la clasificación.

Las guías ACMG- AMP del 2015 son las más utilizadas a la hora de clasificar variaciones genómicas. Estas guías determinan que la localización de una variación en un hotspot ofrece una evidencia moderada de patogenicidad, asignada al criterio PM1. Además, también existen otras guías de clasificación de variaciones genómicas en las que se contempla la evaluación de si una variación se localiza en un hotspot. Algunos ejemplos son:

- La clasificación de variantes en el *Illumina Clinical Services Laboratory (ICSL)*(*Illumina Clinical Services Laboratory Assertion Criteria for Gene Curation*, n.d.)
- La clasificación de variantes del *Praxis for Muangenetik Tübingen* (*Zentrum Für Humangenetik - Experten Für Genetische Diagnostik*, n.d.)
- La clasificación de variantes propuestas conjuntamente por el *Association for Molecular Pathology,* la *American Society of Clinical Ontology* y *College of American Pathologists* (Li et al., 2017).
- La clasificación refinada de las guías ACMG-AMP del 2015 plasmadas en Sherloc (Nykamp et al., 2017).

Todas estas guías se basan en la clasificación que se propone en las ACMG – AMP del 2015 y todas pretenden mejorar, de una forma o de otra, la falta de especificidad que estas presentan o ajustarlas a un campo de conocimiento en concreto, como puede ser la oncología. En el caso de estudio de este Trabajo de Final de Grado, se han tomado las guías ACMG – AMP del 2015 como referencia, por el hecho de ser estas la base de todas las demás clasificaciones de variaciones genómicas.

5.5. BASE CONCEPTUAL: MODELADO CONCEPTUAL DEL GENOMA

Un modelo conceptual en el ámbito de los sistemas de información se define como la descripción del conocimiento del dominio en el cual se desarrollará un sistema de información (Olivé, 2007). Un modelo tiene como objetivo principal comprender a fondo el dominio de estudio al definir las distintas entidades y sus relaciones.

Este trabajo se enfoca en el ámbito genómico, por lo que se requiere de un modelo conceptual que describa este dominio. En el grupo de investigación PROS se ha creado un modelo conceptual, conocido como Modelo Conceptual del Genoma Humano (CSHG), que describe el conocimiento genómico en toda su complejidad. Este modelo surgió a partir de la tesis titulada *Diseño y desarrollo de un sistema de información genómica basado en un modelo conceptual holístico del genoma humano* del Dr. José Fabián Reyes Román (Reyes Román,2018 y Reyes Román et al.,2016). Debido a su adecuación al ámbito de este Trabajo de Fin de Grado, se va a utilizar el CSHG como base para el diseño y desarrollo de la fuente de datos objetivo.

El CSHG ofrece una perspectiva holística los conceptos clave para entender el genoma humano. Es por ello por lo que, desde el instituto PROS de la UPV, se trabaja de forma continua en mantener y mejorar este modelo para adaptarlo a los cambios continuos que aparecen en el dominio a medida que avanza la investigación y el conocimiento sobre el genoma (Bernasconi et al., n.d.) (García S et al., 2022)(Palacio et al., 2018) (Reyes Román et al., 2016)(Alberto Garcia et al., 2021). En su estado actual, el CSHG se divide en cinco vistas que describen las distintas dimensiones del dominio genómico:

1- **Vista Estructural.** En esta se describe la estructura del genoma humano representando los elementos básicos del ADN.
2- **Vista de Variación.** En esta vista se caracterizan los cambios, o variaciones, que se pueden encontrar en la secuencia del genoma.
3- **Vista de Transcripción.** En esta se modela el proceso de transcripción y la síntesis de proteínas.
4- **Vista de Bibliografía.** En esta vista se describen las fuentes de información y la bibliografía relacionadas con las distintas entidades del CSHG.
5- **Vista de Rutas Metabólicas.** En esta se representan los conceptos relacionados con los procesos biológicos del interior de la célula.

Volviendo al caso de estudio, los hotspots, son un término de definición abierta, tal y como se ve en el apartado 5.3 y, por este motivo, el concepto de hotspot todavía no había sido estudiado en profundidad para su inclusión en el CSHG. Es por ello por lo que, en el marco de este Trabajo de Fin de Grado se decidió incluir los hotspots dentro del CSHG y así poder obtener una caracterización precisa de los mismos.

5.6. BASE METODOLÓGICA: MÉTODOLOGÍA SILE

Trabajar con información heterogénea y dispersa de forma eficiente requiere de una metodología que guíe el procesado de la información. En este trabajo se va a utilizar la metodología SILE, creada en el grupo PROS como resultado de la tesis doctoral de la Dra. Ana León Palacio titulada *SILE:A Method for the Efficient Management of Smart Genomic Information*(León Palacio, 2019) . La metodología define 4 etapas que garantizan un procesado eficiente de la información: ***S**earch* (Búsqueda)*, **I**dentify* (Identificación)*, **L**oad* (Carga) y ***E**xplotation* (Explotación). En la Figura 10 se citan las etapas del método SILE y el principal cometido de cada una de ellas.

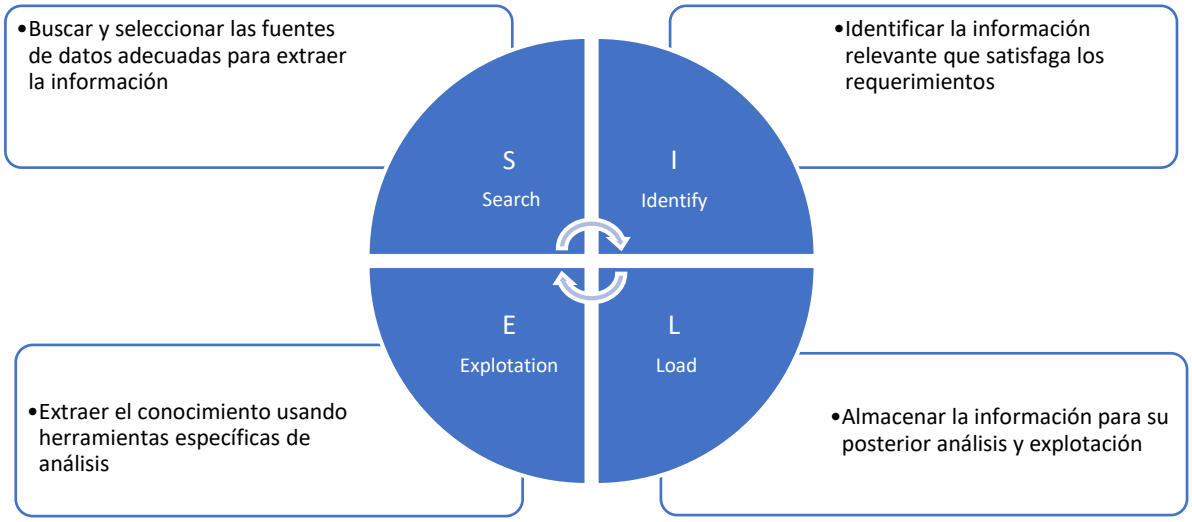


Figura 10: Etapas del método SILE . [Elaboración propia]

En referencia a la Figura 10, la primera etapa del método es la correspondiente con la S o búsqueda. El principal objetivo de esta es seleccionar las fuentes de datos adecuadas para obtener la información de interés. En la actualidad, de acuerdo con NAR Catalog (*NAR Catalogs and Product Brochures | North American Rescue*, n.d.)s | North American Rescue, n.d.), existen 1764 repositorios con información biológica, a lo que se tiene que sumar la información disponible únicamente en la literatura científica(Rigden et al., 2023). Por tanto, esta etapa resulta fundamental para evitar la pérdida de información potencialmente relevante.

La segunda etapa del método es la correspondiente con la I o identificación, en la cual se pretende identificar la información relevante que satisfaga los requerimientos del trabajo. En esta etapa se deben llevar a cabo las siguientes tareas: (1) Identificar los datos relevantes a extraer, tarea que es recomendable esté guiada por modelos conceptuales, (2) Completar los datos con el objetivo de que la información sea lo más completa posible, y (3) Representar la información de forma estandarizada haciendo uso, por ejemplo, de ontologías de dominio[6].

La tercera etapa es la correspondiente con la L o carga de información, en la cual se extrae el conocimiento buscado usando herramientas específicas. Este proceso necesita de una validación de los resultados para garantizar la calidad de la información extraída. Finalmente, la cuarta etapa es la correspondiente con la E o explotación, en la que se almacena la información para su posterior análisis y explotación.

[6] **Ontología de dominio:** Representación de conceptos pertenecientes a una parte específica del mundo, considerándose esta una herramienta de gestión de conocimiento altamente especializado. Estas ontologías dentro del ámbito científico y tecnológico controlan el vocabulario que se emplea para representar y computar el contenido disponible de los recursos digitales.(*Ontologías y Vocabularios Controlados - GNOSS*, n.d.).

CAPÍTULO 6. DISEÑO Y DESARROLLO DE LA SOLUCIÓN

Este capítulo se centra en el diseño y desarrollo de la fuente de datos, explicando más profundamente todos los pasos necesarios para caracterizar un hotspot basado en el modelo conceptual del genoma. Asimismo, se explican cada una de las etapas del método SILE aplicado a nuestro caso de estudio.

6.1. APLICACIÓN DEL MODELO CONCEPTUAL: CARACTERIZACIÓN DE UN HOTSPOT.

Los modelos conceptuales facilitan el entendimiento de dominios complejos tales como el genómico (Jarke & Quix, 2017) y por este motivo, una caracterización precisa de un hotspot pasa por la generación de su modelo conceptual. El empleo del modelo conceptual resulta imprescindible dado que este permite una gestión efectiva y eficiente de los datos genómicos, así como permite el avance en el entendimiento del genoma humano a través de una representación gráfica.

Otra de las ventajas que presenta es la gran capacidad de adaptación ya que, conforme se vayan añadiendo conocimientos y conceptos nuevos acerca de los hotspots mutacionales, el haberlos representado mediante un modelo conceptual, nos permitirá añadir y modificar aspectos de este de una forma rápida y sencilla (Fabián Reyes Román & Pastor López, n.d.).

La propuesta de modelo conceptual asociado al concepto de hotspot es el resultado de la modificación del CSHG realizado por el grupo PROS y el resultado de este se expone en la Figura 11.

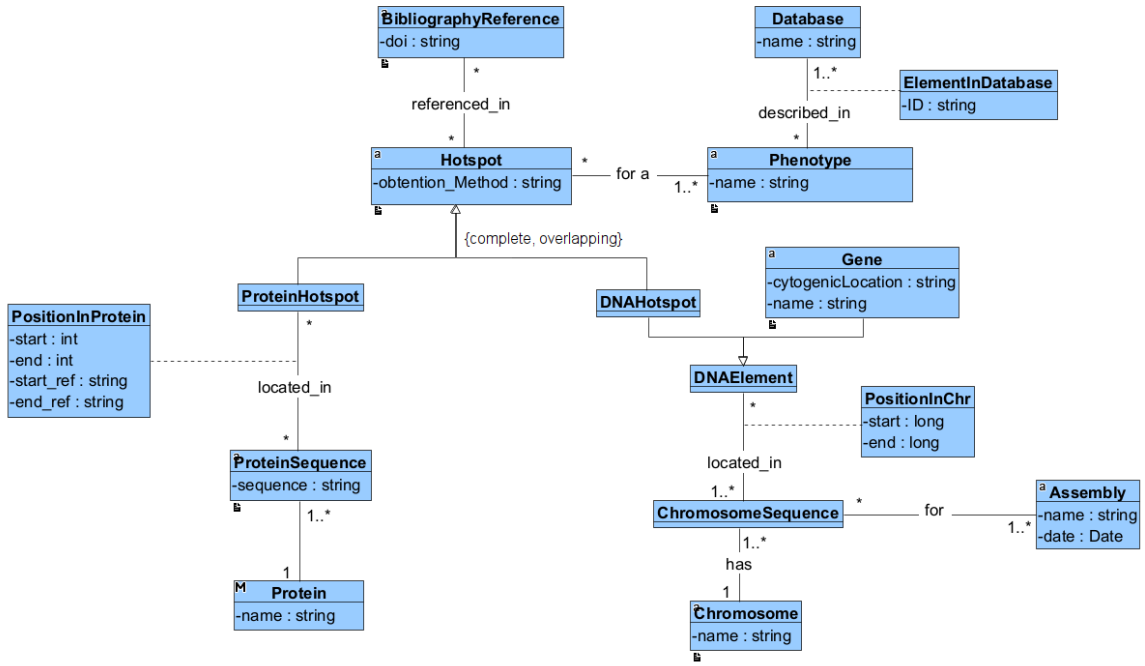


Figura 11: Modelado conceptual de hotspot. [Elaboración propia]

En la propuesta de modelo conceptual de hotspot de este Trabajo de Final de Grado se puede observar que el concepto de hotspot se define dentro de dos marcos: el proteico y el genómico.

El hotspot definido dentro del marco proteico presenta 3 grandes clases: posición en la proteína (*PositioninProtein*), secuencia proteica (*ProteinSequence*) y la propia proteína (*Protein*). Dentro de la

clase de posición en la proteína, para definirlo se necesitan 4 variables: inicio proteico (*start*), fin proteico (*end*), referencia al inicio (*start_ref*) y referencia al fin (*end_ref*) . El inicio y el fin proteico se refiere a la posición del aminoácido de inicio y fin en cada caso. Estas variables se definen como *integers* ya que estos son números enteros. El inicio y fin de referencia se refiere al aminoácido de referencia de inicio y de fin de cada posición. La representación de estas características se presenta como *strings,* ya que estos se expresan mediante su acrónimo - por ejemplo, Gly refiriéndose a Glicerina.

Asimismo, es importante definir también la secuencia respecto a la que se definen dichas posiciones y para ello se hace uso de los ENSP (*sequence*), unos identificadores de secuencia de proteína provistos por UniProtKB y expresados en el modelo como una variable de tipo *string.* La importancia de añadir estos identificadores en el modelo reside en el hecho de que es posible que haya más de un transcrito[7] asociado al mismo gen entonces, se pueden generar distintas versiones o isoformas de esa proteína. Es por ello por lo que el empleo de estos identificadores resulta crucial para evitar confusiones de sobre que secuencia se está definiendo en ese hotspot.

Finalmente, dentro de la definición de un hotspot en el ámbito proteico tenemos la definición de la propia proteína en la que se encuentra. Para caracterizarla se emplea el UniProtID (*name*), ya que este es un identificador único de proteína que nos permite identificar de forma inequívoca la proteína a la que se hace referencia.

En el ámbito genómico también existen diferentes clases imprescindibles para definir un hotspot: gen (*Gene*), posición cromosómica (*PositioninChr*), *assembly (Assembly)* y cromosoma (*Chromosome*). Dentro de la clase gen necesitamos dos características imprescindibles, el nombre de este (*name*) y la localización citogenética (*cytogeneticLocation*). Esta última es un indicador único de la posición dentro del cromosoma en la que se encuentra el gen de interés. Otra variable imprescindible es la posición cromosómica. Esta variable se caracteriza mediante el inicio (*start*) y el fin (*end*) genómico y nos indican el inicio y el fin del hotspot dentro de la secuencia genómica, es por ello por lo que estos atributos son *long* .

Asimismo, tenemos el *assembly* como tercera clase para caracterizar un hotspots dentro del ámbito genómico. Este atributo se refiere a la secuencia de referencia del genoma del cual se ha extraído toda la información anterior. Actualmente tenemos dos opciones, el *assembly* del 37 y el *assembly* del 38. Este trabajo se centra en el *assembly* del 37 puesto que es del que se dispone de más información relacionada. Finalmente, para definir el hotspot necesitamos saber en qué cromosoma sucede, por lo que la cuarto clase que define un hotspot dentro del ámbito genómico es el del cromosoma (*name*).

Es importante recalcar que cada hotspot está asociado a un fenotipo en concreto y que, por tanto, este aspecto también es importante a la hora de caracterizar un hotspot y queda representado por la clase *Phenotype* la cual posee una variable llamada *name* de tipo *string*. Asimismo, este fenotipo [8] debe venir descrito por una base de datos la cual podrá identificarse con su nombre y por una serie de elementos en la base de datos (*ElementInDatabase*) los cuales se identifican mediante un identificador (*ID*). Finalmente, dentro de este modelo conceptual de un hotspot, tenemos que añadir la referencia bibliográfica (*Database*) de la cual hemos obtenido la información, la cual se caracteriza por el doi

[7] **Transcrito**: Molécula de ARN monocatenario que se obtiene inmediatamente después de la transcripción(*Fundación Instituto Roche - Glosario de Genética - Transcrito Primario*, n.d.).

[8] **Fenotipo**: Rasgos observables de una persona determinados a partir de su composición genómica (genotipo) y otros factores (*Fenotipo*, n.d.).

(*name*) del documento científico al que hace referencia . En la Tabla 1, se puede observar todos los atributos empleados para la caracterización de un hotspot y el tipo de información que aportan.

Tabla 1: Atributos escogidos para la caracterización de un hotspot para esta fuente de datos

ATRIBUTO	EXPLICACIÓN DEL ATRIBUTO	TIPO
Cromosoma	Cromosoma en el que se encuentra.	*String*
Localización citogenética	Localización cromosómica en el que se encuentra.	*String*
Gen	Nombre del gen al que pertenece.	*String*
UniProtID	Referencia a la proteína de UniprotKB.	*String*
Comienzo AA[9]	Posición proteica del aminoácido de referencia de inicio.	*Integer*
Final AA	Posición proteica del aminoácido de referencia de final.	*Integer*
Referencia de comienzo AA	Aminoácido asociado de referencia de inicio.	*String*
Referencia de final AA	Aminoácido asociado de referencia de final.	*String*
UniprotKB ENSP	Identificador único de secuencia proteica.	*String*
Comienzo GRCh37	Posición genómica de inicio del *assembly* GRCh37.	*Long*
Final GRCh37	Posición genómica de final del *assembly* GRCh37.	*Long*
Método/ fuente estandarizada	Método de experimentación o fuente de extracción de información estandarizada con las ontologías ECO y OBI.	*String*
Fenotipo estandarizado	Fenotipo asociado estandarizado con las ontologías MONDO y HP.	*String*
DOI	Referencia bibliográfica al documento científico.	*String*

6.2. BÚSQUEDA DE INFORMACIÓN

La búsqueda de la información se corresponde con la primera etapa del método SILE, es decir, con la S. En esta etapa se pretende encontrar la información adecuada para poder desarrollar la fuente de datos, en base a nuestro modelo conceptual de los hotspots. Para acotar el caso de estudio, se parte

[9] **AA** : aminoácido.

de una lista de 457 genes que potencialmente pueden contener hotspots relacionados con cardiopatías y que ha sido proporcionada por el Hospital La Fe de Valencia (Anexo 1).

Esta etapa del método SILE se basa en el diagrama de flujo de la Figura 12 , el cual pretende explicar de forma esquemática todos los pasos seguidos para conseguir buscar toda la información acerca de hotspots de cardiología de la literatura.

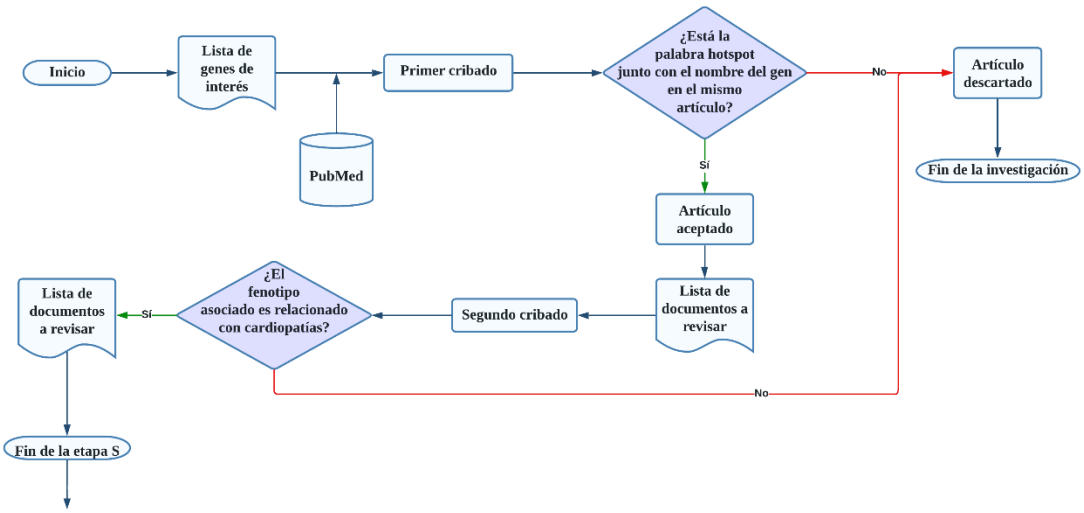


Figura 12:Diagrama de flujo asociado a la etapa de búsqueda del método SILE en este trabajo. [Elaboración propia]

El diagrama de flujo anterior parte de la lista de los 457 genes provenientes del hospital La Fe de Valencia, los cuales constituyen la base del estudio. El primer proceso llevado a cabo es el de la búsqueda de documentación científica específica de los genes de interés provistos para ver si estos presentan hotspots asociados a cardiopatías. En este caso, se acota la búsqueda a PubMed, dado que se trata de una base de datos de acceso libre especializada en ciencias de la salud, con más de 19 millones de referencias bibliográficas (Trueba-Gómez & Estrada-Lorenzo, 2010). Tras esto se define una regla de búsqueda más concreta para encontrar la información que se necesita, siendo esta:

$$(\boldsymbol{hotspot})\ \boldsymbol{AND}\ (\boldsymbol{nombre\ del\ gen})$$

Ecuación 1: Regla de búsqueda en PubMed.

Si el resultado de la aplicación de esta regla resulta positivo y se encuentran artículos científicos de genes relacionados con hotspots, se acepta el artículo estudiado y se añaden esos genes con los documentos a una lista de posibles artículos a revisar. Sin embargo, si el resultado de esta regla es negativo, se descarta el artículo y se acaba el proceso de investigación con ese artículo en concreto.

Una vez se aplica esta regla a todos los artículos obtenidos encontrados, se pasa al segundo cribado. Este consiste en la revisión del resumen de cada documento candidato a obtener información acerca de hotspots relacionados con cardiopatías, que es nuestro caso de estudio. Si el resultado de esta revisión es positivo, se genera otra lista de documentos a revisar y, si es negativo, se descarta el documento y se finaliza el proceso de investigación.

Después de realizar todo este proceso con cada uno de los genes candidatos, se obtiene como resultado de esta fase una lista con documentación científica de genes asociados a hotspots de cardiopatías, dando así por finalizada la etapa de búsqueda. Con esta lista de documentación, se puede pasar a la segunda etapa del método SILE, en la cual se identifica la información relevante según el modelo conceptual del genoma que se precisa para caracterizar un hotspot.

6.3. IDENTIFICACIÓN DE LA INFORMACIÓN

Gracias a la etapa anterior se consigue extraer de la bibliografía los documentos científicos más relevantes con información sobre hotspots de cardiopatías. Sin embargo, toda la información contenida en los documentos no tiene por qué ser clave a la hora de definir un hotspot en base al modelo conceptual. Es por ello por lo que el objetivo de la etapa de identificación es el de seleccionar solo aquella información que necesitamos para caracterizar un hotspot en base al modelo conceptual del mismo. El proceso que se ha seguido para extraer la información queda resumido en el diagrama de flujo de la Figura 13.

El primer paso de la extracción de datos consiste en la lectura de los artículos seleccionados en la etapa previa para ver si contienen la información que describa un hotspot. Si la tienen se procede a la extracción de datos y si no la tienen de forma directa, se estudia el artículo más en profundidad para ver si es posible extraer información clave de forma indirecta. Si la respuesta a la pregunta "¿Se puede encontrar la información?" es positiva, buscamos la información en las cuatro herramientas auxiliares seleccionadas para este propósito: (1) SynVar, (2) RefSeq, (3) Clinvar y (4) UniProtKB .

Una vez extraída la información encontrada en las herramientas auxiliares, se recopila con la encontrada de forma directa para tener una visión global de todos los datos encontrados acerca de ese hotspot. Si por el contrario la respuesta es negativa, descartamos el artículo y se termina el proceso. Finalmente, una vez se ha encontrado toda la información de forma directa e indirecta, se recopila y pasa por un proceso de estandarización mediante el uso de identificadores comunes como son el UniProtID y el ENSP, y el uso de diferentes ontologías como MONDO, ECO, HP y OBI, para garantizar la homogeneización de la información que se ofrece.

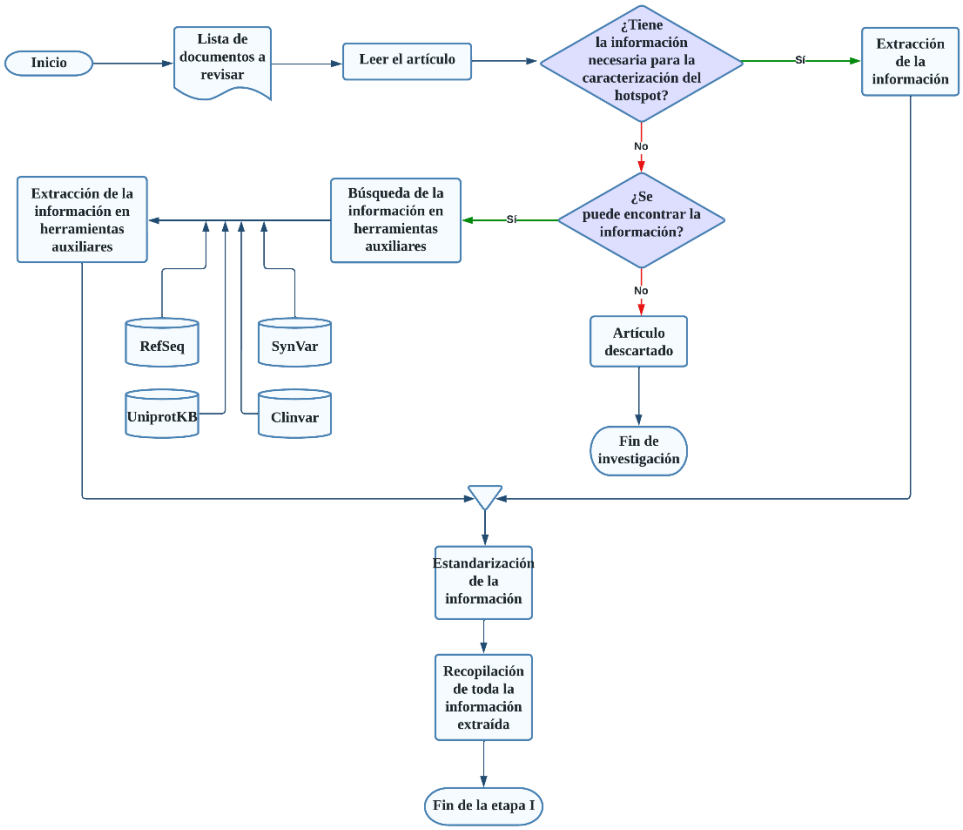


Figura 13:Diagrama de flujo asociado a la etapa de identificación del método SILE en este trabajo. [Elaboración propia]

Este diagrama de flujo anterior puede dividirse en 3 etapas correspondiéndose estas con los procesos que se deben llevar a cabo durante la extracción de datos. La primera etapa se basa en la recopilación de la información, la segunda se basa en la selección de la información en base al modelo conceptual y la tercera se basa en la estandarización de la información. A continuación, se explica cada una de ellas en detalle.

6.3.1. Etapa 1: Recopilación de información disponible

La primera etapa del desarrollo de la fuente de datos consiste en recopilar toda la información descriptiva del hotspot presente en el documento estudiado en cada caso, sin tener en cuenta si esta información se ajusta o no a las clases del modelo conceptual. La extracción de toda la información disponible se debe a que, en muchos casos, la información que precisa el modelo conceptual se puede sacar a partir de información que aparece en el documento.

Una forma muy clara en la que se ve como la información de los documentos puede adecuarse al modelo conceptual es, por ejemplo, la extracción de la base de referencia o la alterada mediante el uso de las notaciones estándar de las variaciones genómicas. En estas notaciones se indican diferentes aspectos como por ejemplo la posición genómica, el tipo de secuencia que se expresa, el tipo de variación y la base de referencia y la base alterada. Es por ello por lo que, sabiendo como se estructuran estas notaciones se puede extraer información primordial para el modelo como es, por ejemplo, la base alterada de la variación.

Es por ello por lo que la fuente de datos en la primera etapa tenía un aspecto un poco distinto al de la etapa final, la cual se rige por los atributos del modelo conceptual explicado en el apartado 6.1. En la primera etapa se incluían otras variables que en su momento se recogieron para no perder información que, a priori, podría ser relevante. La lista de variables que presentaba la primera versión del Excel es:

- **Cromosoma.** Cromosoma en el que se encuentra el hotspot.
- **Localización cromosómica.** Ubicación citogenética en la que se encuentra el hotspot.
- **Gen.** Nombre del gen en el que se encuentra el hotspot.
- **Comienzo aa.** Posición proteica de inicio, en referencia al aminoácido correspondiente de la cadena.
- **Final aa.** Posición proteica de final, en referencia al aminoácido correspondiente de la cadena.
- **aa de comienzo de referencia.** Abreviatura del aminoácido de comienzo de la secuencia de referencia.
- **aa de comienzo alterado.** Abreviatura del aminoácido de comienzo de la secuencia alterada, si presenta.
- **aa de final de referencia.** Abreviatura del aminoácido de final de la secuencia de referencia.
- **SPDI canónico de inicio.** Definición de variantes acorde al modelo de datos SPDI[10] de inicio.
- **SPDI canónico de final.** Definición de variantes acorde al modelo de datos SPDI de final.
- **Secuencia de referencia GRCh37.** Campo que define la posición de referencia y los alelos respecto a la secuencia GRCh37.
- **Comienzo genómico GRCh37 de referencia.** Posición genómica de referencia de inicio del *assembly* 37.
- **Final genómico GRCh37 de referencia.** Posición genómica de referencia de final del *assembly* 37.

[10] **Modelo de datos SPDI:** definición de variantes como secuencia de cuatro atributos: (1)secuencia, (2) posición, (3) deleción y (4) inserción, aplicable a variantes de nucleótidos y proteínas.(*SPDI - NCBI Variation Notation for Variants with Known Breakpoints*, n.d.).

- **Secuencia de referencia GRCh38.** Campo que define la posición de referencia y los alelos respecto a la secuencia GRCh37.
- **Comienzo genómico GRCh38 de referencia.** Posición genómica de referencia de inicio del *assembly* 38.
- **Final genómico GRCh38 de referencia.** Posición genómica de referencia de final del *assembly* 38.
- **Transcrito de referencia.** Identificador único del transcrito de proteína codificante.
- **Inicio de transcrito.** Posición proteica de inicio del transcrito.
- **Final de transcrito.** Posición proteica final del transcrito.
- **Método o fuente.** Método de experimentación o fuente de extracción de la información
- **Fenotipo.** Fenotipo asociado al hotspot definido.
- **DOI.** Referencia bibliográfica.
- **Comentarios.** Comentarios sobre el documento.
- **ENSP.** Identificador proteico estándar.

Tras la extracción de toda la información disponible en los propios documentos, se pasó a la segunda etapa de desarrollo de la fuente de datos. En esta nueva versión se pretende ajustar las variables extraídas al modelo conceptual de hotspot realizado. Para poder conseguir este cometido se recurre a conceptos básicos de genética que permiten asociar información y se emplean herramientas auxiliares para completar la información restante.

6.3.2. Etapa 2: Filtrado de información obtenida usando el modelo conceptual

El objetivo principal de la segunda etapa para el desarrollo de la fuente de datos es completar la información restante. Durante la realización de esta segunda etapa, se presentan dos problemas principalmente. El primer problema, es la falta de información relevante para la definición correcta del hotspot; y el segundo problema es la falta de estandarización de la información presente en el fichero.

Para abordar el primer problema se recurre a herramientas auxiliares que pueden resultar útiles a la hora de completar esa información. En este caso, se emplean cuatro herramientas principalmente para conseguir diferente información: SynVar, RefSeq, Clinvar y UniprotKB. En la Tabla 2 se puede observar un pequeño resumen en el que se expone la información que falta y cuál de las herramientas escogida responde a cada uno.

Tabla 2: Resumen de la utilidad de las diferentes herramientas auxiliares

PROBLEMA	SOLUCIÓN	HERRAMIENTA
Falta de referencia hgvs.	Búsqueda de la referencia sabiendo el gen y la variación en concreto.	SynVar
Falta de posición genómica de exones.	Búsqueda de la secuencia completa para identificar las posiciones de los exones.	RefSeq
Falta de información relevante básica.	Búsqueda de herramienta con información general acerca de variaciones.	ClinVar
1. Falta de aa correspondiente a una posición	1. Búsqueda de la proteína que codifica e identificación de la posición proteica del hotspot para asociarla con un aminoácido.	UniProtKB

proteica. 2. Falta identificador único de secuencia.	2. Empleo de los ENSP.

Dado que cada herramienta auxiliar es capaz de completar un tipo de información diferente, este apartado se divide a su vez en 4 subapartados, uno por cada herramienta auxiliar, en los cuales se explican de forma detallada el modo de empleo de las herramientas auxiliares .

6.3.2.1. SynVar

SynVar es una herramienta que permite la generación y normalización de sinónimos de variantes genómicas que surge como respuesta a la heterogeneidad en la forma de representación de estas variantes(Zahn-Zabal et al., 2020). Estas diferencias surgen por tres motivos: (1) los posibles niveles de representación de variables (genómico, proteico y de transcrito), (2) la alta dependencia de la secuencia de referencia para su representación y (3) la falta de estandarización en la nomenclatura de las variaciones provenientes de la literatura.

El principal cometido de esta herramienta es la generación de sinónimos a partir de un SNP dado, su descripción a nivel genómico, a nivel de transcrito o a nivel proteico. El resultado de la herramienta consta de diferentes partes:

- **Sinónimo.** En esta salida se presentan sinónimos de los nombres de genes y proteínas asociados.
- **Hgvs.** En esta salida se presenta la descripción de la variante en un formato estándar, para utilizarlo como identificador único.
- **Variación sintáctica.** En esta salida se presentan las expresiones de las variantes como se encuentran en la literatura.

Debido a la funcionalidad de SynVar se ha escogido esta herramienta para resolver el problema de falta de <hgvs>, ya que ,gracias a ella, se puede referenciar de una forma estandarizada las posiciones genómicas de inicio y fin del hotspots. Para ello se siguió el siguiente proceso:

1- Identificación del gen de interés.
2- Construcción de la variante en alguna de las diferentes formas: proteica, transcrito, genómica, dbSNP o COSMIC.

Una vez enviada esa información la herramienta nos devuelve un XML con toda la información que queda expuesta anteriormente. Una vez se obtiene el XML se busca la etiqueta <hgvs> y se copia la referencia NC asociada. Las referencias NC se entienden como una forma de representar regiones genómicas completas. Gracias a ellas se puede saber la localización exacta de la variación sobre la secuencia de referencia completa del genoma.En la Figura 14 se puede observar una captura de la página de entrada de la herramienta y los resultados que esta devuelve.

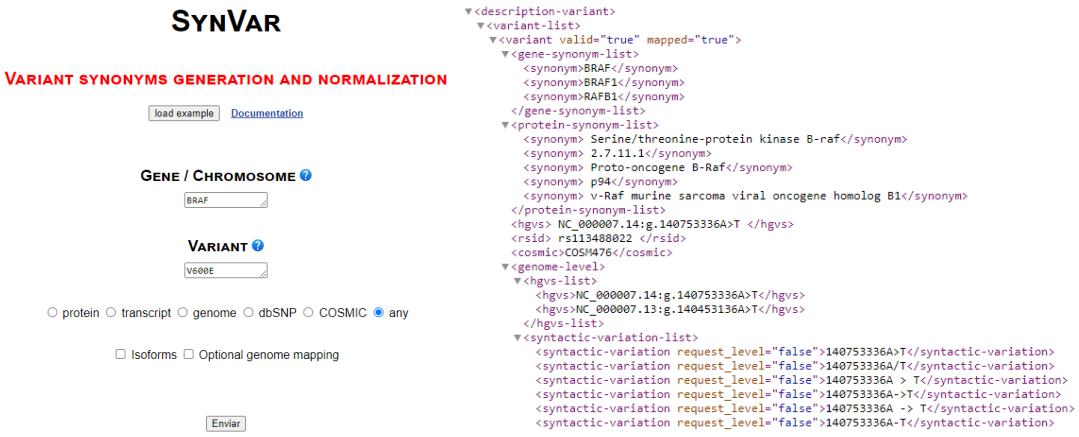


Figura 14: A la derecha la página de entrada de SynVar y a la izquierda la salida de SynVar ante la búsqueda de una variante genómica en un gen en concreto. [Fuente: SynVar]

6.3.2.2. RefSeq

RefSeq es una base de datos de estudios médicos, funcionales y de diversidad. Esto es debido a que la plataforma es capaz de almacenar secuencias completas, integradas, no redundantes y bien anotadas de ADN genómico, de transcritos y de proteínas (McEntyre & Ostell, 2002). Gracias a esta plataforma, se puede dar una referencia estable para la anotación del genoma y la identificación de genes y mutaciones.

Por todo lo expuesto anteriormente, en este Trabajo de Final de Grado se usa esta herramienta con el propósito de completar la información acerca de las posiciones genómicas concretas de los hotspots definidos por exones. El algoritmo empleado para encontrar esa información es:

1- Búsqueda del gen de interés en humanos en RefSeq mediante una regla específica con operadores lógicos.
2- Escoger la secuencia más larga con el *accession* NM_[11](Amberger et al., 2012).
3- Escoger la opción *MANE Ensembl Match* dado que estas secuencias cumplen unos estándares universales empleados en estudios clínicos y de genómica comparativa y evolutiva.
4- Contaje de exones hasta llegar al de interés.
    a. La posición de inicio del exón corresponde con el atributo Comienzo GRCh37.
    b. La posición de final del exón corresponde con el atributo final GRCh37.

En la Figura 14, Figura 15 y la Figura 16 se puede ver la página de RefSeq y todos los pasos seguidos para la extracción de la información de forma ordenada.

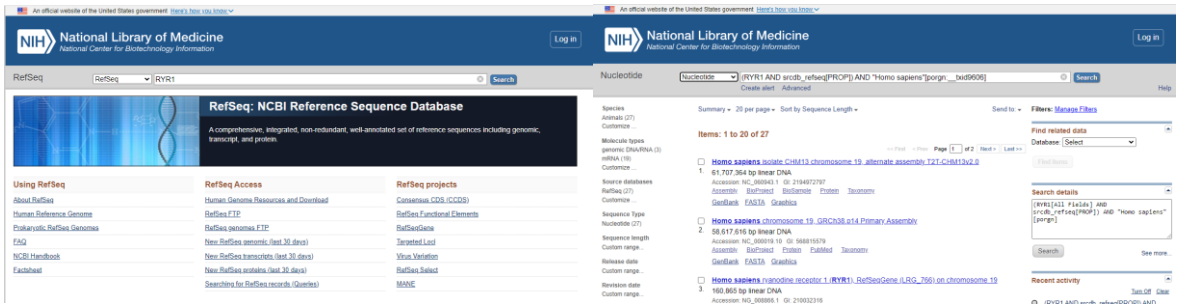
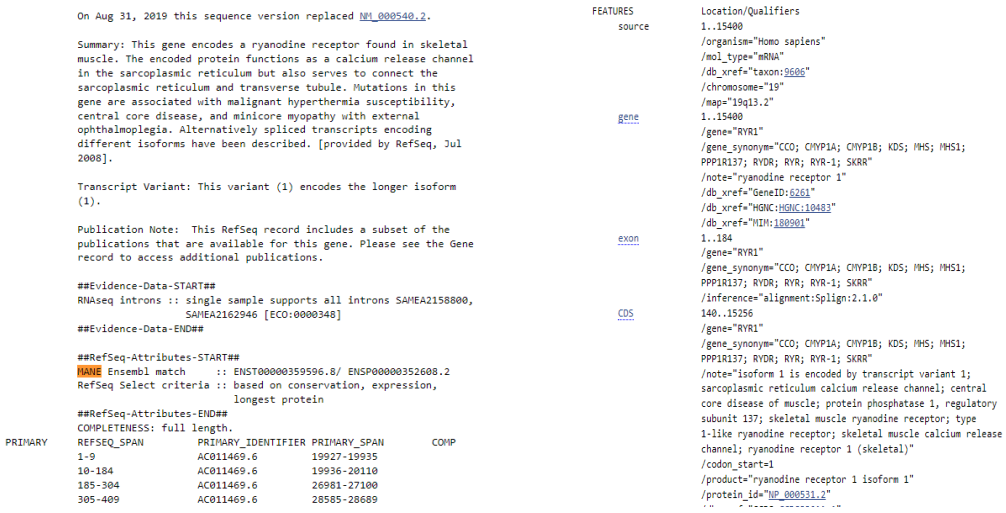


Figura 15: A la derecha la página principal de RefSeq y la izquierda los resultados de la búsqueda de un gen en concreto [Fuente: RefSeq]

[11] **NM_** : Es un tipo de prefijo para designar transcritos codificantes de proteínas.



On Aug 31, 2019 this sequence version replaced NM_000540.2.

Summary: This gene encodes a ryanodine receptor found in skeletal muscle. The encoded protein functions as a calcium release channel in the sarcoplasmic reticulum but also serves to connect the sarcoplasmic reticulum and transverse tubule. Mutations in this gene are associated with malignant hyperthermia susceptibility, central core disease, and minicore myopathy with external ophthalmoplegia. Alternatively spliced transcripts encoding different isoforms have been described. [provided by RefSeq, Jul 2008].

Transcript Variant: This variant (1) encodes the longer isoform (1).

Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications.

##Evidence-Data-START##
RNAseq introns :: single sample supports all introns SAMEA2158800, SAMEA2162946 [ECO:0000348]
##Evidence-Data-END##

##RefSeq-Attributes-START##
MANE Ensembl match :: ENST00000359596.8/ ENSP00000352608.2
RefSeq Select criteria :: based on conservation, expression, longest protein
##RefSeq-Attributes-END##
COMPLETENESS: full length.

PRIMARY	REFSEQ_SPAN	PRIMARY_IDENTIFIER	PRIMARY_SPAN	COMP
	1-9	AC011469.6	19927-19935	
	10-184	AC011469.6	19936-20110	
	185-304	AC011469.6	26981-27100	
	305-409	AC011469.6	28585-28689	

FEATURES	Location/Qualifiers
source	1..15400
	/organism="Homo sapiens"
	/mol_type="mRNA"
	/db_xref="taxon:9606"
	/chromosome="19"
	/map="19q13.2"
gene	1..15400
	/gene="RYR1"
	/gene_synonym="CCO; CMYP1A; CMYP1B; KDS; MHS; MHS1; PPP1R137; RYDR; RYR; RYR-1; SKRR"
	/note="ryanodine receptor 1"
	/db_xref="GeneID:6261"
	/db_xref="HGNC:HGNC:10483"
	/db_xref="MIM:180901"
exon	1..184
	/gene="RYR1"
	/gene_synonym="CCO; CMYP1A; CMYP1B; KDS; MHS; MHS1; PPP1R137; RYDR; RYR; RYR-1; SKRR"
	/inference="alignment:Splign:2.1.0"
CDS	140..15256
	/gene="RYR1"
	/gene_synonym="CCO; CMYP1A; CMYP1B; KDS; MHS; MHS1; PPP1R137; RYDR; RYR; RYR-1; SKRR"
	/note="isoform 1 is encoded by transcript variant 1; sarcoplasmic reticulum calcium release channel; central core disease of muscle; protein phosphatase 1, regulatory subunit 137; skeletal muscle ryanodine receptor; type 1-like ryanodine receptor; skeletal muscle calcium release channel; ryanodine receptor 1 (skeletal)"
	/codon_start=1
	/product="ryanodine receptor 1 isoform 1"
	/protein_id="NP_000531.2"

Figura 16: A la derecha podemos ver dónde encontrar el MANE Ensembl Match y a la izquierda donde encontrar la información relativa a los exones. [Fuente: RefSeq]

6.3.2.3. Clinvar

Tal y como queda expuesto en el apartado 4.1, Clinvar es un archivo público y de libre acceso de informes sobre las relaciones entre las variaciones y los fenotipos humanos con pruebas de apoyo. Por este motivo se escoge Clinvar como base de datos de referencia para completar los datos restantes. Sin embargo, y a diferencia de RefSeq o SynVar, el empleo de Clinvar no es de forma concreta sino más bien se emplea de una forma complementaria. Esta afirmación se refiere a que Clinvar se usó para completar la información que faltaba gracias al hecho de que esta base contiene información básica de variaciones genómicas . Es por ello por lo que, de esta herramienta se saca información sobre las posiciones proteicas, sobre los comienzos y finales de cada *assembly* etc.

6.3.2.4. UniProtKB

UniProt es un repositorio completo de datos sobre secuencias y anotaciones de proteínas. Este recurso se subdivide en tres grandes bases de datos: UniProtKB, UniRef y UniParc. Este recurso es el considerado de referencia a nivel proteico y por ello es este el que se emplea para sacar la información complementaria relacionada con la definición a nivel proteico de un hotspot. Asimismo, el hecho de que UniProtKB sea la base de datos principal y las otras dos las complementarias ha supuesto la focalización del empleo de este recurso a esta base de datos únicamente.

UniProtKB es una base de datos de proteínas comprensiva que consiste en dos secciones principales, UniProtKB/Swiss – Prot y UniProtKB/TrEMBL. La principal diferencia entre ambas secciones es la forma de añadir entradas a la base ya que en la primera lo hacen manualmente y en la segunda mediante computadores. Por un lado, UniProtKB/Swiss – Prot se caracteriza por contener una amplia notación manual, con redundancia mínima, integración con otras bases de datos y la asociación a documentación científica (Boutet et al., 2007). Por el otro lado, UniProt/TrEMBL se caracteriza principalmente por constar de traducciones de todos los CDS propuestos por EMBL/GenBank/ DNADatabank of Japan no integradas por la sección anterior (Leinonen et al., 2006) .

Tal y como queda descrito anteriormente, UniProtKB es capaz de encontrar toda la información relativa a la caracterización a nivel proteico de las proteínas de interés. Es por ello por lo que se ha empleado esta herramienta para encontrar tres tipos de información: (1) los aminoácidos de

referencia (*start_ref* y *end_ref* en la clase *PositionProt*), (2) ENSP (*sequence* en la clase *ProteinSequence*) y (3) UniprotID (*name* en la clase *Protein*).

En primer lugar, la búsqueda de información acerca de los aminoácidos de referencia restantes, los cuales corresponden con las variables *start_ref* y *end_ref* en el modelo conceptual, se basó en la estrategia de búsqueda expuesta en el algoritmo a continuación:

1- Búsqueda del gen de interés en la barra de buscadores.
2- Filtrado de la búsqueda en *Status* a *Reviewed* (Swiss - Prot) y en *Taxonomy* a *Human*.
3- Ir a la sección *Sequence and Isoforms.*
4- Búsqueda de la posición proteica de interés.
5- Selección del aminoácido correspondiente en forma abreviada.

En la Figura 17, la Figura 18, la Figura 19 y la Figura 20 se puede observar paso a paso la forma de proceder para encontrar el aminoácido de referencia restante.

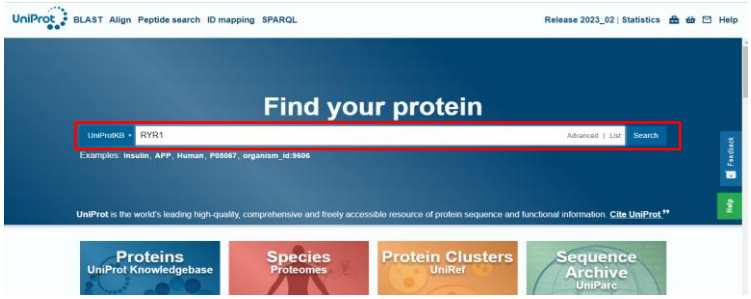


Figura 17: Página de entrada de UniProtKB. [Fuente: UniProtKB]

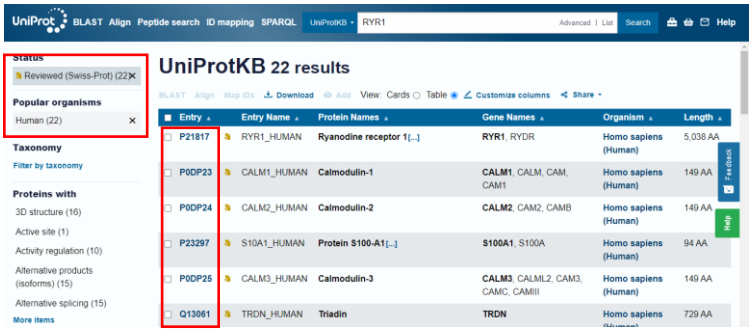


Figura 18: Resultado de búsqueda de un gen con los filtros aplicados. [Fuente: UniProtKB]

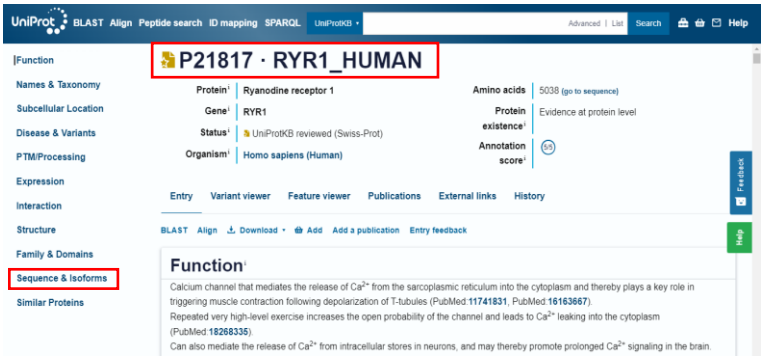


Figura 19: Página principal tras seleccionar una de las entradas de la página principal de resultado de la búsqueda . [Fuente: UniProtKB]

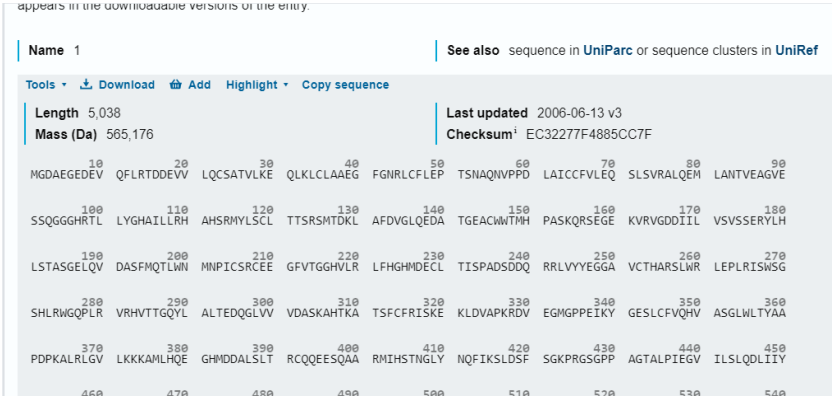


Figura 20:Subapartado de la página principal de la proteína estudiada (Sequence & Isoforms).[Fuente: UniProtKB]

En segundo lugar, esta herramienta también sirve para la búsqueda de los identificadores ENSP, los cuales se presentan en el modelo como la variable *sequence* dentro de la clase *ProteinSequence*. El algoritmo de búsqueda de estos identificadores es el siguiente:

1- Búsqueda del gen de interés en la barra de buscadores (Figura 17 izquierda).
2- Filtrado de la búsqueda en *Status* a *Reviewed* (Swiss - Prot) y en *Taxonomy* a *Human* (Figura 17 derecha).
3- Ir a la sección *Sequence and Isoforms* (Figura 19).
4- Búsqueda del subapartado llamado *Genome annotation databases* (Figura 21).
5- Selección del ENSP asociado a esta proteína.

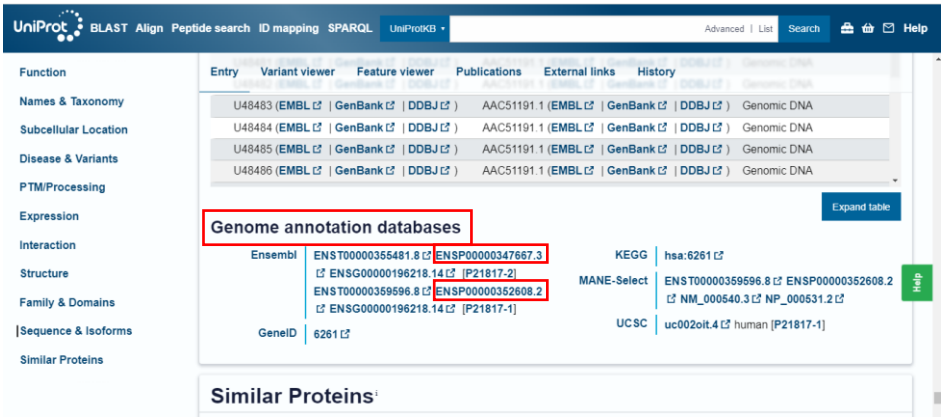


Figura 21: Vista del apartado de Genome annotation database de UniProtKB. [Fuente: UniProtKB]

En tercer lugar, UniProtKB también se emplea para obtener la variable *name,* correspondiente al UniprotID, dentro de la clase *Protein* del modelo conceptual. El motivo principal de añadir esta variable al modelo reside en la necesidad de identificar de una forma estandarizada la proteína que se codifica en cada caso. Para conseguir este atributo se usa la página web UniProtKB y se busca la proteína que se codifica en cada caso. En la Figura 19 puede observarse en la parte izquierda del nombre del gen, el identificador proteico.

Una vez se tiene recopilada toda la información , se pasa a la optimización de esta fuente de datos mediante el filtrado de todas las variables usando el modelo conceptual. Es por ello por lo que pasamos de tener 27 variables a tan solo 15, las cuales se corresponden con las definidas en el modelo conceptual. Una vez se consigue extraer toda la información se pasa a la última etapa de esta parte de la extracción de la información, la estandarización de la información.

6.3.3. Etapa 3: Estandarización de la información

Esta última etapa de la parte de extracción de la información consiste en la estandarización de la información recogida. La estandarización de la información en este caso consiste en el ajuste y la adaptación de las características de ciertas variables del modelo para garantizar una representación común con el objetivo de mejorar la coherencia interna y futura tratabilidad de los datos.

El primer paso para la estandarización de la información consiste en la obtención de la posición genómica de inicio (*start* en la clase *PositionChr*) y fin (*end* en la clase *PositionChr*) de todos los hotspots identificados. De esta forma se consigue estandarizar la forma en la que se presenta la información relativa a la posición genómica de estos. Para conseguir este cometido se recurre a los conocimientos genéticos adquiridos y a la plataforma RStudio para generar un código de R que semi automatice el proceso. El código empleado para la estandarización de la posición genómica del hotspot de interés puede consultarse en el Anexo 2.

El segundo paso para la estandarización se basa en la búsqueda de un marco común para determinar distintas variables en base a ontologías. El motivo principal del empleo de las ontologías reside en la necesidad de tener una información coherente, unificada y estandarizada. Este hecho resulta fundamental para conseguir cubrir todas las necesidades descritas en el apartado 5.2. Las variables del modelo escogidas para estandarizar mediante ontologías son el fenotipo asociado (*name* dentro de la clase *Phenotype*) y el método o fuente (*obtention_method* dentro de clase *Hotspot*).

Las ontologías empleadas para estandarizar la variable fenotipo son MONDO y HP. MONDO *Disease Ontology* es una ontología construida de forma semiautomática que fusiona múltiples recursos de enfermedades(Vasilevsky et al., n.d.). HP o *Human Phenotype Ontology* es una ontología que proporciona vocabulario de anormalidades fenotípicas y características clínicas encontradas en enfermedades humanas basándose en literatura médica sacada de Orphanet (https://www.orpha.net/), DECIPHER (https://www.deciphergenomics.org/) y OMIM(https://www.omim.org/) (*Human Phenotype Ontology*, n.d.). Por este motivo se escogieron MONDO y HP como las ontologías de referencia para la estandarización de los fenotipos. La estrategia de búsqueda de los identificadores ontológicos se basó en poner el fenotipo descrito en el artículo científico en la barra del buscador de la ontología para así, encontrar el identificador más adecuado.

Asimismo, las ontologías empleadas para la estandarización de la variable método o fuente son ECO y OBI. ECO o *Evidence and Conclusion Ontology*, describe los tipos de pruebas y métodos de aserción en proceso de biocuración para ratificar pruebas biológicas para poder, posteriormente, rastrear la procedencia de las anotaciones, establecer medidas de control de calidad y consultar pruebas (Giglio et al., 2019). Por otro lado, la ontología OBI o *Ontology for Biomedical Investigation* define investigaciones biomédicas, diseños de estudio, protocolos e instrumentación empleada, así como los datos generados y los tipos de análisis realizados con los datos (Bandrowski et al., 2016).

En la Figura 13 se puede observar el diagrama de flujo asociado a esta etapa del método SILE. En él, se puede observar cómo se integran los pasos realizados para conseguir las diferentes versiones de la fuente de datos, hasta obtener la extracción completa de toda la información con la correspondiente estandarización de esta.

6.4. WEB (L Y E)

La tercera etapa del método SILE, tiene como principal objetivo convertir la información extraída en un formato consultable que permita su posterior gestión y explotación. En el caso de estudio del trabajo, esta etapa consiste en la propia creación de la fuente de datos de hotspots de cardiopatías, dado que ya se han definido los parámetros que se van a utilizar y se ha seleccionado las fuentes de información de las cuales se va a extraer el conocimiento.

Es importante tener en cuenta cuales van a ser los usos de esta fuente de datos ya que, según sus aplicaciones, la forma de almacenar la información encontrada puede variar. En este caso en concreto, es crucial tener en mente la futura automatización de esta fuente de datos. Por tanto, resulta imprescindible presentar la información de forma que, en un futuro, un algoritmo de inteligencia artificial o IA sea capaz de extraer la información para múltiples aplicaciones tales como la interpretación de variaciones. Por todo lo expuesto anteriormente, el formato utilizado para la carga de estos datos es el .xslx gracias a que este nos permite una gran versatilidad y nos permite extraer, importar y cargar la información de forma sencilla.

La cuarta y última etapa del método SILE consiste en explotar el conocimiento de la información almacenada en la fuente de datos creada. En esta se deben definir mecanismos de consulta de la información, estadísticas y análisis que aporten conocimiento a nivel clínico de una forma rápida y sencilla. Para conseguir este cometido, en este Trabajo de Fin de Grado se ha colaborado activamente en el diseño de una página web en la que se consigue integrar toda la información extraída y se presenta de una forma amigable y sencilla para los beneficiarios de este recurso.

La página web tiene como objetivo principal proporcionar toda la información almacenada en la fuente de datos de una forma clara, inteligible y ordenada para que el usuario sea capaz de entenderla de manera intuitiva. Es por ello por lo que, el diseño de la página web no es trivial, ya que este aboga por la sencillez para evitar posibles confusiones por parte del usuario. Esta web contendrá de una forma más visual toda la información que se ha extraído haciendo uso de figuras y esquemas interactivos.

Tras la implementación de todos los pasos del método SILE, se consigue cumplir con todos los objetivos asociados a esta etapa del ciclo regulativo del diseño propuestos en el trabajo. En la Figura 22 se puede observar el diagrama de flujo asociado a todo el proceso de diseño y desarrollo de la fuente de datos mediante la aplicación del método SILE.

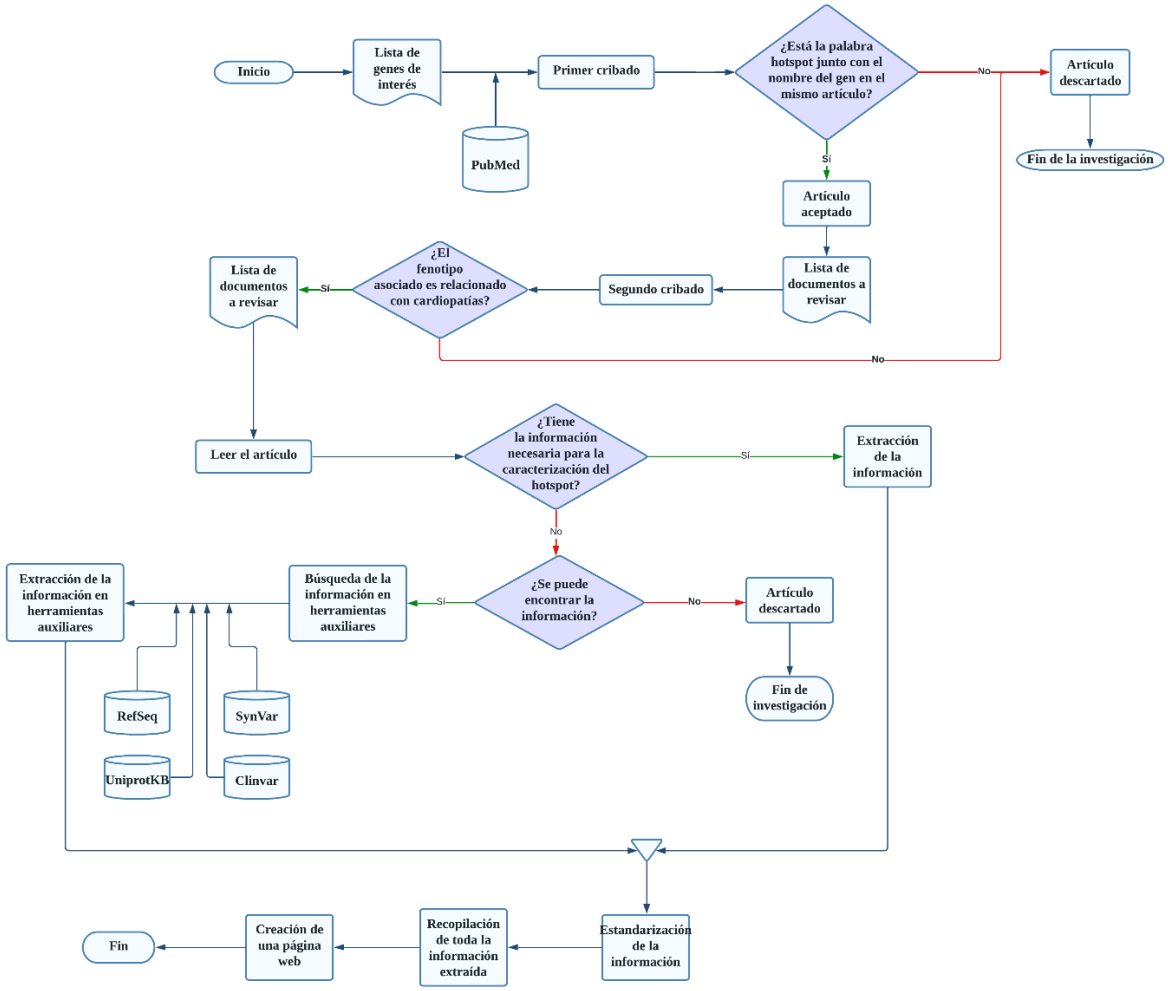


Figura 22: Diagrama de flujo integrado de la fase de diseño y desarrollo de la fuente de datos aplicando el método SILE. [Elaboración propia]

CAPÍTULO 7. RESULTADOS

En este capítulo se van a exponer los resultados del proceso de diseño y desarrollo de la fuente de datos para conocer la información que hay disponible acerca de los hotspots de cardiopatías. A lo largo de todo el capítulo se hace un análisis sobre los resultados obtenidos en cada parte del método SILE, así como la fuente de datos obtenida finalmente. Asimismo, se exponen los resultados del caso de uso de la fuente de datos con datos de pacientes reales de cardiología del Hospital La Fe de Valencia.

7.1. ANÁLISIS DE LOS RESULTADOS

Una vez diseñada toda la estrategia, se desarrolló la fuente de datos de hotspots y se tomó como referencia la lista de genes de interés proporcionada por el hospital La Fe de Valencia (Anexo 1). Para analizar los resultados de una forma estructurada, este subapartado se divide a su vez en tres secciones en las que se va a centrar este análisis: (1) Resultados de la etapa S (apartado 6.2), (2) Resultados de la etapa I (apartado 6.3) y (3) Resultado de las etapas L y E (apartado 6.4).

7.1.1. Resultados de la etapa S

Los resultados obtenidos del primer cribado, el cual tiene el objetivo de encontrar información relativa a hotspots de cardiopatías a través del empleo de la Ecuación 1 en PubMed, se resumen en la Figura 23.

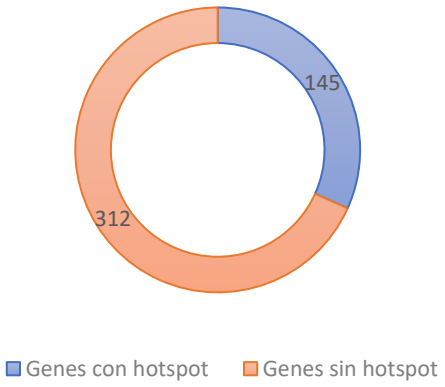


Figura 23: Diagrama circular en el que se muestra el resultado del primer cribado de genes candidatos. [Elaboración propia]

En el diagrama circular superior se puede observar cómo solo el 32% de los genes candidatos contienen alguna información acerca de hotspots dentro de cualquier ámbito clínico. Este porcentaje representa un total de 145 genes de los 457 genes estudiados inicialmente. Este resultado supone un descarte de más de la mitad de los genes iniciales, dado que no se ha encontrado ningún artículo relacionado, y acota aún más el caso de estudio. Este hecho suscita dos hipótesis acerca de los resultados; la primera hipótesis es que no hay mucha información documentada acerca de la existencia de hotspots, y la segunda es que no hay muchos genes de interés cardiológico que presenten hotspots.

Una vez descartados los genes que no contenían información acerca de hotspots en ningún artículo , se pasan los artículos restantes asociados a los posibles genes con hotspots a la siguiente fase. Esta se corresponde con el segundo paso del diagrama de flujo de la Figura 12, es decir, con la lectura de los resúmenes de cada artículo para determinar si estos hablan de hotspots relacionados con cardiopatías o no. Los resultados del segundo cribado se resumen en la Figura 24. Se puede observar cómo solo

quedaron 85 genes de los 145 que habían pasado el primer cribado. A nivel porcentual supone que solo el 41% de los genes con hotspots están relacionados con las cardiopatías.

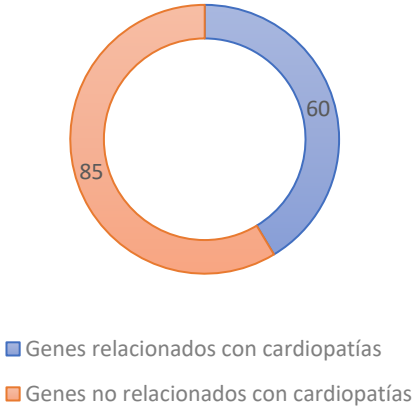


Figura 24: Diagrama circular resultante del segundo cribado de genes candidatos. [Elaboración propia]

Tras este segundo cribado se puede dar por finalizada la primera etapa del método SILE, ya que ya se cuenta con los artículos de interés que, sirven para extraer la información relevante para la caracterización de hotspots. En la Figura 25, se puede ver un resumen de todo el proceso de búsqueda de información con los resultados obtenidos en este caso de estudio.

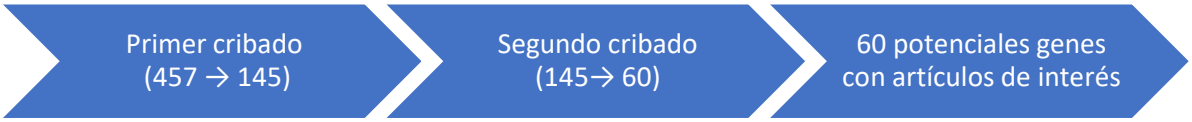


Figura 25: Resumen de los resultados obtenidos de la etapa de obtención de la información. [Elaboración propia]

7.1.2. Resultados de la etapa I

Tras el diseño de la fuente de datos se pasó al análisis de los resultados extraídos con la intencionalidad de encontrar información relevante. El primer paso realizado fue la aceptación o descarte de los documentos científicos extraídos de la etapa anterior. Tal y como queda explicado en el subapartado 6.3, el primer proceso realizado en esta etapa es el de lectura del documento para ver si este contiene toda la información o, por lo menos, parte de ella. Los resultados de este proceso quedan expuestos en la Figura 26.

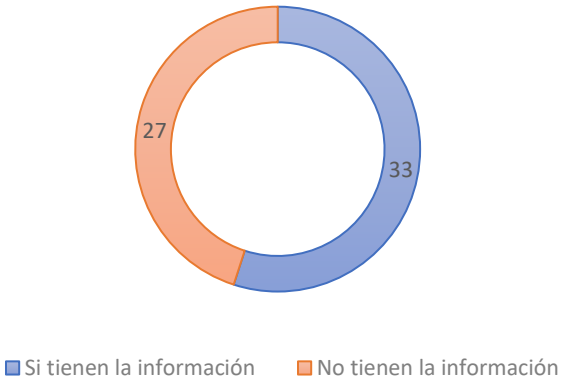


Figura 26: Genes diferentes con información o no en la literatura sobre hotspots. [Elaboración propia]

Para conseguir los resultados expuestos en la Figura 26 se recurrió a un Excel de control en el que se enumeraban todos los artículos a revisar. En este documento auxiliar se planteaban posibles motivos de aceptación o rechazo de un artículo científico tales como: (1) eliminado por no contener la información relativa al modelo, (2) paper repetido, (3) contiene información acerca de hotspots de recombinación y no mutacionales, (4) no se tiene acceso al paper y (5) documento aceptado. Los resultados de la clasificación de estos documentos en el Excel de control se exponen en la Figura 27 .

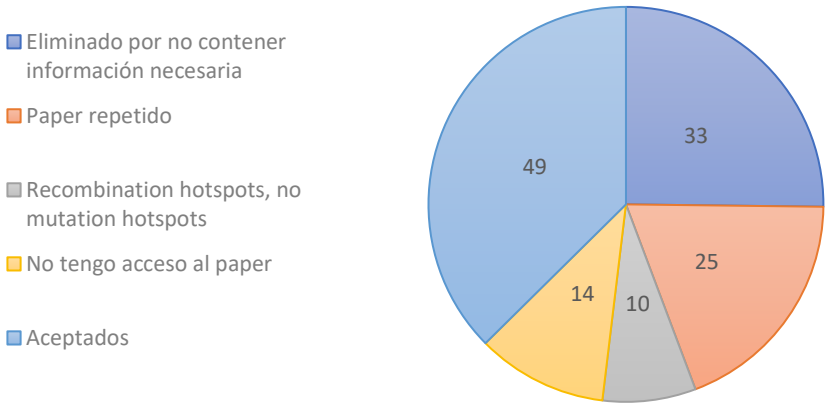


Figura 27: Diagrama circular que muestra los motivos de aceptación o rechazo de un documento científico. [Elaboración propia]

Los resultados que se observan en el diagrama superior son reveladores, dado que se puede observar como a penas el 40 % de los documentos se aceptaron, dejando patente así la falta de información existente en la documentación científica a la hora de caracterizar un hotspot. Asimismo, es importante comentar, que un 25% de toda la documentación extraída no contenía la información necesaria para caracterizar un hotspot, siendo este un motivo más por el que se ratifica la problemática acerca de los hotspot.

También existen otras casuísticas como el hecho de tener el paper repetido motivado por contener información acerca de diferentes genes en un mismo documento, siendo este caso el 19% de los documentos estudiados. De igual manera, también se rechazaron acerca del 8% de los artículos por contener información acerca de hotspots de recombinación y no sobre hotspots mutacionales. Cabe añadir que el 11% de los documentos encontrados no pudieron ser estudiados por no tener acceso a dichos documentos.

Una vez terminado el proceso de lectura de los documentos se sigue con el proceso de extracción de la información de forma directa e indirecta – a través de las herramientas auxiliares – según lo expuesto en el apartado 6.3.2. El empleo de estas herramientas resultó exitoso ya que se emplearon todas ellas para completar información de distintos hotspots que quedaban incompletos.

RefSeq se empleó para caracterizar las posiciones genómicas de 27 hotspots que venían definidos mediante exones. Estos 27 hotspots se enmarcaban en 9 genes en concreto: RYR1, JUP, MYH7, KCNH2, COG8, TTN, ACTC1, NKX2-5 y MYBPC3.

SynVar fue la segunda herramienta más empleada ya que, contribuyó en la extracción de información relativa al hgvs de 23 de los hotspots asociados a 12 genes distintos como son el CACNA1C, DNAJB6, GNE, KCNH2, TBX5, MYBPC3, RYR2, RYR1, ABCC6, MYH7, GATA4 y KCNJ8.

Asimismo, el empleo de UniProtKB en relación con la determinación de los aminoácidos de referencia fue también bastante extendida ya que, la mayoría de los artículos mencionaban la posición proteica

pero no el aminoácido al que correspondía en cada caso. Más concretamente, se buscó el aminoácido de referencia en 20 de los hotspots definidos dentro de la fuente de datos, repartidos estos en 8 genes en concreto como son el APOA1, DSP, RYR1, KCNQ1, RYR2, SCN5A, RBM20 y KCNH2.Finalmente, el empleo de Clinvar fue muy habitual debido a que la mayoría de los hotspots definidos tenían alguna variable del modelo sin completar y Clinvar, fue la herramienta encargada de rellenar ese vacío.

Una vez extraídos los datos, se realizó un análisis para obtener una visión global de los hotspots obtenidos. En concreto, se ha analizado la distribución de los hotspots en base a tres de las variables más importantes de nuestra fuente de datos: el gen, el cromosoma y el fenotipo.

El resultado del primer caso de estudio puede observarse en la Figura 28, siendo esta la distribución de los hotspots extraídos en función del gen en el que se presentan. Con esta figura podemos observar como la mayor concentración de hotspots identificados se asocian al gen RYR1, acumulando este un total de 27 hotspots. Tras este gen le siguen TTN, con un total de 10 hotspots, RYR2 con 9 hotspots y KCNH2 y TNNT2 con 6 hotspots cada uno. Estos resultados son evidentes por el hecho de ser los genes más comunes estudiados dentro del ámbito de cardiología. El resto de los genes estudiados presentan pocos hotspots asociados, lo que indica que no existe demasiada información en la literatura de estos genes dentro del ámbito de las cardiopatías. Estos resultados invitan a la reflexión y al estudio profundo de estos genes con alta tasa de hotspots dentro de este caso de estudio ya que, probablemente contengan más información relevante que pueda ser reveladora a la hora del entendimiento y diagnóstico de cardiopatías hereditarias.

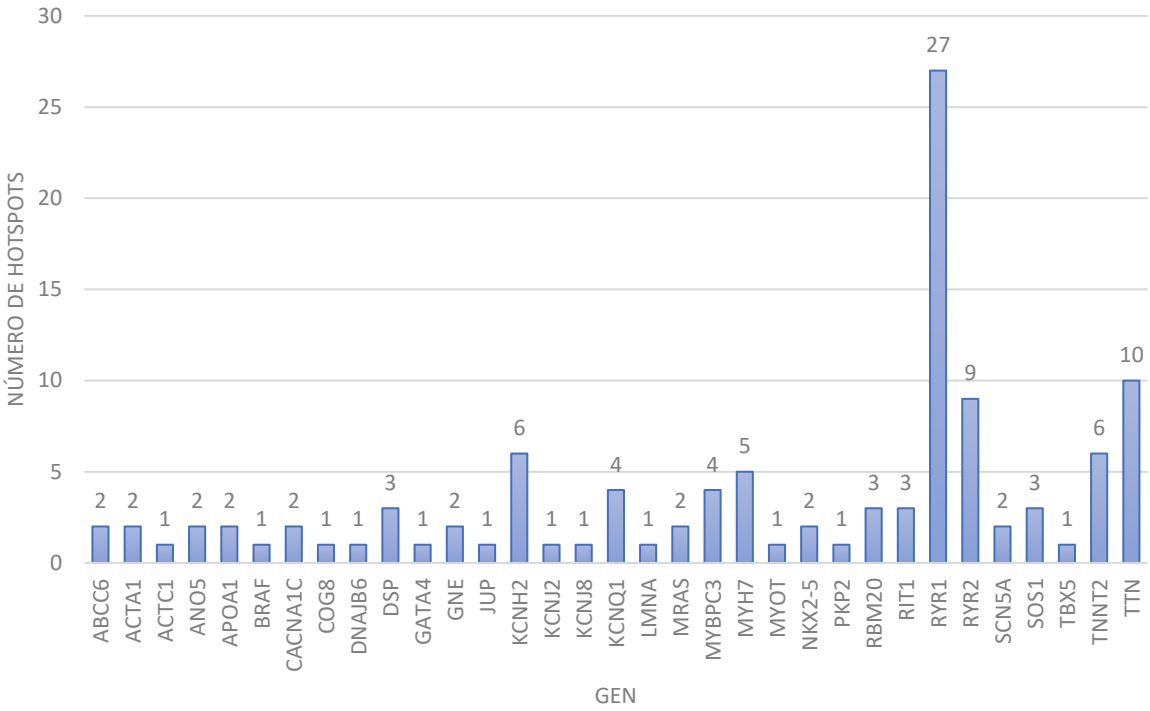


Figura 28: Distribución de hotspots en función del gen de interés mediante un diagrama de barras. [Elaboración propia]

El segundo caso de estudio es el de la distribución de la existencia de hotspots en función del cromosoma en el que se encuentre. Este análisis puede revelar que cromosomas son más susceptibles a la presencia de hotspots y, por tanto, puede suponer el inicio de una futura investigación acerca de estos cromosomas. En la Figura 29 se puede observar dicha distribución.

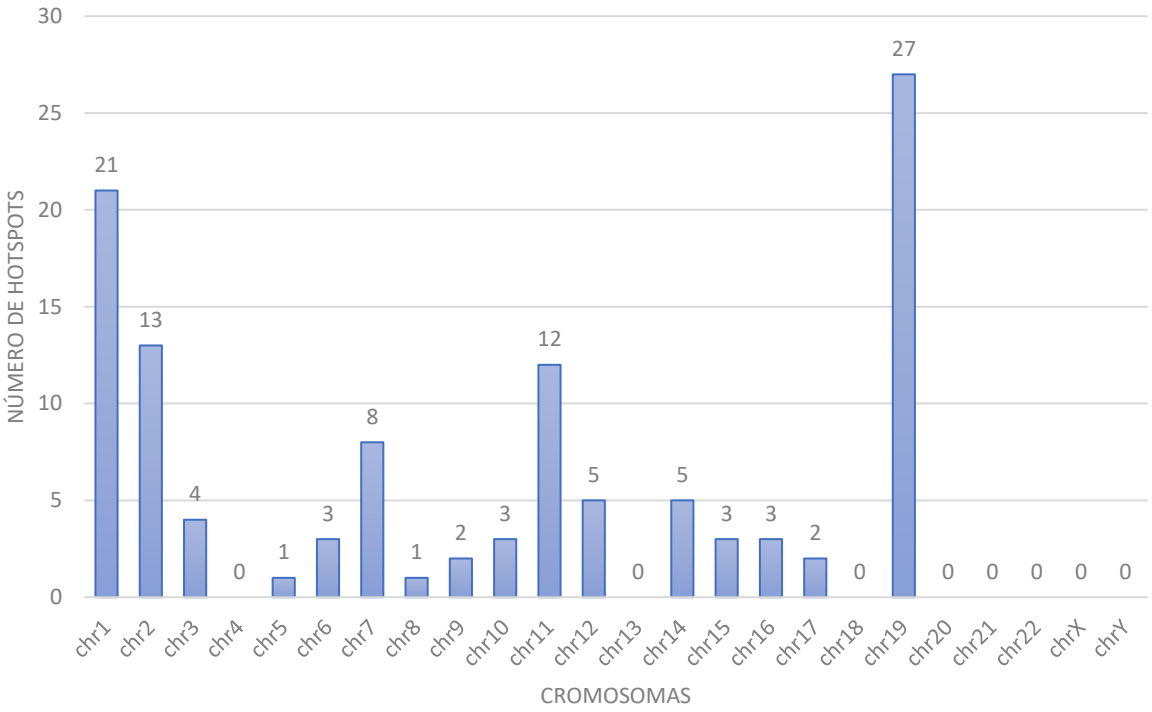


Figura 29: Distribución de hotspots en función de cromosomas mediante un diagrama de barras. [Elaboración propia]

Analizando los resultados de la distribución anterior se puede decir que el cromosoma 19 es el más susceptible a presentar hotspots en su estructura. Esta afirmación concuerda con el diagrama de la Figura 28 ya que, este afirma que el gen RYR1 es el que más hotspots presenta y este se sitúa en el gen 19 precisamente. Tras este cromosoma le sigue el cromosoma 1, con un total de 21 hotspots en su estructura. Este se corresponde con los genes RYR2 (con 9 hotspots), TNNT2 (con 6 hotspots), RIT1 (con 3 hotspots) y ACTA1 (con 2 hotspots). Es importante resaltar también que el cromosoma 2 cuenta con un total de 13 hotspots en su estructura, siendo 10 de ellos pertinentes al gen TTN. Estos resultados reflejan indicios de que cromosomas son los más vulnerables a la presencia de hotspots y, por tanto, son más susceptibles de generar cardiopatías familiares.

El tercer caso, es el de la distribución de hotspots en función del fenotipo que tienen asociado. En la Figura 30 se puede observar el resultado de dicho análisis.

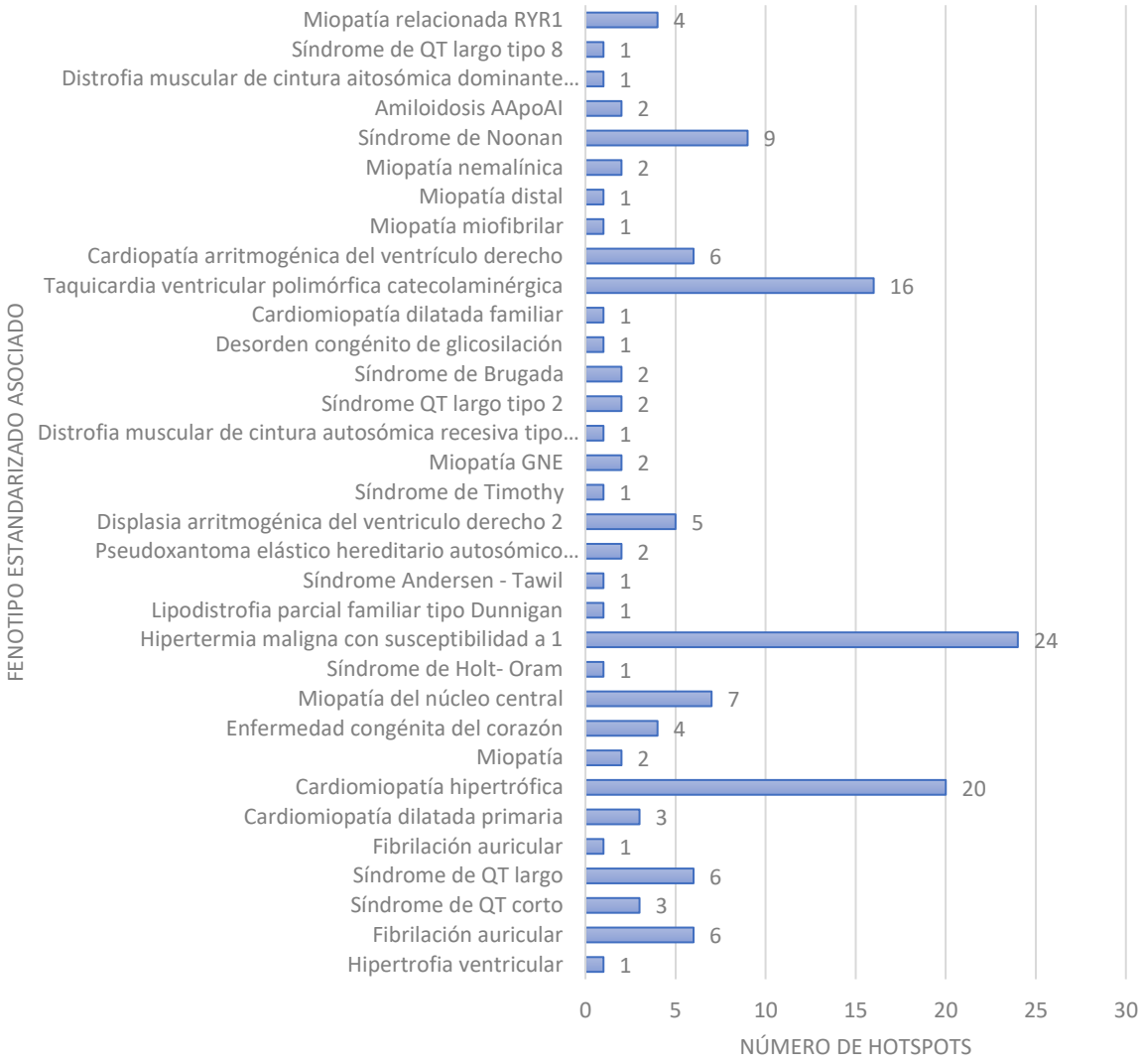


Figura 30:Distribución de hotspots en función del fenotipo asociado. [Elaboración propia]

Tras conocer estas asociaciones se puede pasar al análisis de estos resultados. Según la Figura 30, se puede ver como el fenotipo que más hotspots tiene identificados en la bibliografía es el de la Hipertermia maligna de susceptibilidad a 1 (MONDO:0007783), con un total de 24 hotspots. Esta enfermedad no había sido descrita en el apartado 1.2 debido a que su incidencia es muy baja (1:16000 – 1:250000 ratio de incidencia) debido a que se presenta principalmente como respuesta hipermetabólica a anestésicos volátiles y relajantes musculares despolarizables en individuos susceptibles por presentar mutaciones (Janet et al., 2023) . Es por ello por lo que la mayoría de los casos de Hipertermia Maligna se conoce la causa y son reportados. Esta enfermedad presenta las mutaciones genéticas en tres genes, RYR1,CACNA1S y STAC3, siendo el primero de ellos uno de los más frecuentes de nuestra base de datos.

Al fenotipo más común le sigue la Miocardiopatía Hipertrófica o MH (MONDO:0005045) sumando este 20 hotspots y la Miocardiopatía arritmogénica del ventrículo derecho (MONDO:0016587) con un total de 16 hotspots. Asimismo, tenemos dos fenotipos bastante comunes como el Síndrome de Noonan (MONDO:0018997) y la Miocardiopatía de núcleo central (MONDO:007294), con 9 y 7 hotspots asociados respectivamente.

7.1.3. Resultados de las etapas L y E

Una vez se termina la etapa de identificación del método SILE, se pasa las etapas L y E. En la etapa de carga, se obtuvo finalmente un documento Excel (.xslx) en el que se encuentra toda la información recopilada y dispuesta según el modelo conceptual elaborado de hotspot. El resultado de esta fuente de datos queda expuesto en el Anexo 3. Esta fuente se presenta de forma clara, unificada y estandarizada y su aspecto es el que puede observarse en la Figura 31 y la Figura 32. En este documento se recoge información acerca de 111 hotspots diferentes extraídos de la literatura dispuestos en un formato de tabla de 15 columnas y 111 filas, una para cada hotspot.

CHROMOSOME	CHROMOSOMAL LO	GENE	Uniprot I	START AA	END AA	REF START AA	REF END AA	UNIPROTKB'S ENSP	START GRCh37
chr1	1q22	LMNA	P02545	482	482	Arg	Arg	ENSP00000355292	156106775
chr1	1q21.2	RIT1	Q92963	57	57	Ala	Ala	ENSP00000357306	155874588
chr1	1q21.2	RIT1	Q92963	82	82	Phe	Phe	ENSP00000357306	155874285
chr1	1q21.2	RIT1	Q92963	95	95	Gly	Gly	ENSP00000357306	155874246
chr1	1q43	RYR2	Q92736	3778	4201	Leu	Gln	ENSP00000355533	237923082
chr1	1q43	RYR2	Q92736	44	466	Asn	Pro	ENSP00000355533	237433878
chr1	1q43	RYR2	Q92736	77	466	Leu	Pro	ENSP00000355533	237494238
chr1	1q43	RYR2	Q92736	2246	2534	Ser	Leu	ENSP00000355533	237798236
chr1	1q43	RYR2	Q92736	3778	4201	Leu	Gln	ENSP00000355533	237923082
chr1	1q43	RYR2	Q92736	4497	4959	Arg	Arg	ENSP00000355533	237954741
chr1	1q43	RYR2	Q92736	164	164	Pro	Pro	ENSP00000355533	237540649
chr1	1q43	RYR2	Q92736	169	169	Arg	Arg	ENSP00000355533	237540664
chr1	1q43	RYR2	Q92736	176	176	Arg	Arg	ENSP00000355533	237540685
chr1	1q32	TNNT2	P45379	92	92	Arg	Arg	ENSP00000236918	201334741
chr1	1q32	TNNT2	P45379	79	79	Ile	Ile	ENSP00000236918	201334780
chr1	1q32	TNNT2	P45379	110	110	Phe	Phe	ENSP00000236918	201334385
chr1	1q32	TNNT2	P45379	130	130	Arg	Arg	ENSP00000236918	201334325
chr1	1q32	TNNT2	P45379	278	278	Arg	Arg	ENSP00000236918	201328753
chr1	1q32	TNNT2	P45379	286	286	Arg	Arg	ENSP00000236918	201328362
chr10	10q25.3	RBM20	Q5T481	634	634	Arg	Arg	ENSP00000358532	112572055
chr10	10q25.3	RBM20	Q5T481	638	638	Arg	Arg	ENSP00000358532	112572067
chr10	10q25.3	RBM20	Q5T481	628	655	Tyr	Cys	ENSP00000358532	112572037
chr11	11p15.1	ANO5	Q75V66	61	95	Phe	Glu	ENSP00000315371	22242643
chr11	11p15.1	ANO5	Q75V66	746	805	Ala	Arg	ENSP00000315371	22296115
chr11	11q23.3	APOA1	P02647	50	93	Trp	Gln	ENSP00000236850	116707717
chr11	11q23.3	APOA1	P02647	170	178	Leu	Leu	ENSP00000236850	116706794
chr11	11p15.5	KCNQ1	P51787	341	341	Pro	Pro	ENSP00000155840	2604764
chr11	11p15.5	KCNQ1	P51787	344	344	Thr	Thr	ENSP00000155840	2604773
chr11	11p15.5	KCNQ1	P51787	160	202	Glu	Asp	ENSP00000155840	2591858
chr11	11p15.5	KCNQ1	P51787	261	307	Glu	Val	ENSP00000155840	2594076
chr11	11p11.2	MYBPC3	Q14896	847	847	Tyr	Tyr	ENSP00000442795	47359003
chr11	11p11.2	MYBPC3	Q14896	258	274	Glu	Thr	ENSP00000442795	47369975

Figura 31: Fuente de datos de hotspots de cardiopatías familiares (I) . [Elaboración propia]

END GRCh37	STANDARDIZED METHOD/SOURCE	STANDARDIZED PHENOTYPE	DOI	COMMENTS
156106777	author statement without traceable s	familial partial lipodystrophy Dunnigan t	DOI:10.1016/j.diabet.2018.09.006	
155874590	polymerase chain reaction assay (OBI_	Noonan syndrome (MONDO:0018997)	DOI:10.1038/gim.2016.32	
155874287	polymerase chain reaction assay (OBI_	Noonan syndrome (MONDO:0018997)	DOI:10.1038/gim.2016.32	
155874248	polymerase chain reaction assay (OBI_	Noonan syndrome (MONDO:0018997)	DOI:10.1038/gim.2016.32	
237947615	author statement without traceable s	catecholaminergic polymorphic ventricul	DOI:10.1074/jbc.M116.75652	The source states that the functional impa
237617796	imported information (ECO:0000311)	catecholaminergic polymorphic ventricul	DOI:10.1016/j.hrthm.2021.07	Test in catecholaminergic polymorphic ve
237617796	manual assertion (ECO:0000218)|aut	catecholaminergic polymorphic ventricul	DOI: 10.1161/CIRCRESAHA.11	| The hotspot is located within the N-terr
237813266	imported information (ECO:0000311)	catecholaminergic polymorphic ventricul	DOI:10.1016/j.hrthm.2021.07	Test in catecholaminergic polymorphic ve
237947615	imported information (ECO:0000311)	catecholaminergic polymorphic ventricul	DOI:10.1016/j.hrthm.2021.07	Test in catecholaminergic polymorphic ve
237995920	imported information (ECO:0000311)	catecholaminergic polymorphic ventricul	DOI:10.1016/j.hrthm.2021.07	Test in catecholaminergic polymorphic ve
237540651	polymerase chain reaction (OBI_0000	arrhythmogenic right ventricular cardiom	DOI:10.1016/j.jmb.2013.08.01	The hotspot is P164S in the HS-LOOP of th
237540666	polymerase chain reaction (OBI_0000	arrhythmogenic right ventricular cardiom	DOI:10.1016/j.jmb.2013.08.01	The hotspot is R169Q in the HS-LOOP of tl
237540687	polymerase chain reaction (OBI_0000	arrhythmogenic right ventricular cardiom	DOI:10.1016/j.jmb.2013.08.01	The hotspot is R176Q in the HS-LOOP of t
201334743	RNA sequencing assay (OBI:0001177),	Ventricular hypertrophy (HP:0001714)|A	DOI:10.1016/j.yexcr.2019.111736|DOI:10.3389/fphys.2022.864547	
201334782	DNA sequencing assay (OBI_0000626)	Atrial fibrillation (HP:0005110), Hypertro	DOI:10.3389/fphys.2022.864547	
201334387	DNA sequencing assay (OBI_0000626)	Atrial fibrillation (HP:0005110), Hypertro	DOI:10.3389/fphys.2022.864547	
201334327	DNA sequencing assay (OBI_0000626)	Atrial fibrillation (HP:0005110), Hypertro	DOI:10.3389/fphys.2022.864547	
201328755	DNA sequencing assay (OBI_0000626)	Atrial fibrillation (HP:0005110), Hypertro	DOI:10.3389/fphys.2022.864547	
201328364	DNA sequencing assay (OBI_0000626)	Atrial fibrillation (HP:0005110), Hypertro	DOI:10.3389/fphys.2022.864547	
112572057	DNA sequencing assay (OBI_0000626)	Primary dilated cardiomyopathy (MOND(	DOI:10.1111/j.1752-8062.201	The hotspot is in the exon 9 of the RBM2(
112572069	DNA sequencing assay (OBI_0000626)	Primary dilated cardiomyopathy (MOND(	DOI:10.1111/j.1752-8062.201	The hotspot is in the exon 9 of the RBM2(
112572120	polymerase chain reaction (OBI_0000	familial dilated cardiomyopathy (MONDC	DOI:10.1016/j.jacc.2009.05.0	The hotspot is located in the RS-domain o
22242747	DNA sequencing assay (OBI_0000626)	autosomal recessive limb-girdle muscula	DOI:10.1016/j.nmd.2015.03.0	The hotspot is the exon 5 of the ANO5 ge
22297640	DNA sequencing assay (OBI_0000626)	autosomal recessive limb-girdle muscula	DOI:10.1016/j.nmd.2015.03.0	The hotspot is the exon 20 of the ANO5 g
116707127	DNA sequencing assay (OBI_0000626)	AApoAI amyloidosis (MONDO:0019731)	DOI:10.2353/jmoldx.2009.080161	
116706820	DNA sequencing assay (OBI_0000626)	AApoAI amyloidosis (MONDO:0019731)	DOI:10.2353/jmoldx.2009.080161	
2604766	molecule detection assay evidence (E	long QT syndrome (MONDO:0002442)	DOI:10.1161/01.cir.100.10.1077	
2604775	molecule detection assay evidence (E	long QT syndrome (MONDO:0002442)	DOI:10.1161/01.cir.100.10.1077	
2592556	polymerase chain reaction (OBI_0000	long QT syndrome (MONDO:0002442)	DOI:10.1097/PAF.000000000(	The hotspot is located in exon 3 of the KC
2594216	polymerase chain reaction (OBI_0000	long QT syndrome (MONDO:0002442)	DOI:10.1097/PAF.000000000(	The hotspot is located in exon 6 of the KC
47359005	DNA sequencing assay (OBI_0000626)	Hypertrophic cardiomyopathy (MONDO:(	DOI:10.20452/pamw.15130	Tested in Polish population
47369231	polymerase chain reaction (OBI_000(	Hypertrophic cardiomyopathy (MONDO:(	DOI:10.1080/ac.67.1.2146562	

Figura 32: Fuente de datos de hotspots de cardiopatías familiares (II) . [Elaboración propia]

Centrándose en la etapa de explotación, y tal y como se ha explicado en el apartado 6.4, esta se materializa en una web a la que los usuarios objetivos pueden recurrir cuando necesiten consultar esta información. Actualmente, esta página web sigue en proceso de desarrollo y, es por ello por lo que todavía no está accesible para su uso. La página web consta de una pantalla principal en la que se dispone de un ideograma de los 22 autosomas [12] y los 2 sexuales pertenecientes al ser humano (Figura 33).

Figura 33: Página principal de la web *CardioHotspotsDatabase*. [Elaboración propia]

Si se desea acceder a información acerca de un hotspot simplemente se tiene que clicar en la zona donde se encuentra el triángulo rojo, ya que este indica la presencia de un hotspot en esa región. La decisión de emplear este símbolo y este color para determinar la posición dentro del cromosoma del hotspot no es trivial ya que, se pretende que el usuario sea capaz de identificar de forma rápida y sencilla donde se encuentran los hotspots disponibles en esta web.

Una vez se ha escogido la zona de interés, aparecen en la zona inferior de la pantalla 2 esquemas: (1) representando la cadena proteica de aminoácidos correspondientes a esa zona en la parte izquierda y (2) el diagrama circular que describe la secuencia proteica en la parte derecha. En la Figura 34 se puede observar cómo queda la web después de escoger una zona de interés. Justo debajo del ideograma se expone el gen en el que nos encontramos dentro de cada cromosoma y después se exponen los dos esquemas.

[12] **Autosomas**: cromosomas comprendidos entre el par 1 y el par 22.(*Autosoma*, n.d.).

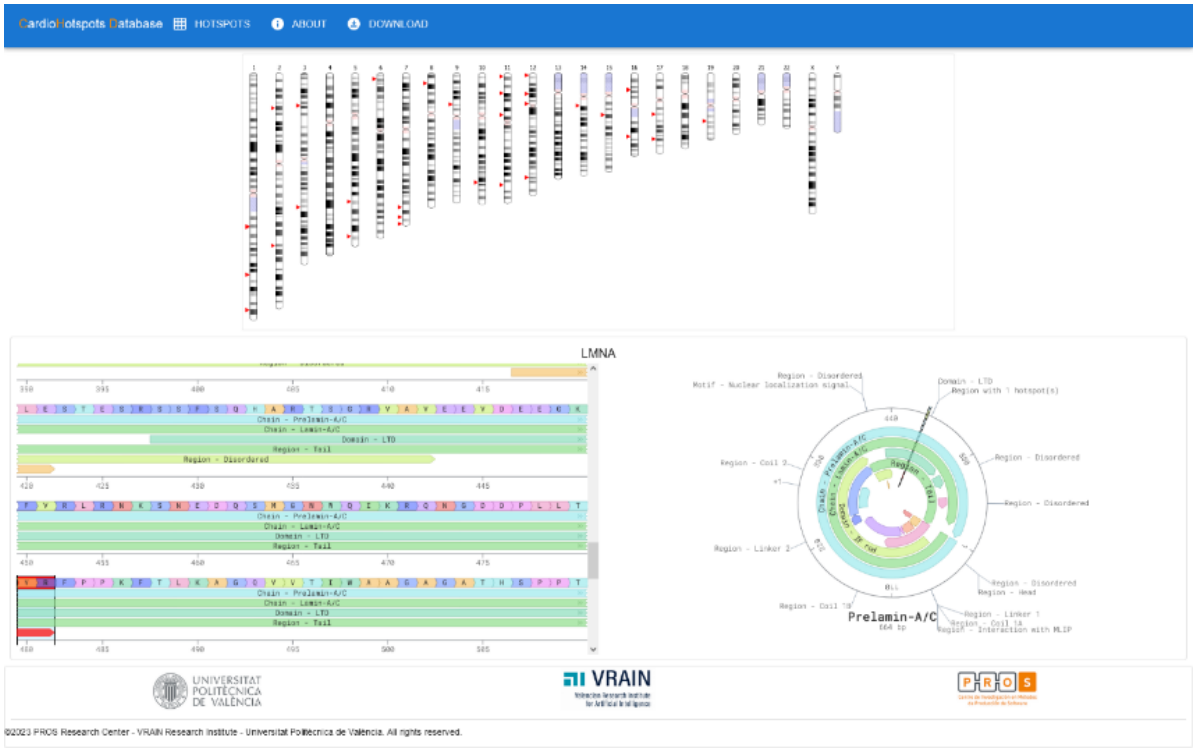


Figura 34: Vista de *CardioHotspotsDatabase* una vez se ha escogido una zona del ideograma. [Elaboración propia]

La forma del primer esquema es lineal, ya que se muestra la cadena de aminoácidos correspondiente a esa zona como una secuencia lineal representada por las abreviaciones de cada aminoácido. Arriba de los aminoácidos se observa la referencia proteica, es decir, la posición proteica en la que se encuentra cada aminoácido. Asimismo, debajo de la secuencia de aminoácidos tenemos los diferentes dominios y regiones proteicas que se codifican, así como la presencia de un hotspot o no en una región.

La forma del segundo esquema es circular y representa la secuencia proteica de una forma más visual. En esta representación se detallan aspectos como los dominios, la proteína codificada en cada caso y las distintas regiones de interés. El esquema de la parte izquierda de la interfaz está relacionado con el diagrama circular ya que, si quieres ver en qué parte del diagrama está una zona en concreto, simplemente tienes que clicar en la zona de interés y automáticamente te saldrá el diagrama acotado por dos separadores indicando a que zona se corresponde (Figura 35).

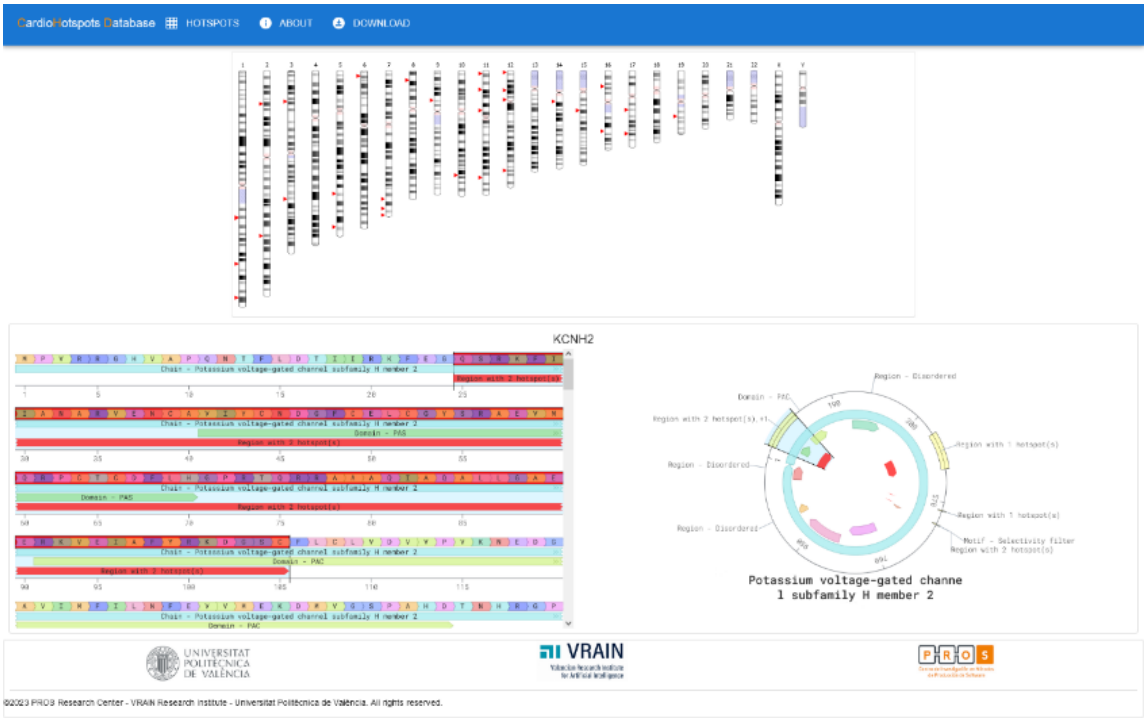


Figura 35: Vista de *CardioHotspotsDatabase* una vez escogida una zona de la representación de la cadena proteica en el diagrama circular descriptor de la secuencia proteica especificada por separadores. [Elaboración propia]

Esta página web tiene otras pestañas a parte de la principal y, una de las más importantes es la de HOTSPOTS. En la parte superior izquierda de la página web, se puede observar cómo existe una opción que pone HOTSPOTS. Si se clica sobre ese elemento, aparece una interfaz con una apariencia como la de la Figura 36 y la Figura 37. La interfaz se compone con una tabla con 9 columnas en las cuales se describen las siguientes características para caracterizar un hotspot: cromosoma, localización cromosómica, gen, inicio de aa, final de aa, método, fenotipo, referencia bibliográfica y comentarios.

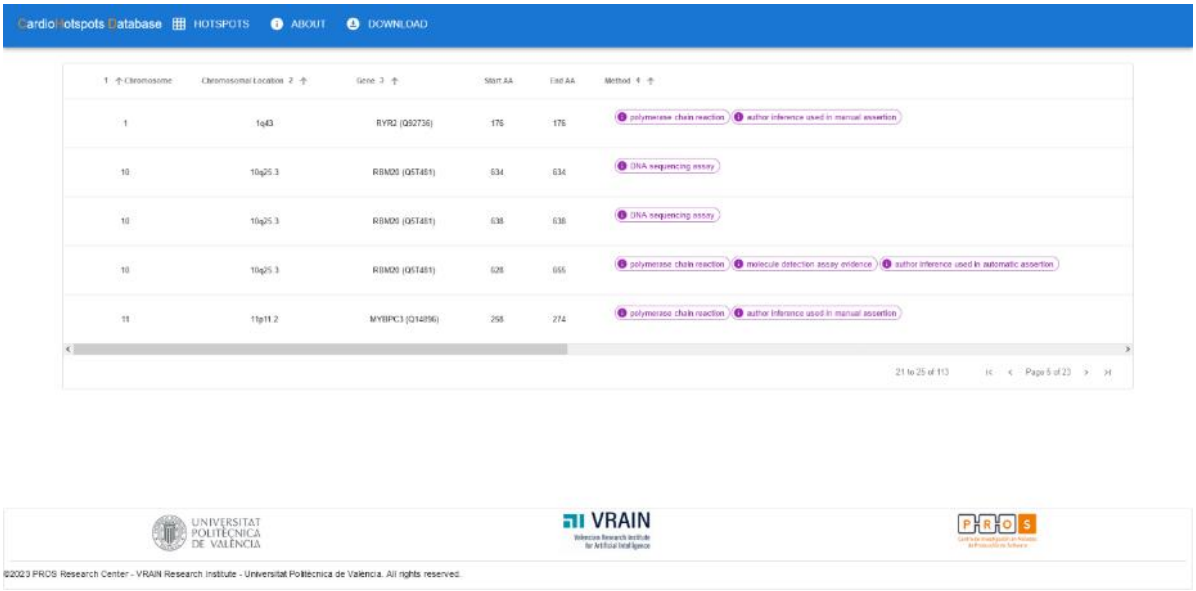


Figura 36: Interfaz de usuario al clicar HOTSPOTS. [Elaboración propia]

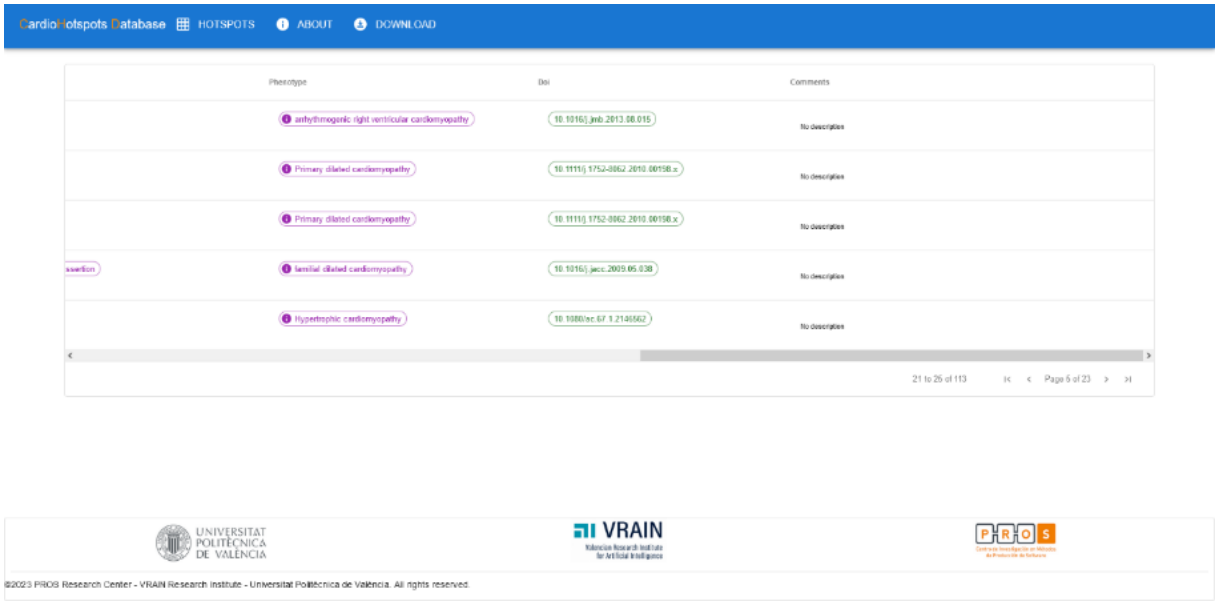


Figura 37: Apariencia de la tabla de hotspots al *scrollear*. [Elaboración propia]

Finalmente, la página web tiene una opción llamada ABOUT la cual contiene información acerca de los creadores de la página web y el propósito de esta, tal y como puede observarse en la Figura 38.

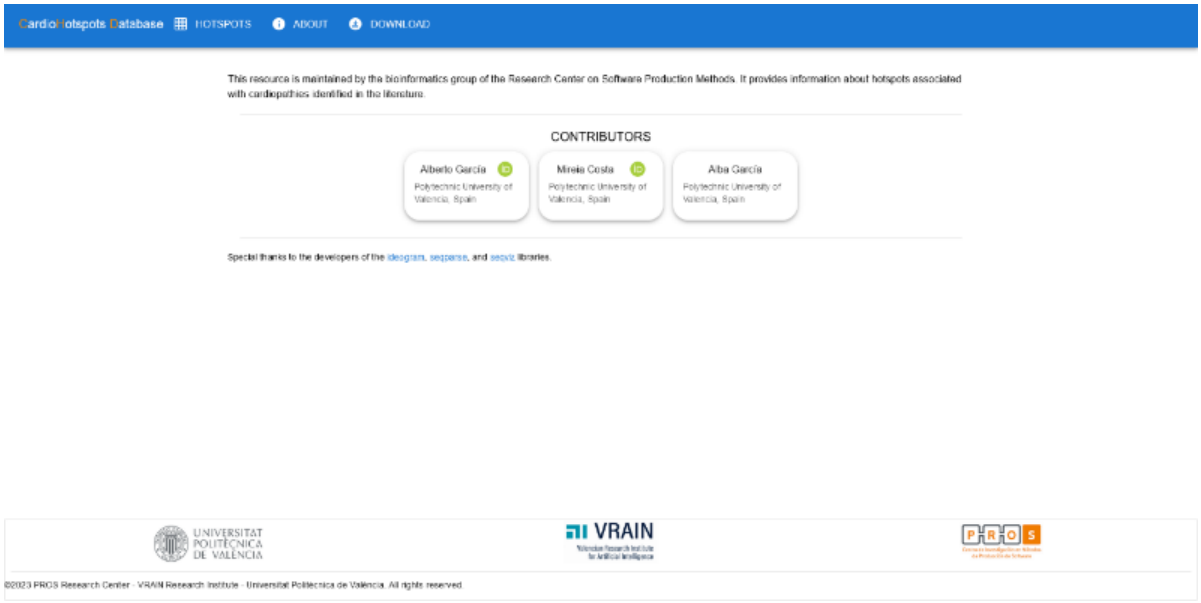


Figura 38: Interfaz de usuario al clicar en ABOUT . [Elaboración propia]

Tras el desarrollo de la página web se puede dar por finalizado el método SILE .Sin embargo, es crucial la validación de la fuente de datos realizada. La validación de estos resultados se explica en el apartado a continuación.

7.2. CASO DE USO

Para evaluar la exactitud y precisión de la fuente de datos realizada se comprueba la existencia de variaciones genómicas de pacientes reales dentro de los hotspots definidos por la propia fuente de

datos. Para conseguirlo se recurre al estudio de ficheros VCF[13] de pacientes del hospital La Fe de Valencia, provistos por el proyecto OGMIOS.

La evaluación de los resultados se basa en la búsqueda de relaciones entre las variaciones genómicas detectadas y reportadas a través de los ficheros VCF de los hospitales y los hotspots detectados en este Trabajo de Final de Grado. La estrategia de evaluación se basa en la comparación de los datos obtenidos y los provistos para obtener similitudes. Dicha comparación se lleva a cabo mediante un código comparativo elaborado en Python, con el cual, mediante el uso de comparadores lógicos y aritméticos, se pretende obtener una lista de las variaciones de los VCF de cada paciente que se encuentran dentro de alguno de los hotspots definidos en la fuente de datos. El código empleado puede consultarse en el Anexo 4. Por otra parte, en la Figura 39 se puede ver un esquema de la estrategia de evaluación.

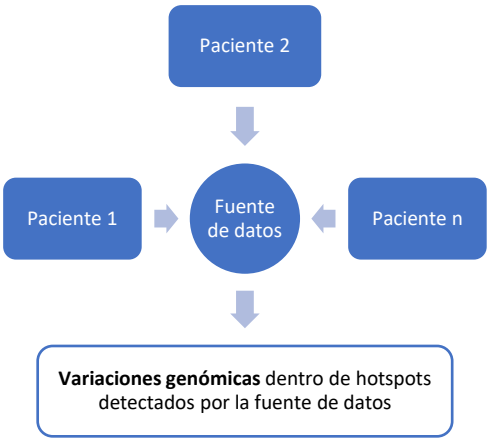


Figura 39: Esquema de la estrategia de evaluación de los resultados. [Elaboración propia]

Los resultados obtenidos a raíz de esta etapa de evaluación fueron muy positivos ya que, todos los pacientes contenían variaciones genómicas dentro de hotspots en los ficheros VCF a excepción de dos de ellos. La suma de variaciones genómicas encontradas dentro de hotspots caracterizados en la fuente de datos llega a las 970. Para organizar todas estas variaciones, en la Figura 40 y la Figura 41 se puede observar la cantidad de variaciones genómicas encontradas en los ficheros VCF en cada paciente.

[13] **Variant Call Format o VCF**: Fichero de texto usado en Bioinformática para almacenar variaciones de la secuencia de genes y su información.

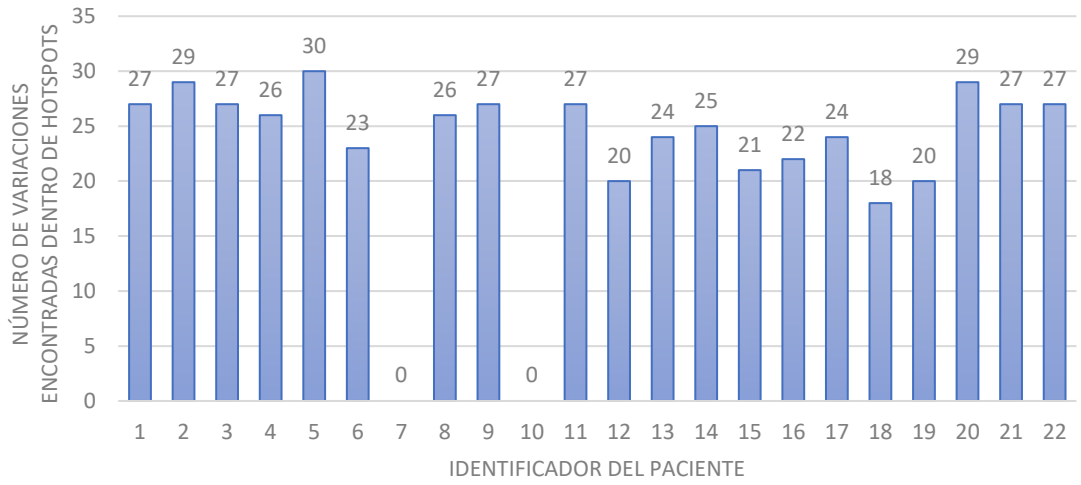


Figura 40: Distribución de variaciones encontradas dentro de hotspots por la fuente de datos de pacientes reales (I) . [Elaboración propia]

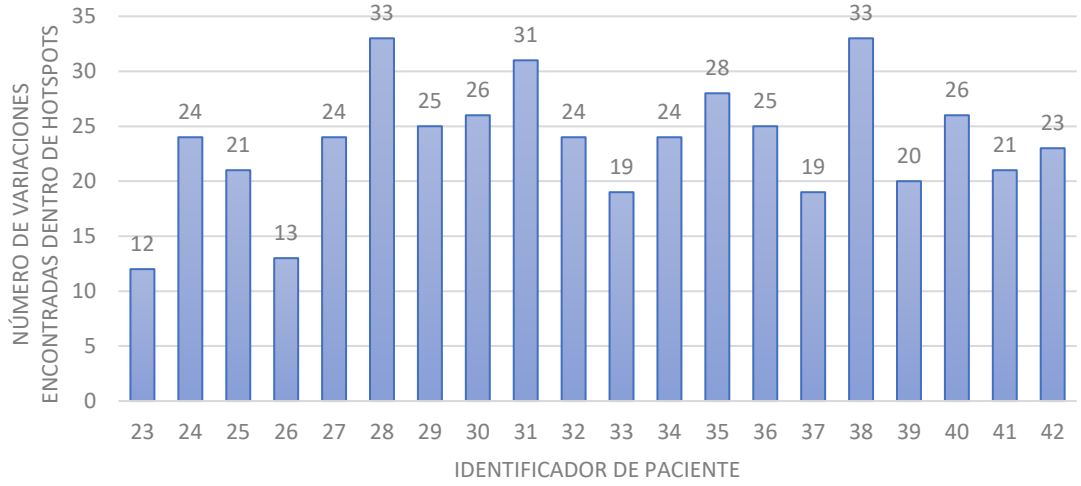


Figura 41:Distribución de variaciones encontradas dentro de hotspots por la fuente de datos de pacientes reales (II) . [Elaboración propia]

Si analizamos los resultados obtenidos en la Figura 40 y la Figura 41 podemos observar como la gran mayoría de los pacientes presentan alrededor de 20 variaciones genómicas dentro de hotspots en sus ficheros VCF. Estos resultados son indicativos de que la fuente de datos realmente contiene hotspots reales en su archivo ya que, muchos pacientes tenían variaciones genómicas en alguno de esos hotspots definidos por la literatura. Una de las posibles aplicaciones de este conocimiento sería estudiar estas variaciones genómicas identificadas dentro de un hotspot más a fondo ya que, según los criterios de clasificación de las guías ACMG-AMP del 2015 (apartado 5.4), son más susceptibles a ser patogénicas.

Otro tipo de análisis interesante a estudiar consiste en analizar la distribución de las variaciones genómicas encontradas dentro de hotspots en pacientes reales en función de diferentes parámetros como el cromosoma o el gen al que afectan. Si se centra la atención en el estudio de las variaciones genómicas encontradas dentro de hotspots en función del cromosoma al que pertenecen, los resultados son los expuestos en la Figura 42.

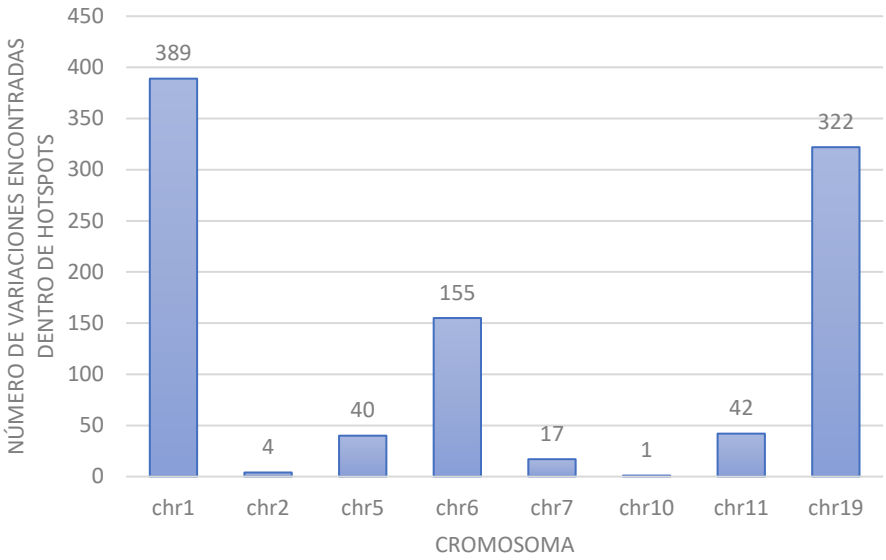


Figura 42: Distribución de variaciones genómicas encontradas dentro de pacientes en función del cromosoma al que pertenecen. [Elaboración propia]

Estos resultados muestran que, de los 16 cromosomas con hotspots encontrados en la literatura, tan solo 8 de ellos contienen variaciones genómicas dentro hotspots en pacientes reales. Estos cromosomas son 1,2,5,6,7,10,11 y 19. Estos resultados resultan bastante evidentes ya que, la mayoría de información encontrada durante el diseño y desarrollo de la fuente de datos es sobre esos cromosomas (Figura 29). Si nos centramos en un análisis porcentual de estos resultados podemos observar como el cromosoma 1 y el cromosoma 19 son los más identificados dentro de pacientes reales, suponiendo un 40,10% y un 33,20% de las variaciones genómicas encontradas respectivamente. Tras estos cromosomas, les siguen el cromosoma 6 con un 15,98%, el cromosoma 11 con un 4,33% y el cromosoma 5 con un 4,12%. Finalmente tenemos el cromosoma 7 con un 1,75%, el cromosoma 2 con un 0,42% y el cromosoma 10 con un 0,1%.

Por otro lado, si se analiza la distribución de estas variaciones en función del gen al que afectan se obtienen los resultados de la Figura 43. Uno de los hechos más destacables de este análisis es ver como de los 33 genes con hotspots detectados en la fuente de datos, solo 9 de ellos aparecen en los ficheros VCF de pacientes. Los genes que aparecen en los ficheros VCF son : ANO5, DSP, KCNH2, KCNQ1, MYOT, RBM20, RYR1, RYR2 y TTN. Es importante recalcar que esos genes se corresponden con los cromosomas encontrados anteriormente. El gen ANO5 se corresponde con el cromosoma 11, el gen DSP con el cromosoma 6, el gen KCNH2 con el 7, el KCNQ1 con el 11, el MYOT con el 5, el RBM20 con el 10, el RYR1 con el 19, el RYR2 con el 1 y el TTN con el 2. Estos resultados demuestran que la mayoría de los cromosomas afectados por variaciones genómicas dentro de hotspots se enmarcan dentro del mismo gen.

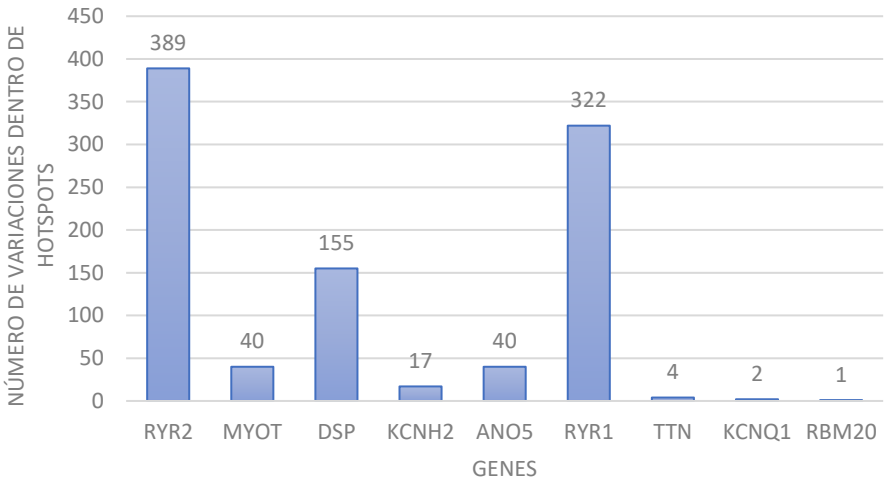


Figura 43: Distribución de variaciones genómicas en hotspots en función del gen al que afecta. [Elaboración propia]

Si analizamos los resultados obtenidos en la Figura 43, se puede observar cómo la distribución de variaciones por genes es muy parecida a la distribución de variaciones por cromosomas (Figura 42). En cuanto a los porcentajes de las variaciones distribuidas por genes se tiene que el 40,10% se corresponde con el gen RYR2, el 33,20% se corresponde con el gen RYR1 y el 15,98% con el gen DSP. Asimismo, se tienen dos porcentajes del 4,13% correspondientes con el gen MYOT y ANO5.En menor medida tenemos el gen KCNH2 representando el 1,76% de las variaciones, y los genes TTN, KCNQ1 y RBM20, representados en un 0,43%, 0.22% y 0,11% respectivamente.

De este análisis se puede concluir que los genes más susceptibles a contener variaciones genómicas dentro de hotspots son el RYR2, RYR1 y DSP. Estos resultados demuestran que los genes de los que más información se tiene son los más detectados en pacientes reales. Por este motivo resulta fundamental este Trabajo de Final de Grado, ya que este es el inicio de una línea de investigación acerca de estos genes en relación con cardiopatías y hotspots.

El hecho de que estos genes hayan sido los más identificados parece razonable por la naturaleza de estos. El gen RYR1 es el encargado de codificar los receptores de rianodina 1. Estos receptores son proteínas esenciales para el correcto funcionamiento de los músculos ya que se encargan de la correcta liberación de calcio en el retículo sarcoplásmico (Witherspoon & Meilleur, n.d.). Esta función de los receptores es muy relevante ya que el calcio es uno de los iones que regula la contracción cardíaca y, por tanto, su correcto funcionamiento resulta crucial para garantizar un corazón sano. Las mutaciones que se presentan en este gen suelen estar asociadas a fenotipos como la hipertermia maligna (Rosenberg et al., 2015), enfermedades de núcleo central (Shepherd et al., 2004)y miopatía minicore con oftalmoplejía externa (Kizer et al., 1976).

El gen RYR2 es el encargado de codificar los receptores de rianodina 2. Estos receptores se encuentran en el retículo sarcoplásmico del músculo cardiaco y estos, al igual que los receptores de rianodina 1, son uno de los componentes de los canales de calcio (Priori & Napolitano, 2005). Estos receptores junto con los anteriores forman los canales de calcio y, por tanto, son muy importantes en la contracción cardíaca. Las mutaciones en este gen se asocian con distintas cardiopatías como la taquicardia ventricular polimórfica (Wleklinski et al., 2020) y la displasia arritmogénica del ventrículo derecho (Meurs et al., 2006).

El tercer gen con más variaciones genómicas dentro de pacientes reales es el DSP. Este gen codifica una proteína encargada de anclar los filamentos intermedios de las placas desmosómicas y la forma

de uno de los componentes de los desmosomas funcionales (Kiselev et al., 2016). Los desmosomas son las uniones que vinculan una célula con otra adyacente (Delva et al., n.d.). Este gen regula la expresión profibrótica de los cardiomiocitos, es decir, se encarga de la expresión de más o menos fibra en los cardiomiocitos (Den Haan et al., 2009). Este hecho resulta relevante a la hora de la contracción muscular, ya que las fibras musculares son un componente fundamental del proceso de contracción. Las mutaciones de este gen son causantes de varias cardiomiopatías y queratodermias (Yao & Winship, 2020).

Toda esta información biológica acerca de estos genes puede resumirse como que estos tres genes tienen una función primordial para el correcto desarrollo de la función contráctil del corazón. Además, los genes RYR1 e RYR2 tienen un papel primordial en el control de paso de iones a través de los canales eléctricos ya que, ambos dos, son los encargados de codificar las proteínas encargadas del desarrollo de dichos canales. Es por ello por lo que las mutaciones que se lleven a cabo en dichos genes pueden desarrollar miocardiopatías y canalopatías principalmente. Dado que la mayoría de las variaciones genómicas encontradas en los pacientes que están dentro de hotspots se encuentran en estos tres genes, y teniendo en cuenta la definición de hotspot, podemos plantear como hipótesis que los hotspots asociados a cardiopatías familiares se centran en los fenotipos relacionados con las miocardiopatías y las canalopatías concretamente.

Para hacer un análisis más completo de la evaluación de los resultados, se hizo un estudio comparativo de los diagnósticos hechos por los clínicos y sus variaciones, y los resultados de variaciones contenidos en hotspots. El principal objetivo de este análisis reside en ver si las variaciones causantes de las enfermedades, determinadas por los expertos clínicos, se encuentran dentro de regiones descritas como hotspots por nuestra fuente de datos. Los resultados de este análisis (Anexo 5) dejan a relucir que ninguna de las variaciones definidas como diagnósticas por parte de los expertos clínicos se encuentran dentro de los hotspots de la fuente de datos. Aunque, a priori puedan parecer malos resultados, la realidad es que el hecho de que una variación pueda ser considerada patogénica depende de distintos factores y no únicamente de si se encuentra dentro de un hotspot o no . Asimismo, es importante comentar que muchos de los pacientes no se habían sometido a pruebas genéticas diagnósticas y, por tanto, no se han podido analizar.

Cabe recalcar que, el hecho de que una variación genómica se encuentre dentro de un hotspot nos proporciona una evidencia moderada de patogenicidad, tal y como queda descrito en las guías ACMG – AMP del 2015 y, por tanto, no se espera que todas ellas sean las causantes de la enfermedad. Este hecho simplemente advierte al clínico sobre posibles indicadores diagnósticos, pero no hay que olvidar que no son factores discriminantes para diagnosticar cualquier tipo de enfermedad.

Sin embargo, se puede observar que muchos de los genes sí que coinciden con los descritos en la fuente de datos. Esto suscita diferentes hipótesis; por un lado, se puede pensar que hay más regiones con alta frecuencia de mutagénesis no plasmadas en la literatura que son las causantes de ciertas enfermedades, y por el otro lado, que las variaciones genómicas causantes de estas enfermedades no se encuentran dentro de hotspots.

Asimismo , es importante comentar que muchos de los pacientes no se han podido analizar debido a diferentes motivos; uno de ellos es que los clínicos no les han realizado pruebas genéticas para averiguar si existen mutaciones, otro es debido a que las pruebas se han realizado en laboratorios privados o por el hecho de no haber encontrado mutaciones en las pruebas diagnósticas. Este último hecho representa la posibilidad de que ciertas cardiopatías no tienen por qué tener un patrón hereditario o genético.

CAPÍTULO 8. CONCLUSIONES Y TRABAJOS FUTUROS

Tras la realización de este trabajo se ha afirmado una vez más la envergadura de la problemática del caos genómico de datos al que nos enfrentamos actualmente. La comunidad científica se enfrenta al reto de unificar toda la información genética que tenemos almacenada. La búsqueda de la información a través de las diferentes bases de datos resulta un reto ya que la información se encuentra almacenada de forma muy heterogénea y, en muchos casos, de forma incompleta. La homogeneización de toda esa información podría suponer un paso hacia delante dentro de la comunidad científica en el ámbito de la genética ya que resolvería muchos problemas de interpretación de esa información. Es por ello por lo que el diseño y desarrollo de una fuente de datos de hotspots de cardiopatías familiares realizada mediante la identificación y caracterización de hotspots a través del modelo conceptual cobra sentido por sí misma.

Durante este Trabajo de Final de Grado y, más concretamente en el capítulo 2, se han determinado 3 objetivos específicos acordes a las dos primeras etapas del ciclo regulativo propuesto por el *Design Science* además de una tercera sobre la evaluación de la fuente de datos con preguntas de investigación asociadas a cada uno. Con estos objetivos se pretende realizar una correcta investigación determinando parámetros importantes tales como los usuarios objetivos o la base metodológica y conceptual que va a seguirse durante el trabajo.

En el apartado 8.1. se van a resumir las respuestas específicas que se le ha dado a cada una de las preguntas de investigación a lo largo de este Trabajo de Final de Grado y en la sección 8.2. se detallan futuras líneas de trabajo derivadas de este Trabajo de Final de Grado.

8.1. PREGUNTAS DE INVESTIGACIÓN

Esta sección se divide a su vez en 3 subsecciones, una por cada objetivo, en las que se van a resolver de forma detallada cada una de las preguntas de investigación.

8.1.1. Objetivo 1. Investigación del problema

PI1. ¿Cuáles son los usuarios objetivo?

En el apartado 5.1 se han definido los usuarios objetivos de la fuente de datos de hotspots desarrollada. En esa sección puede observarse como se presentan dos potenciales usuarios; los expertos clínicos y los analistas de datos genómicos. Por un lado, los expertos clínicos se benefician de esta fuente de datos con fines diagnósticos mientras que, por el otro lado, los analistas de datos se benefician de esta porque les facilita la gestión de información genómica para favorecer la creación de conocimiento.

PI2. ¿Qué necesidades de estos usuarios se espera cubrir con la fuente de datos?

En el apartado 5.2 se han definido cuatro necesidades básicas que cubre esta fuente de datos y son la estandarización de la información, el fácil acceso a la información, la integración de la información y la caracterización de un hotspot.

PI3. ¿Qué definición se adopta como hotspot?

En el apartado 5.3 se exponen las diferentes definiciones de hotspot existentes en la literatura. Para este trabajo se ha empleado la definición de hotspot 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 se observan mutaciones con más frecuencia de lo habitual".

PI4. ¿Cuál es la base conceptual del trabajo?

La base conceptual del trabajo es el modelado conceptual, tal y como queda expuesto en el apartado 5.5 de la memoria. En este se explica en qué consiste el modelado conceptual del genoma humano y como se estructura este.

PI5.¿Cómo interpretamos una variación genómica dentro de un hotspot?

En el apartado 5.4 se explica que la interpretación de las variaciones genómicas por excelencia se basa en el uso de las guías ACMG- AMP del 2015. Más concretamente el término de hotspot va asociado al criterio PM1 de estas guías, el cual les atribuye una patogenicidad moderada.

PI6. ¿Qué base conceptual se va a emplear para caracterizar un hotspot?

La base conceptual para la caracterización de un hotspot es el modelo conceptual debido a las prestaciones que esta base nos ofrece, tal y como se explica en el apartado 5.5.

PI7. ¿Cuál es la base metodológica que debería utilizarse en la fuente de datos?

En el apartado 5.6 se presenta el método SILE como base metodológica de este trabajo. Este método es el resultado de la tesis doctoral de la Dra. Ana León Palacio titulada: *SILE: A method for the Efficient Management of Smart Genomic Information* .

8.1.2. Objetivo 2. Diseño y desarrollo de la fuente de datos

PI8. ¿Qué estructura tiene la fuente de datos?

En el apartado 6.1 se presenta la estructura de la fuente de datos basada en el modelo conceptual que caracteriza un hotspot. Alguno de los aspectos más importantes es la definición de este en base a dos marcos: el proteico y el genómico.

PI9.¿Cómo deben seleccionarse los artículos?

Tal y como queda explicado en el apartado 6.2, el criterio de selección de artículos se basa en distintos cribados con procesos de selección intermedios. El primer cribado parte de la base de una regla de búsqueda en PubMed mediante su opción avanzada empleando operadores lógicos. El segundo cribado consiste en la identificación del tipo de fenotipo que presenta cada artículo para discernir si se trata de un fenotipo relacionado con cardiopatías o no.

PI10. ¿Cómo debe realizarse la extracción de datos?

En el apartado 6.3 se detalla el proceso de identificación de la información relevante para definir un hotspot. Este proceso se caracteriza por la sucesión de una serie de procesos tales como la lectura del artículo para sacar información, el estudio de la viabilidad de la extracción de la información de este

(apartado 6.3.1), el empleo de herramientas auxiliares para completar la información (apartado 6.3.2) y la estandarización de la misma (apartado 6.3.3).

PI11. ¿Cómo debe de estandarizarse la información?

El proceso de estandarización de la información queda expuesto en el apartado 6.3.3. En este se explica la importancia del empleo de ontologías para estandarizar los atributos del fenotipo y el método o fuente. Asimismo, se estandarizan las posiciones genómicas – enmarcada dentro del *assembly* 37 - y proteicas mediante el empleo de un código de R.

PI12. ¿Cómo se presentará esta fuente de datos hacia el usuario?

En el apartado 6.4 se expone la presentación de la fuente de datos a través de un formato web al usuario. Esta interfaz permite una mejor interoperabilidad y manejo de los usuarios objetivo con la información debido a que el uso de la fuente de datos en crudo puede resultar complicada y confusa.

8.1.3. Objetivo 3. Evaluación de la fuente de datos

PI13. ¿Cuáles han sido los resultados obtenidos?

Los resultados obtenidos del diseño y desarrollo de la fuente de datos de cardiopatías familiares se han descrito en el apartado 7.1. En concreto, se han descrito los resultados obtenidos en cada una de las etapas del método SILE: *Search*(sección 7.1.1), *Identify*(sección 7.1.2) y *Load and Explotation* (sección 7.1.3).

PI14. ¿La fuente de datos cumple los objetivos expuestos?

Se concluye que la fuente de datos sí que cumple con los objetivos expuestos tal y como puede observarse en los resultados de validación del apartado **Error! Reference source not found.**, en el cual s e exponen los resultados de estudiar ficheros VCF en base a la fuente de datos.

8.2. TRABAJOS FUTUROS

En esta sección se van a detallar las futuras líneas de trabajo para garantizar la continuidad del estudio de la fuente de datos desarrollada en este Trabajo de Final de Grado.

La primera aproximación de vías de trabajo futuro implica la ampliación de la propia fuente de datos, gracias al empleo de otros buscadores de artículos científicos tales como *ScienceDirect*. Esta herramienta es el inicio de una fuente de datos que se pretende actualizar de forma periódica para garantizar que la información que esta contenga sea actualizada y completa.

Asimismo, también es interesante estudiar en la revisión y mejora continua del modelo conceptual a medida que el nuevo conocimiento acerca de hotspots se haga disponible. El modelo conceptual será mejor cuantas menos variables precise para definir un hotspots en su representación y, por ello, el estudio más extenso de los hotspots mutacionales dentro del ámbito de la genética resulta crucial a la hora de mejorar la fuente de datos.

Por otro lado, el ampliar el análisis de los datos obtenidos mediante el uso de estudios estadísticos puede ser una vía de mejora y un posible trabajo futuro. Estos test son capaces de buscar relaciones entre las variaciones, los genes y los hotspots para intentar acercarse más al entendimiento de estas regiones del genoma. Asimismo, esta implementación pretende proporcionar una propuesta de valor

añadida a nuestra fuente de datos más allá de la presentación de una forma clara y ordenada de la información existente acerca de los hotspots de cardiopatías.

Otra vía de mejora es la actualización de la web y la inserción de nuevas funcionalidades que puedan ser útiles para los usuarios objetivo. Para conseguir este cometido se debe hacer un estudio en profundidad y conjunto con los potenciales usuarios para así poder garantizar que la página web cubre todas sus necesidades. Una de las propuestas de mejora de esta es la de añadir la funcionalidad de subir la información genética que el usuario está estudiando y que la página web sea capaz de encontrar coincidencias o diferencias para poder hacer un análisis de los resultados. Gracias a esa actualización se conseguiría semiautomatizar el proceso de búsqueda de variaciones dentro de hotspots y les permitiría a los clínicos agilizar el proceso de diagnóstico de enfermedades. Todas las herramientas que puedan ofrecer rapidez a la hora del diagnóstico resultan atractivas para los expertos clínicos y, por tanto, esta sería una buena funcionalidad que añadir a la página web.

Además, también sería interesante ampliar esta fuente a otros ámbitos clínicos como son la oncología o las retinopatías, dado que estos ámbitos son en los que se centra el grupo de investigación. Es importante recordar que el estudiar solo las cardiopatías familiares es debido a los proyectos con los que se colabora (OGMIOS y CARDIOVAL) y que hay muchas otras enfermedades genéticas que resulta interesante estudiar.

Finalmente, otra posible propuesta de futuro se basaría en la automatización del proceso de extracción de información. Es importante recalcar que este proceso se ha hecho de forma manual y que supone una cantidad de horas de trabajo por parte del personal muy significativa. Es por ello por lo que sería muy interesante desarrollar una forma de automatizar esa extracción de la información de artículos mediante técnicas basadas en inteligencia artificial empleando el procesado del lenguaje natural. La propuesta de emplear este tipo de técnicas reside en que estas pretenden comprender textos no estructurados para que extraigan información sobre ellos y, precisamente eso es lo que pretendemos automatizar.

En conclusión, el diseño y desarrollo de esta fuente de datos es una pequeña contribución al ordenamiento del caos de datos genómicos dentro del proyecto OGMIOS y CARDIOVAL. Es por ello por lo que esta fuente de datos solo es el inicio de esta línea de investigación que tiene su continuidad en el Trabajo de Final de Máster que se va a realizar el curso siguiente gracias a las competencias que voy a adquirir en el Máster en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones.

CAPÍTULO 9: REFERENCIAS BIBLIOGRÁFICAS

▷ *Valor Q: definición y ejemplos en 2023 → STATOLOGOS®*. (n.d.). Retrieved July 1, 2023, from https://statologos.com/valor-q/

Ackerman, M. J., Marcou, C. A., & Testera, D. J. (2013). Medicina personalizada: diagnóstico genético de cardiopatías/canalopatías hereditarias. *Rev. Esp. Cardiol. (Ed. Impr.)*, 298–307.

Alberto Garcia, S., Palacio, A. L., Roman, J. F. R., Casamayor, J. C., & Pastor, O. (2021). A Conceptual Model-Based Approach to Improve the Representation and Management of Omics Data in Precision Medicine. *IEEE Access*, *9*, 154071–154085. https://doi.org/10.1109/ACCESS.2021.3128757

Amberger, J., Bocchini, C., & Hamosh, A. (2012). The Reference Sequence (RefSeq) Database. *Human Mutation*, *32*(5), 564–567. https://doi.org/10.1002/HUMU.21466

Argelia Medeiros-Domingo, D., & Medeiros-Domingo, A. (2009). Genética de la taquicardia ventricular polimorfa catecolaminérgica; conceptos básicos. *Arch Cardiol Mex*, *79*, 13–17. www.elsevier.com.mx

Autosoma. (n.d.). Retrieved July 2, 2023, from https://www.genome.gov/es/genetics-glossary/Autosoma

Ayerza Casas, A., López Ramón, M., Palanca Arias, D., & Jiménez Montañés, L. (2017). Afectación cardiovascular en el síndrome de Loeys-Dietz. *Anales de Pediatría*, *86*(1), 54–55. https://doi.org/10.1016/J.ANPEDI.2015.11.009

Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M. A., He, Y., Heiskanen, M., Hernandez-Boussard, T., … Zheng, J. (2016). The Ontology for Biomedical Investigations. *PLOS ONE*, *11*(4), e0154556. https://doi.org/10.1371/JOURNAL.PONE.0154556

Barriales-Villa, R., García-Giustiniani, D., & Monserrat, L. (2011). Genética del síndrome de Marfan. *Cardiocore*, *46*(3), 101–104. https://doi.org/10.1016/J.CARCOR.2011.05.001

Benito, B., Brugada, J., Brugada, R., & Brugada, P. (2009). Brugada Syndrome. *Revista Española de Cardiología (English Edition)*, *62*(11), 1297–1315. https://doi.org/10.1016/S1885-5857(09)73357-2

Bernasconi, A., García, A., Ceri, S., & Pastor, O. (n.d.). *A comprehensive approach for the conceptual modeling of genomic data*.

Boccia, A., Petrillo, M., di Bernardo, D., Guffanti, A., Mignone, F., Confalonieri, S., Luzi, L., Pesole, G., Paolella, G., Ballabio, A., & Banfi, S. (2005). DG-CST (Disease Gene Conserved Sequence Tags), a database of human–mouse conserved elements associated to disease genes. *Nucleic Acids Research*, *33*(suppl_1), D505–D510. https://doi.org/10.1093/NAR/GKI011

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot. *Methods in Molecular Biology (Clifton, N.J.)*, *406*, 89–112. https://doi.org/10.1007/978-1-59745-535-0_4

Cardentey, M. C., Rosabal, A. M., Vigoa, A. V., & Cruz, A. V. (2009). Síndrome de QT corto. *Clínica e Investigación En Arteriosclerosis*, *21*(4), 193–197. https://doi.org/10.1016/S0214-9168(09)72046-7

Cardiopatías familiares y genética - Fundación Española del Corazón. (n.d.-a). Retrieved June 28, 2023, from https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/cardiopatias-familiares-y-genetica.html

Cardiopatías familiares y genética - Fundación Española del Corazón. (n.d.-b). Retrieved July 1, 2023, from https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/cardiopatias-familiares-y-genetica.html

Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., Gao, J., Socci, N. D., Solit, D. B., Olshen, A. B., Schultz, N., & Taylor, B. S. (2015). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature Biotechnology*, *34*(2). https://doi.org/10.1038/nbt.3391

Chang, M. T., Bhattarai, T. S., Schram, A. M., Bielski, C. M., Donoghue, M. T. A., Jonsson, P., Chakravarty, D., Phillips, S., Kandoth, C., Penson, A., Gorelick, A., Shamu, T., Patel, S., Harris, C., Gao, J. J., Sumer, S. O., Kundra, R., Razavi, P., Li, B. T., … Taylor, B. S. (2018). Accelerating discovery of functional mutant alleles in cancer. *Cancer Discovery*, *8*(2), 174–183. https://doi.org/10.1158/2159-8290.CD-17-0321/333221/AM/ACCELERATING-DISCOVERY-OF-FUNCTIONAL-MUTANT

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., Gabriel, S., … Yohannes, M. T. (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *BioRxiv*, 2022.03.20.485034. https://doi.org/10.1101/2022.03.20.485034

Choi, K., & Henderson, I. R. (2015). Meiotic recombination hotspots - A comparative view. *Plant Journal*, *83*(1), 52–61. https://doi.org/10.1111/TPJ.12870

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/NAR/GKAB1049

Database Summary Paper Categories. (n.d.). Retrieved July 1, 2023, from https://www.oxfordjournals.org/nar/database/cat/1

Delva, E., Tucker, D. K., & Kowalczyk, A. P. (n.d.). *The Desmosome*. https://doi.org/10.1101/cshperspect.a002543

Den Haan, A. D., Tan, B. Y., Zikusoka, M. N., Lladó, L. I., Jain, R., Daly, A., Tichnell, C., James, C., Amat-Alarcon, N., Abraham, T., Russell, S. D., Bluemke, D. A., Calkins, H., Dalal, D., & Judge, D. P. (2009). Comprehensive desmosome mutation analysis in North Americans with arrhythmogenic right ventricular dysplasia/cardiomyopathy. *Circulation: Cardiovascular Genetics*, *2*(5), 428–435. https://doi.org/10.1161/CIRCGENETICS.109.858217

Escobar Cervantes, C., Echarri Carrillo, R., Amador Borrego, A., Tarancón Zubimendi, B., Salido Tahoces, L., & Barrios Alonso, V. (2005). Síndrome de QT largo congénito: revisión de las diferentes

variantes y tratamientos. *Revista Costarricense de Cardiología*, *7*(1), 23–29. http://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S1409-41422005000100005&lng=en&nrm=iso&tlng=es

Fabián Reyes Román, J., & Pastor López, Ó. (n.d.). *Dirigida por: Febrero 2018*.

Fenotipo. (n.d.). Retrieved July 2, 2023, from https://www.genome.gov/es/genetics-glossary/Fenotipo

FireBrowse. (n.d.). Retrieved July 1, 2023, from http://firebrowse.org/

Fundación Instituto Roche - Glosario de genética - Puntos calientes de mutación. (n.d.). Retrieved July 2, 2023, from https://www.institutoroche.es/recursos/glosario/Puntos+calientes+de+mutaci%C3%B3n

Fundación Instituto Roche - Glosario de genética - transcrito primario. (n.d.). Retrieved July 2, 2023, from https://www.institutoroche.es/recursos/glosario/transcrito+primario

García S, A., Costa, M., Leon, A., & Pastor, O. (2022). The challenge of managing the evolution of genomics data over time: a conceptual model-based approach. *BMC Bioinformatics*, *23*(11), 1–33. https://doi.org/10.1186/S12859-022-04944-Z/FIGURES/15

Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., Mitraka, E., Schriml, L. M., Gaudet, P., Hobbs, E. T., Erill, I., Siegele, D. A., Hu, J. C., Mungall, C., & Chibucos, M. C. (2019). ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Research*, *47*(D1), D1186–D1194. https://doi.org/10.1093/NAR/GKY1036

Ginesi, A., Santillán, J., Puebla, N., Bertolino, T., Abud, M., Marín, J., Ciambrone, G., & Nogués, I. (n.d.). *MONOGRAFÍA SELECCIONADA Antiagregación en implantación valvular aórtica transcatéter: doble versus simple ARTÍCULOS ORIGINALES Seguimiento a 8 años de aneurismas de aorta abdominal. Registro unicéntrico CASOS CLÍNICOS Disección arterial coronaria espontánea IMÁGENES EN CARDIOLOGÍA Una causa infrecuente de disnea El síndrome del QT corto congénito: avances en los últimos años*. Retrieved July 2, 2023, from www.conarec.org

Guía de aplicación clínica de la secuenciación masiva en síndromes mielodisplásicos y leucemia mielomonocítica crónica – GCECGH. (n.d.). Retrieved July 2, 2023, from https://www.gcecgh.org/guia-de-aplicacion-clinica-de-la-secuenciacion-masiva-en-sindromes-mielodisplasicos-y-leucemia-mielomonocitica-cronica/

Human Phenotype Ontology. (n.d.). Retrieved July 2, 2023, from https://hpo.jax.org/app/

Illumina Clinical Services Laboratory Assertion Criteria for Gene Curation. (n.d.). Retrieved July 2, 2023, from https://www.ncbi.nlm.nih.gov/clinvar/

Janet, D., Ortiz-Bautista, G., Colín-Hernández, D. J., León-Álvarez, D. E., & Pediatra, A. (2023). *www.medigraphic.org.mx Artículo de revisión Hipertermia maligna Malignant hyperthermia*. *46*(1), 38–45. https://doi.org/10.35366/108621

Jarke, M., & Quix, C. (2017). On Warehouses, Lakes, and Spaces: The Changing Role of Conceptual Modeling for Data Integration. *Conceptual Modeling Perspectives*, 231–245. https://doi.org/10.1007/978-3-319-67271-7_16

Kiselev, A., Mikhaylov, E., Parmon, E., Sjoberg, G., Sejersen, T., Tarnovskaya, S., Nugnyi, P., Mitrofanova, L., Lebedev, D., & Kostareva, A. (2016). Progressive cardiac conduction disease

associated with a DSP gene mutation. *International Journal of Cardiology*, *216*, 188–189. https://doi.org/10.1016/j.ijcard.2016.04.164

Kizer, J. S., Muth, E., & Jacobowitz, D. M. (1976). The effect of bilateral lesions of the ventral noradrenergic bundle on endocrine-induced changes of tyrosine hydroxylase in the rat median eminence. *Endocrinology*, *98*(4), 886–893. https://doi.org/10.1210/endo-98-4-886

Kong, F., Zhu, J., Wu, J., Peng, J., Wang, Y., Wang, Q., Fu, S., Yuan, L.-L., & Li, T. (n.d.). *dbCRID: a database of chromosomal rearrangements in human diseases*. https://doi.org/10.1093/nar/gkq1038

Koshy, R., Ranawat, A., & Scaria, V. (2017). al mena: a comprehensive resource of human genetic variants integrating genomes and exomes from Arab, Middle Eastern and North African populations. *Journal of Human Genetics*, *62*, 889–894. https://doi.org/10.1038/jhg.2017.67

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/NAR/GKX1153

Leinonen, R., Nardone, F., Zhu, W., & Apweiler, R. (2006). UniSave: the UniProtKB Sequence/Annotation Version database. *Bioinformatics*, *22*(10), 1284–1285. https://doi.org/10.1093/BIOINFORMATICS/BTL105

León Palacio, A. (2019). *SILE: A Method for the Efficient Management of Smart Genomic Information*. https://doi.org/10.4995/THESIS/10251/131698

Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., Tsimberidou, A. M., Vnencak-Jones, C. L., Wolff, D. J., Younes, A., & Nikiforova, M. N. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *The Journal of Molecular Diagnostics : JMD*, *19*(1), 4. https://doi.org/10.1016/J.JMOLDX.2016.10.002

Lichten, M., & Goldman, A. S. H. (2003). MEIOTIC RECOMBINATION HOTSPOTS. *Https://Doi.Org/10.1146/Annurev.Ge.29.120195.002231*, *29*, 423–444. https://doi.org/10.1146/ANNUREV.GE.29.120195.002231

López-López, D., Roldán, G., Fernández-Rueda, J. L., Bostelmann, G., Carmona, R., Aquino, V., Perez-Florido, J., Ortuño, F., Pita, G., Núñez-Torres, R., González-Neira, A., Group, C., Peña-Chilet, M., & Dopazo, J. (2023). *A crowdsourcing database for the copy-number variation of the Spanish population*. https://doi.org/10.1186/s40246-023-00466-8

McEntyre, J., & Ostell, J. (2002). *The NCBI Handbook*. https://www.ncbi.nlm.nih.gov/books/NBK21101/

Medeiros-Domingo, A., Iturralde-Torres, P., & Ackerman, M. J. (2007). Clínica y genética en el síndrome de QT largo. *Revista Española de Cardiología*, *60*(7), 739–752. https://doi.org/10.1157/13108280

Medicina de precisión. (n.d.). Retrieved June 27, 2023, from https://www.genome.gov/es/genetics-glossary/Precision-Medicine

Meurs, K. M., Lacombe, V. A., Dryburgh, K., Fox, P. R., Reiser, P. R., & Kittleson, M. D. (2006). Differential expression of the cardiac ryanodine receptor in normal and arrhythmogenic right ventricular cardiomyopathy canine hearts. *Hum Genet*, *120*, 111–118. https://doi.org/10.1007/s00439-006-0193-2

Miocardiopatía arritmogénica - Fundación Española del Corazón. (n.d.). Retrieved June 28, 2023, from https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/cardiopatias-familiares-y-genetica/miocardiopatias/miocardiopatia-arritmogenica.html

Miocardiopatía dilatada - Fundación Española del Corazón. (n.d.). Retrieved June 28, 2023, from https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/cardiopatias-familiares-y-genetica/miocardiopatias/miocardiopatia-dilatada.html

Miocardiopatía hipertrófica - Fundación Española del Corazón. (n.d.). Retrieved June 28, 2023, from https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/cardiopatias-familiares-y-genetica/miocardiopatias/miocardiopatia-hipertrofica.html

Miocardiopatía no compactada - Fundación Española del Corazón. (n.d.). Retrieved July 1, 2023, from https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/cardiopatias-familiares-y-genetica/miocardiopatias/miocardiopatia-no-compactada.html

Miocardiopatía restrictiva - Fundación Española del Corazón. (n.d.). Retrieved July 1, 2023, from https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/cardiopatias-familiares-y-genetica/miocardiopatias/miocardiopatia-restrictiva.html

Mutágeno. (n.d.). Retrieved July 1, 2023, from https://www.genome.gov/es/genetics-glossary/Mutageno

Nachtegael, C., Gravel, B., Dillen, A., Smits, G., Nowé, A., Papadimitriou, S., & Lenaerts, T. (2022). Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database. *Database*, *2022*, 1–15. https://doi.org/10.1093/DATABASE/BAAC023

NAR Catalogs and Product Brochures | North American Rescue. (n.d.). Retrieved July 1, 2023, from https://www.narescue.com/nar-catalogs

Nucleótido. (n.d.). Retrieved June 27, 2023, from https://www.genome.gov/es/genetics-glossary/Nucleotido

Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y. Y., Kobayashi, Y., Patil, N., Thusberg, J., Westbrook, M., & Topper, S. (2017). Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. *Genetics in Medicine 2017 19:10*, *19*(10), 1105–1117. https://doi.org/10.1038/gim.2017.37

Oliva, P., Moreno A, R., Toledo, M. I., Montecinos, A., & Molina, J. (2006). Síndrome de Marfán. *Rev Méd Chile*, *134*, 1455–1464.

Olivé, A. (2007). Conceptual modeling of information systems. *Conceptual Modeling of Information Systems*, 1–455. https://doi.org/10.1007/978-3-540-39390-0/COVER

Ontologías y vocabularios controlados - GNOSS. (n.d.). Retrieved July 2, 2023, from https://www.gnoss.com/ontologias-vocabularios

Palacio, A. L., López, Ó. P., & Ródenas, J. C. C. (2018). A method to identify relevant genome data: Conceptual modeling for the medicine of precision. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11157 LNCS*, 597–609. https://doi.org/10.1007/978-3-030-00847-5_44/FIGURES/4

Paul, P., Nag, D., & Chakraborty, S. (2016). Recombination hotspots: Models and tools for detection. *DNA Repair*, *40*, 47–56. https://doi.org/10.1016/J.DNAREP.2016.02.005

Primaria, A., Javier, F., Zurián, V., Martín Gutiérrez, V., Sorlí, J. V, Castillo, M. M., Doménech, I. E., Ortiz Uriarte, R., & García Ribes, M. (n.d.). *Síndrome de Marfan Marfan's syndrome*. https://doi.org/10.1016/j.aprim.2008.07.015

Priori, S. G., & Napolitano, C. (2005). Cardiac and skeletal muscle disorders caused by mutations in the intracellular Ca2+ release channels. *Journal of Clinical Investigation*, *115*(8), 2033–2038. https://doi.org/10.1172/JCI25664

Recombinación homóloga. (n.d.). Retrieved July 1, 2023, from https://www.genome.gov/es/genetics-glossary/Recombinacion-homologa

Reyes Román, J. F., Pastor, Ó., Casamayor, J. C., & Valverde, F. (2016). Applying conceptual modeling to better understand the human genome. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9974 LNCS*, 404–412. https://doi.org/10.1007/978-3-319-46397-1_31/COVER

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & 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*. https://doi.org/10.1038/gim.2015.30

Rigden, D. J., Xos´, X., Fernández, X. M., & Fernández, F. (2023). The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Research*, *51*. https://doi.org/10.1093/nar/gkac1186

Rodrigues, A. S. L. (2013). Hotspots. *Encyclopedia of Biodiversity: Second Edition*, 127–136. https://doi.org/10.1016/B978-0-12-384719-5.00410-X

Rogozin, I. B., & 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. https://doi.org/10.1016/S1383-5742(03)00032-2

Rosenberg, H., Pollock, N., Schiemann, A., Bulger, T., & Stowell, K. (2015). Malignant hyperthermia: a review. *Orphanet Journal of Rare Diseases*, *10*(1), 1–19. https://doi.org/10.1186/S13023-015-0310-1

Secuencia de referencia del genoma humano. (n.d.). Retrieved June 28, 2023, from https://www.genome.gov/es/genetics-glossary/Human-Genome-Reference-Sequence

Shepherd, S., Ellis, F., Halsall, J., Hopkins, P., & Robinson, R. (2004). RYR1 mutations in UK central core disease patients: more than just the C-terminal transmembrane region of the RYR1 gene. *Journal of Medical Genetics*, *41*(3). https://doi.org/10.1136/JMG.2003.014274

Síndrome de Brugada – Cardiopatias Familiares. (n.d.). Retrieved July 1, 2023, from https://cardiopatiasfamiliares.es/sindrome-de-brugada/

Síndrome de Loeys-Dietz - Stanford Medicine Children's Health. (n.d.). Retrieved July 1, 2023, from https://www.stanfordchildrens.org/es/service/cardiovascular-connective-tissue/loeys-dietz-syndrome

Síndrome de Marfan - Fundación Española del Corazón. (n.d.). Retrieved July 1, 2023, from https://fundaciondelcorazon.com/informacion-para-pacientes/enfermedades-cardiovasculares/sindrome-de-marfan.html

Síndrome de QT largo | The Texas Heart Institute. (n.d.). Retrieved July 1, 2023, from https://www.texasheart.org/heart-health/heart-information-center/topics/sindrome-de-qt-largo/

Síndrome de QT largo congénito: Revisión de la literatura. (n.d.). *Médicas UIS*.

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., … Harris, L. W. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, *51*(D1), D977–D985. https://doi.org/10.1093/NAR/GKAC1010

SPDI - NCBI Variation Notation for Variants with Known Breakpoints. (n.d.). Retrieved July 2, 2023, from https://www.ncbi.nlm.nih.gov/variation/notation/

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Iny Stein, T., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, *54*(1), 1.30.1-1.30.33. https://doi.org/10.1002/CPBI.5

Trevino, V. (2000). *HotSpotAnnotations-a database for hotspot mutations and annotations in cancer*. *81*(3000). https://doi.org/10.1093/database/baaa025

Trueba-Gómez, R., & Estrada-Lorenzo, J. M. (2010). La base de datos PubMed y la búsqueda de información científica. *Seminarios de La Fundación Española de Reumatología*, *11*(2), 49–63. https://doi.org/10.1016/J.SEMREU.2010.02.005

Universitat, A., & València, P. De. (2021). *Proyectos estratégicos en cooperación*.

Vasilevsky, N. A., Roncaglia, P., & Ross, J. E. (n.d.). *Mondo: Unifying diseases for the world, by the world*. https://doi.org/10.1101/2022.04.13.22273750

Wieringa, R. J. (2014). Design Science Methodology for Information Systems and Software Engineering. *Design Science Methodology: For Information Systems and Software Engineering*, 1–332. https://doi.org/10.1007/978-3-662-43839-8

Witherspoon, J. W., & Meilleur, K. G. (n.d.). *Review of RyR1 pathway and associated pathomechanisms*. https://doi.org/10.1186/s40478-016-0392-6

Wleklinski, M. J., Kannankeril, P. J., & Knollmann, B. C. (2020). Molecular and tissue mechanisms of catecholaminergic polymorphic ventricular tachycardia. *The Journal of Physiology*, *598*(14), 2817–2834. https://doi.org/10.1113/JP276757

Yao, J. V., & Winship, I. (2020). More than meets the eye: Palmoplantar keratoderma and arrhythmogenic right ventricular cardiomyopathy in a patient with loss of the DSP gene. *JAAD Case Reports*, *6*(9), 804–806. https://doi.org/10.1016/J.JDCR.2020.06.025

Zahn-Zabal, M., Michel, P. A., Gateau, A., Nikitin, F., Schaeffer, M., Audot, E., Gaudet, P., Duek, P. D., Teixeira, D., De Laval, V. R., Samarasinghe, K., Bairoch, A., & Lane, L. (2020). The neXtProt knowledgebase in 2020: Data, tools and usability improvements. *Nucleic Acids Research*, *48*(D1), D328–D334. https://doi.org/10.1093/NAR/GKZ995

Zentrum für Humangenetik - Experten für genetische Diagnostik. (n.d.). Retrieved July 2, 2023, from https://www.humangenetik-tuebingen.de/

PRESUPUESTO:

Diseño y desarrollo de una fuente de datos sobre hotspots asociados al criterio PM1 de las guías ACMG – AMP del 2015 aplicado a cardiopatías familiares

Documento II

Alumna: Alba García Zarzoso

Tutora: Mireia Costa Sánchez

Cotutor: Óscar Pastor López

Grado en Ingeniería Biomédica

Curso 2022 - 2023

CAPÍTULO 1. PRESUPUESTO

1. OBJETIVOS DEL PRESUPUESTO

El propósito del presupuesto es la determinación del coste económico del proyecto llevado a cabo en este Trabajo de Final de Grado. Para ello, se debe calcular el costo total del proyecto teniendo en cuenta los gastos de personal, los costos de software y hardware utilizados.

Para calcular el costo atribuible – amortizaciones – al software y al hardware, se utilizará la siguiente formula:

$$Coste\ imputable\ (\sin IVA) = t \cdot \frac{C}{T}$$

Ecuación 1:Fórmula asociada al coste imputable

Donde "t" representa el tipo de uso en meses, "C" es el costo del equipo o la licencia de software y "T" es el tiempo de amortización en meses. A continuación, se presenta un desglose del presupuesto, que divide los costos por personal, software y hardware

2. PRESUPUESTO DESGLOSADO

En esta sección se presenta el desglose detallado del presupuesto asociado a este trabajo. Para ello, se muestran tres tablas que resumen los costos relacionados al personal (),al software () y al hardware.

2.1. Costes de personal

Para calcular los costos de personal se toman en cuenta las horas dedicadas por el ingeniero encargado del trabajo, las horas empleadas por los tutores y otros miembros de PROS que han asesorado a lo largo del proyecto, así como el costo unitario por hora de cada profesional involucrado.

Los costos unitarios por hora se determinaron en base a los criterios de elaboración de presupuestos de I-D publicados en 2018 por la Universidad Politécnica de Valencia que establecen recomendaciones para los costos de personal según su categoría profesional.

Según estos criterios, se establece un costo unitario de 24,99 €/hora para el ingeniero biomédico. En el caso de los tutores y otros miembros del grupo PROS, se establecen costos de 31,45 €/hora para un doctor contratado, 51,40 €/hora, para un catedrático de Universidad y 24,99 €/hora para el investigador. Con esta información, y considerando el número de horas invertidas al proyecto por cada participante, se obtienen los resultados de Tabla 3.

Tabla 3: Costes de personal

COSTES DE PERSONAL			
Personal	**Coste Unitario**	**Horas Totales**	**Coste total**
Ingeniero Biomédico: Titulado Superior	24,99 €/hora	300 h	7.497 €
Tutor: Investigador Predoctoral	24,99 €/hora	80 h	1.999,2 €
Cotutor: Catedrático de Universidad	51,40 €/hora	45h	2.056 €
Miembro del grupo PROS: Investigador Predoctoral	24,99 €/hora	10h	249,9 €
Coste total de personal		11.802,1 €	

Según se muestra en la Tabla 3, los costes totales de personal asociados a este Trabajo Fin de Grado ascienden a un total de once mil ochocientos y dos € con diez céntimos (11.802,1€).

2.2. Costes de software

Para el coste de software se va a emplear la fórmula del coste imputable del apartado 1 de este documento. En este Trabajo de Final de Grado se han usado diferentes programas de software: Microsoft Windows 10 Home, Microsoft Office Hogar y Estudiantes 2019, diferentes bases de datos genómicas, la plataforma RStudio y VisualStudioCode para realizar códigos de R y Python respectivamente. Todos los datos empleados para el cálculo de los costes imputables a cada software, así como los costes totales de software se resumen en la Tabla 4.

Tabla 4: Costes de software

COSTE DE SOFTWARE					
Programa	**Coste de Licencia**	**Número de Licencias**	**Periodo de Uso**	**Duración de la Licencia**	**Coste Imputable**
Sistema Operativo Microsoft Windows 10	145 €	1	6 meses	Indefinida (uso 3 años)	24,17 €
Microsoft Office Hogar y Estudiantes	149 €	1	6 meses	Indefinida (uso 3 años)	24,83 €
Bases de datos	0 €	Acceso libre	6 meses	-	0 €
RStudio	0 €	1	6 meses	Indefinida (uso 3 años)	0 €
VisualStudioCode	0 €	1	6 meses	Indefinida (uso 3 años)	0 €
Coste Total de Software				49 €	

Según se muestra en la Tabla 4, los costes totales de software asociados a este Trabajo Final de Grado ascienden a un total de cuarenta y nueve € (49 €).

2.3. Costes de hardware

Finalmente, se han determinado los costes imputables de hardware siguiendo la fórmula del apartado 1 del presupuesto. Más concretamente, para este proyecto el hardware utilizado ha sido un ordenador portátil con un procesador Intel Core i7 de octava generación y 16 GB de RAM cuyo periodo de vida útil o periodo de amortización se ha determinado en 3 años. Todos los datos empleados para el cálculo de los costes imputables a cada software, así como los costes totales de hardware se resumen en la Tabla 5.

Tabla 5: Costes de hardware

COSTES DE HARDWARE					
Equipo	**Coste Unitario**	**Unidades**	**Periodo de amortización**	**Periodo de uso**	**Coste imputable**
Ordenador portátil: Asus Vivobook 15 X512	900 €	1	3 años	6 meses	150 €
Coste total de Hardware			150 €		

Según se muestra en la Tabla 5,los costes totales de hardware asociados a este Trabajo Fin de Máster ascienden a un total de ciento cincuenta y ocho € con treinta y tres céntimos (158,33 €). En la siguiente sección se va a presentar, teniendo en cuenta los costes presentados, el presupuesto total necesario para el desarrollo de este Trabajo Fin de Máster.

3. PRESUPUESTO TOTAL

En primer lugar, se debe calcular el Presupuesto por Ejecución Material correspondiéndose este a la suma de los costes de personal (apartado 2.1), costes de software (apartado 2.2) y costes de hardware (apartado 2.3). en base a esto se calculan los gastos generales y el beneficio industrial del Proyecto.

Los gastos generales son el 13% del Presupuesto de Ejecución Material, y se corresponde con el capital necesario para llevar a cabo tareas en la empresa no relacionadas con el producto directamente y que suponen más beneficios de la empresa. El beneficio industrial supone un 6% del presupuesto por Ejecución Material y son los beneficios reales que la empresa consigue por el producto.

La suma del valor del Presupuesto por Ejecución Material, los gastos generales y el beneficio industrial dan lugar al conocido Presupuesto de ejecución por Contrata. Sumándole a este el IVA (21%) se obtendrá el presupuesto neto total necesario para este Trabajo de Final de Grado . los resultados de este presupuesto total se recogen en la Tabla 6.

Tabla 6: Presupuesto total

COSTE DEL PRESUPUESTO TOTAL	
Coste total de personal	11.802,1 €
Coste total de software	49,00 €
Coste total de hardware	150,00 €
TOTAL, PRESUPUESTO DE EJECUCIÓN MATERIAL	12.001,1 €
Gastos generales (13%)	1.560,14 €
Beneficio industrial (6%)	720,07 €
TOTAL PRESUPUESTO DE EJECUCIÓN POR CONTRATA	14.281,31 €
IVA (21%)	2.999,08 €
PRESUPUESTO TOTAL	17.280,39 €

ANEXOS:

Diseño y desarrollo de una fuente de datos sobre hotspots asociados al criterio PM1 de las guías ACMG – AMP del 2015 aplicado a cardiopatías familiares

Documento III

Alumna: Alba García Zarzoso

Tutora: Mireia Costa Sánchez

Cotutor: Óscar Pastor López

Grado en Ingeniería Biomédica

Curso 2022 - 2023

CAPÍTULO 1. ANEXOS

1. LISTA DE GENES DE INTERÉS DE CARDIOPATÍAS

- ADSL
- AFG3L2
- AGL
- AIFM1
- AKAP9
- AKT1
- A2ML1
- AARS
- ABAT
- ACADM
- ACADS
- ACADSB
- ACADVL
- ACAT1
- ACO2
- ACTN2
- ANK2
- ANKRD1
- ANO5
- APOA1
- ATAD3A
- ATP5A1
- AUH
- BAG3
- ALG6
- ALMS1
- BCS1L
- BRAF
- BSCL2
- BTD
- C10orf2
- C12orf65
- CACNA1C
- CACNA1D
- CACNB2
- CALM1
- CALM2
- CALR
- CALR3
- CASZ1
- CAV3
- CDH2
- COQ2
- COQ5
- COQ6
- COX10
- COX14
- COX15
- COX6B1
- CEP89
- CHRM2
- CNBP
- COG4
- COG5
- COG6
- DLAT
- DLD
- DMD
- DMPK
- CPT1A
- CPT2
- CRYAB
- CSRP3
- CYCS
- D2HGDH
- DHDDS
- DSC2
- DSG2
- DSP
- DTNA
- FAH
- ECHS1
- ELAC2
- EMD
- ETFA
- ETFB
- DNAJB6
- DNAJC19
- DNM1L
- DPAGT1
- DPYD
- 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
- HRAS
- GRHPR
- GSK3B
- GUSB
- GYG1
- HADH
- HADHA
- HADHB
- HEXA
- HFE
- HIBCH
- HLCS
- HMGCS2
- HSD17B10
- HTRA2
- HTT
- IARS2
- IDH2
- IDH3B
- IDS
- IDUA
- ILK
- KCNQ1
- LZTR1
- L2HGDH
- LAMP2
- LDB3
- LETM1
- LMNA
- JPH2
- JUP
- KCNA5
- KCND3
- KCNE1
- KCNE2
- KCNE3
- KCNH2
- KCNJ2
- KCNJ5
- KCNJ8
- MTFMT
- MAP2K1
- MAP2K2
- MEF2C
- MFF
- MFN2
- MGAT2
- MGME1
- MIB1
- MPC1
- MPDU1
- MPV17
- MRPS22
- MYLK2
- MYO6
- MYOM1
- MYOT

- MYPN
- NARS2
- OPA1
- OPA3
- NDUFA2
- NDUFAF2
- NDUFAF5
- NDUFAF6
- NDUFV1
- NDUFV2
- NEBL
- NEXN
- NF1
- NKX2-5
- NOS1AP
- NPC1
- NPC2
- NRAS
- NUBPL
- MYBPC3
- MYL2
- MYL3
- PKD2
- PKP2
- PANK2
- PC
- PCK1
- PDHA1
- PDHB
- PDHX
- PGM1
- PHKA1
- PHYH
- PRDM16
- PRKAG2
- PRKCSH
- PSEN1
- PSEN2
- PTPN11
- PTRF
- PUS1
- PLN
- PMM2
- PNPT1
- SACS
- RAF1
- RIT1
- RRM2B
- RYR1
- RYR2
- SEMA3A
- SERAC1
- SLC6A8
- SGCA
- SGCB
- SGCD
- SGCG
- SHOC2
- SLC19A3
- SLC22A5
- SLC25A10
- SLC25A12
- SLC25A24
- SLC25A3
- SLC25A38
- SLC25A4
- SLC35D1
- SCN10A
- SCN1B
- SCN3B
- SCN5A
- SCO1
- SCO2
- SDHA
- SDHAF1
- SDHAF2
- SDHB
- SDHC
- SDHD
- SEC23B
- SOS1
- SOS2
- SRD5A3
- TRMU
- COX7B
- ST3GAL3
- WRN
- XK
- YARS
- AARS2
- APTX
- DGUOK
- ETHE1
- GATAD1
- KCNE5
- STAT2
- STT3A
- SUCLA2
- SUCLG1
- MLYCD
- NADK2
- PDK3
- SLC25A19
- MURC
- NDUFA13
- NDUFA4
- NDUFA6
- NDUFA8
- NDUFB1
- NDUFB6
- NDUFC2
- NDUFV3
- SURF1
- SYNE1
- SYNE2
- NDUFB9
- TACO1
- TAZ
- TNNI3K
- TBX5
- TCAP
- ALG11
- CA5A
- MTO1
- ACAD9
- AGK
- COQ4
- TGFB3
- NGLY1
- MYOZ2
- POLG2
- TK2
- ATP5E
- ATPAF2
- COA5
- DNA2
- SLC25A1
- KARS
- DPM2
- EARS2
- ISCU
- LRPPRC
- MRPL3
- NDUFA1
- NDUFA10
- NDUFA11
- NDUFA12
- NDUFA7
- NDUFA9
- NDUFAB1
- NDUFAF1
- NDUFAF3
- NDUFAF4
- NDUFB3
- NFU1
- ALG8
- DPM1
- TMEM43
- TMEM70
- TMPO
- RARS2
- RMND1
- SLC35A2
- TPK1
- UQCRB
- UQCRQ
- AGXT
- BOLA3
- DOLK
- HCCS
- CARS2
- TNNC1
- TNNI3
- TNNT2
- MRPL44
- 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
- TTN
- TTR
- LIAS
- LYRM4
- MOGS
- MPI
- MRPS16
- PCK2
- PDP1
- SFXN4
- SLC25A32
- SLC35C1
- TARS2
- TMEM165
- VARS2
- COX4I1
- COX7A1
- COX7A2
- TUFM
- TUSC3
- TXN2
- TXNRD2
- TYMP
- PGM3
- TRNT1
- HARS2
- UPB1
- LARS2
- ATP6V0A2
- UQCRC2
- UQCRH
- VCL
- SMPD1
- SNTA1
- LONP1
- AGPAT2
- ABCC6
- ABCC9
- ACTA1
- ACTC1
- ATP6AP2
- CASQ2
- CBL
- COL7A1
- DES
- ETFDH
- EYA4
- GDAP1
- GPD1L
- HCN4
- HMGCL
- GLA
- GLB1
- HSPD1
- KRAS
- LAMA2
- LAMA4
- MAN1B1
- NDUFS1
- NDUFS2
- NDUFS3
- NDUFS4
- NDUFS7
- NDUFS8
- MYH6
- MYH7
- OXCT1
- PDSS2
- PPOX
- POLG
- SAMHD1
- RBM20
- SCN4B
- SPG7
- KLF10
- DARS2
- MARS2
- TBX20
- XPNPEP3
- PDLIM3
- NDUFS6
- NAA10
- TSFM
- PDSS1
- RFT1
- RPIA
- TIMM8A

2. CÓDIGO DE R QUE TRANSFORMA POSICIÓN PROTEICA EN POSICIÓN GENÓMICA

```
install.packages("BiocManager")
BiocManager::install("ensembldb")
BiocManager::install("IRanges")
BiocManager::install("EnsDb.Hsapiens.v75")
library(ensembldb)
library(IRanges)
library (EnsDb.Hsapiens.v75)

edbx<- EnsDb.Hsapiens.v75
gene_name <- "RYR1"
start_aa <- 3916
end_aa <-4942

protein_ensdbid <- proteins(edbx, filter = GeneNameFilter(gene_name))$protein_id
prt_position <- IRanges(start= start_aa, end = end_aa, names=protein_ensdbid)

gnm_position <- proteinToGenome (prt_position,edbx)
start_gnm = gnm_position[[protein_ensdbid]]@ranges@start
end_gnm = start_gnm + gnm_position[[protein_ensdbid]]@ranges@width-1
```

3. FUENTE DE DATOS

CRC	LOCAL	GEN	UNIPF	INIC	FIN	REF	REFER	UNIPROTKB'S ES	INICIO GR	FIN GRC	MÉTODO O FUENTE ESTANDARIZADA	FENOTIPO ESTANDARIZADO	DOI	COMENTARIOS
chr1	1q22	LMNA	P02545	482	482	Arg	Arg	ENSP00000355292	156106775	156106777	author statement without traceable sup	familial partial lipodystrophy Dunnigan type (MONDO:0007906	DOI:10.1016/j.diabet.2018.09.006	
chr1	1q21.2	RIT1	Q92963	57	57	Ala	Ala	ENSP00000357306	155874588	155874590	polymerase chain reaction assay (OBI_0	Noonan syndrome (MONDO:0018997)	DOI:10.1038/gim.2016.32	
chr1	1q21.2	RIT1	Q92963	82	82	Phe	Phe	ENSP00000357306	155874285	155874287	polymerase chain reaction assay (OBI_0	Noonan syndrome (MONDO:0018997)	DOI:10.1038/gim.2016.32	
chr1	1q21.2	RIT1	Q92963	95	95	Gly	Gly	ENSP00000357306	155874246	155874248	polymerase chain reaction assay (OBI_0	Noonan syndrome (MONDO:0018997)	DOI:10.1038/gim.2016.32	
chr1	1q43	RYR2	Q92736	3778	4201	Leu	Gln	ENSP00000355533	237923082	237947615	author statement without traceable sup	catecholaminergic polymorphic ventricular tachycardia (MOND	DOI:10.1074/jbc.M116.756528	The source states that the functional impact of mutations locat
chr1	1q43	RYR2	Q92736	44	466	Asn	Pro	ENSP00000355533	237433878	237617796	imported information (ECO:0000311)|in	catecholaminergic polymorphic ventricular tachycardia (MOND	DOI:10.1016/j.hrthm.2021.07.061, DOI: 10.1161/CIRCRE	Test in catecholaminergic polymorphic ventricular tachycardia
chr1	1q43	RYR2	Q92736	77	466	Leu	Pro	ENSP00000355533	237494238	237617796	manual assertion (ECO:0000218)|author	catecholaminergic polymorphic ventricular tachycardia (MOND	DOI: 10.1161/CIRCRESAHA.110.226845|DOI:10.1242/jcs.	| The hotspot is located within the N-terminal region
chr1	1q43	RYR2	Q92736	2246	2534	Ser	Leu	ENSP00000355533	237798236	237813266	imported information (ECO:0000311) | i	catecholaminergic polymorphic ventricular tachycardia (MOND	DOI:10.1016/j.hrthm.2021.07.061, DOI: 10.1161/CIRCRE	Test in catecholaminergic polymorphic ventricular tachycardia
chr1	1q43	RYR2	Q92736	3778	4201	Leu	Gln	ENSP00000355533	237923082	237947615	imported information (ECO:0000311)|in	catecholaminergic polymorphic ventricular tachycardia (MOND	DOI:10.1016/j.hrthm.2021.07.061, DOI: 10.1161/CIRCRE	Test in catecholaminergic polymorphic ventricular tachycardia
chr1	1q43	RYR2	Q92736	4497	4959	Arg	Arg	ENSP00000355533	237954741	237995920	imported information (ECO:0000311) | a	catecholaminergic polymorphic ventricular tachycardia (MOND	DOI:10.1016/j.hrthm.2021.07.061, DOI: 10.1161/CIRCRE	Test in catecholaminergic polymorphic ventricular tachycardia
chr1	1q43	RYR2	Q92736	164	164	Pro	Pro	ENSP00000355533	237540649	237540651	polymerase chain reaction (OBI_000041	arrhythmogenic right ventricular cardiomyopathy (MONDO:001	DOI:10.1016/j.jmb.2013.08.015	The hotspot is P164S in the HS-LOOP of the RYR2 protein.
chr1	1q43	RYR2	Q92736	169	169	Arg	Arg	ENSP00000355533	237540664	237540666	polymerase chain reaction (OBI_000041	arrhythmogenic right ventricular cardiomyopathy (MONDO:001	DOI:10.1016/j.jmb.2013.08.015	The hotspot is R169Q in the HS-LOOP of the RYR2 protein.
chr1	1q43	RYR2	Q92736	176	176	Arg	Arg	ENSP00000355533	237540685	237540687	polymerase chain reaction (OBI_000041	arrhythmogenic right ventricular cardiomyopathy (MONDO:001	DOI:10.1016/j.jmb.2013.08.015	The hotspot is R176Q in the HS-LOOP of the RYR2 protein.
chr1	1q32	TNNT2	P45379	92	92	Arg	Arg	ENSP00000236918	201334741	201334743	RNA sequencing assay (OBI:0001177), au	Ventricular hypertrophy (HP:0001714)|Atrial fibrillation (HP:00	DOI:10.1016/j.yexcr.2019.111736|DOI:10.3389/fphys.2022.864547	
chr1	1q32	TNNT2	P45379	79	79	Ile	Ile	ENSP00000236918	201334780	201334782	DNA sequencing assay (OBI_0000626), a	Atrial fibrillation (HP:0005110), Hypertrophic cardiomyopathy (	DOI:10.3389/fphys.2022.864547	
chr1	1q32	TNNT2	P45379	110	110	Phe	Phe	ENSP00000236918	201334385	201334387	DNA sequencing assay (OBI_0000626), a	Atrial fibrillation (HP:0005110), Hypertrophic cardiomyopathy (	DOI:10.3389/fphys.2022.864547	
chr1	1q32	TNNT2	P45379	130	130	Arg	Arg	ENSP00000236918	201334325	201334327	DNA sequencing assay (OBI_0000626), a	Atrial fibrillation (HP:0005110), Hypertrophic cardiomyopathy (	DOI:10.3389/fphys.2022.864547	
chr1	1q32	TNNT2	P45379	278	278	Arg	Arg	ENSP00000236918	201328753	201328755	DNA sequencing assay (OBI_0000626), a	Atrial fibrillation (HP:0005110), Hypertrophic cardiomyopathy (	DOI:10.3389/fphys.2022.864547	
chr1	1q32	TNNT2	P45379	286	286	Arg	Arg	ENSP00000236918	201328362	201328364	DNA sequencing assay (OBI_0000626), a	Atrial fibrillation (HP:0005110), Hypertrophic cardiomyopathy (	DOI:10.3389/fphys.2022.864547	
chr10	10q25.3	RBM20	Q5T481	634	634	Arg	Arg	ENSP00000358532	112572055	112572057	DNA sequencing assay (OBI_0000626)	Primary dilated cardiomyopathy (MONDO:0005021)	DOI:10.1111/j.1752-8062.2010.00198.x	The hotspot is in the exon 9 of the RBM20 gene, from R634 to R
chr10	10q25.3	RBM20	Q5T481	638	638	Arg	Arg	ENSP00000358532	112572067	112572069	DNA sequencing assay (OBI_0000626)	Primary dilated cardiomyopathy (MONDO:0005021)	DOI:10.1111/j.1752-8062.2010.00198.x	The hotspot is in the exon 9 of the RBM20 gene, from R634 to R
chr10	10q25.3	RBM20	Q5T481	628	655	Tyr	Cys	ENSP00000358532	112572037	112572120	polymerase chain reaction (OBI_000041	familial dilated cardiomyopathy (MONDO:0016333)	DOI:10.1016/j.jacc.2009.05.038	The hotspot is located in the RS-domain of the RBM20 protein.
chr11	11p15.1	ANO5	Q75V66	61	95	Phe	Glu	ENSP00000315371	22242643	22242747	DNA sequencing assay (OBI_0000626)	autosomal recessive limb-girdle muscular dystrophy type 2L (M	DOI:10.1016/j.nmd.2015.03.011	The hotspot is the exon 5 of the ANO5 gene.
chr11	11p15.1	ANO5	Q75V66	746	805	Ala	Arg	ENSP00000315371	22296115	22297640	DNA sequencing assay (OBI_0000626)	autosomal recessive limb-girdle muscular dystrophy type 2L (M	DOI:10.1016/j.nmd.2015.03.011	The hotspot is the exon 20 of the ANO5 gene.
chr11	11q23.3	APOA1	P02647	50	93	Trp	Gln	ENSP00000236850	116707717	116707127	DNA sequencing assay (OBI_0000626)	AApoAI amyloidosis (MONDO:0019731)	DOI:10.2353/jmoldx.2009.080161	
chr11	11q23.3	APOA1	P02647	170	178	Leu	Leu	ENSP00000236850	116706794	116706820	DNA sequencing assay (OBI_0000626)	AApoAI amyloidosis (MONDO:0019731)	DOI:10.2353/jmoldx.2009.080161	
chr11	11p15.5	KCNQ1	P51787	341	341	Pro	Pro	ENSP00000155840	2604764	2604766	molecule detection assay evidence (EC	long QT syndrome (MONDO:0002442)	DOI:10.1161/01.cir.100.10.1077	
chr11	11p15.5	KCNQ1	P51787	344	344	Thr	Thr	ENSP00000155840	2604773	2604775	molecule detection assay evidence (EC	long QT syndrome (MONDO:0002442)	DOI:10.1161/01.cir.100.10.1077	
chr11	11p15.5	KCNQ1	P51787	160	202	Glu	Asp	ENSP00000155840	2591858	2592556	polymerase chain reaction (OBI_000041	long QT syndrome (MONDO:0002442)	DOI:10.1097/PAF.0000000000000411	The hotspot is located in exon 3 of the KCNQ1 gene.
chr11	11p15.5	KCNQ1	P51787	261	307	Glu	Val	ENSP00000155840	2594076	2594216	polymerase chain reaction (OBI_000041	long QT syndrome (MONDO:0002442)	DOI:10.1097/PAF.0000000000000411	The hotspot is located in exon 6 of the KCNQ1 gene.
chr11	11p11.2	MYBPC3	Q14896	847	847	Tyr	Tyr	ENSP00000442795	47359003	47359005	DNA sequencing assay (OBI_0000626)	Hypertrophic cardiomyopathy (MONDO:0005045)	DOI:10.20452/pamw.15130	Tested in Polish population
chr11	11p11.2	MYBPC3	Q14896	258	274	Glu	Thr	ENSP00000442795	47369975	47369231	polymerase chain reaction (OBI_000041	Hypertrophic cardiomyopathy (MONDO:0005045)	DOI:10.1080/ac.67.1.2146562	

chr11	11p11.2	MYBPC3	Q14896	274	284	Thr	Ser	ENSP00000442795	47369408	47369030	polymerase chain r	Hypertrophic cardio	DOI:10.1080/ac.67.1.2146562
chr11	11p11.2	MYBPC3	Q14896	364	408	Ala	Ser	ENSP00000442795	47367758	47364813	polymerase chain r	Hypertrophic cardio	DOI:10.1080/ac.67.1.2146562
chr12	12p13.33	CACNA1C	Q13936	851	860	Pro	Arg	ENSP00000266376	2702399	2702428	cryogenic electron	long qt syndrome 8 (	doi: 10.1093/europace/e	The hotspot is lo
chr12	12p13.33	CACNA1C	Q13936	518	518	Arg	Arg	ENSP00000266376	2675631	2675633	imported informati	Timothy syndrome (	doi: 10.1016/j.ijcard.201	The hotspot is ch
chr12	12p12.1	KCNJ8	Q15842	422	422	Ser	Ser	ENSP00000240662	21918666	21918668	DNA sequencing as	Brugada syndrome (	DOI:10.1016/j.hrthm.2011.10.035	
chr12	12p11	PKP2	Q99959	852	852	Asn	Asn	ENSP00000070846	32945599	32945601	polymerase chain r	arrhythmogenic righ	DOI:10.1016/j.amjcard.2	Tested in Chines
chr12	12q24.1	TBX5	Q99593	279	279	Arg	Arg	ENSP00000309913	1,15E+08	1,15E+08	DNA sequencing as	Holt-Oram syndrom	DOI:10.1590/S1415-47572010005000051	
chr14	14q11.2-q	MYH7	P12883	403	403	Arg	Arg	ENSP00000347507	23898486	23898488	polymerase chain r	Hypertrophic cardio	DOI:10.1016/j.scr.2021.102245	
chr14	14q11.2-q	MYH7	P12883	723	723	Arg	Arg	ENSP00000347507	23895021	23895023	DNA sequencing as	Hypertrophic cardio	DOI:10.1177/147323001003800308	
chr14	14q11.2-q	MYH7	P12883	894	974	Glu	Lys	ENSP00000347507	23893116	23893358	chain termination s	Hypertrophic cardio	DOI: 10.1002/jcla.22303	The hotspot is ex
chr14	14q11.2-q	MYH7	P12883	380	418	Glu	Val	ENSP00000347507	23898984	23898556	polymerase chain r	Hypertrophic cardio	DOI:10.1080/ac.67.1.2146562	
chr14	14q11.2-q	MYH7	P12883	894	974	Glu	Lys	ENSP00000347507	23893116	23893358	polymerase chain r	Hypertrophic cardio	DOI:10.1080/ac.67.1.2146562	
chr15	15q14	ACTC1	P68032	1	43	Met	Gln	ENSP00000290378	35086881	35087009	polymerase chain r	congenital heart dis	DOI:10.3390/ijerph19165	The hotspot is ex
chr15	15q34	NKX2-5	P52952	1	112	Met	Glu	ENSP00000327758	1,73E+08	1,73E+08	polymerase chain r	congenital heart dis	DOI:10.3390/ijerph19165	The hotspot is ex
chr15	15q34	NKX2-5	P52952	112	324	Glu	Trp	ENSP00000327758	1,73E+08	1,73E+08	polymerase chain r	congenital heart dis	DOI:10.3390/ijerph19165	The hotspot is ex
chr16	16p13.11	ABCC6	O95255	378	378	Gln	Gln	ENSP00000205557	16295900	16295902	polymerase chain r	autosomal recessive	DOI:10.1111/j.1752-8062	The mutation Q3
chr16	16p13.11	ABCC6	O95255	1339	1339	Arg	Arg	ENSP00000205557	16248754	16248756	polymerase chain r	autosomal recessive	DOI:10.1111/j.1752-8062	The mutation R1
chr16	16q22.1	COG8	Q96MW5	528	612	Gly	Pro	ENSP00000305459	69366617	69364998	chain termination s	congenital disorder	DOI:10.1002/ajmg.a.610	The exon 5 of th
chr17	17q21	JUP	P14923	675	675	His	His	ENSP00000377508	39913688	39913690	imported informati	arrhythmogenic righ	DOI:10.1136/jmedgenet	The hotspot is in
chr17	17q24.3	KCNJ2	P63252	218	218	Arg	Arg	ENSP00000243457	68171832	68171834	chain termination s	Andersen-Tawil syn	DOI:10.12688/f1000research.11610.1	
chr19	19q13.1	RYR1	P21817	35	614	Cys	Arg	ENSP00000352608	38931442	38948187	imported informati	RYR1-related myopa	DOI:10.1007/s13311-018	The hotspot is th
chr19	19q13.1	RYR1	P21817	4550	4940	Val	Ala	ENSP00000352608	39058546	39076594	imported informati	RYR1-related myopa	DOI:10.1007/s13311-018	The hotspot is th
chr19	19q13.1	RYR1	P21817	1	552	Met	Arg	ENSP00000352608	38924470	38946170	imported informati	malignant hyperthe	DOI:10.1038/s41436-021	The hotspot is th
chr19	19q13.1	RYR1	P21817	2101	2458	Met	Arg	ENSP00000352608	38985018	38991296	imported informati	malignant hyperthe	DOI:10.1038/s41436-021	The hotspot is th
chr19	19q13.1	RYR1	P21817	4631	4991	Tyr	Phe	ENSP00000352608	39062803	39077168	imported informati	malignant hyperthe	DOI:10.1038/s41436-021	The hotspot is th
chr19	19q13.1	RYR1	P21817	4550	4940	Val	Ala	ENSP00000352608	39058546	39076594	imported informati	central core myopat	DOI:10.1002/humu.2205	The hotspot is th
chr19	19q13.1	RYR1	P21817	1	614	Met	Arg	ENSP00000352608	38924470	38948187	imported informati	central core myopat	DOI:10.1213/ANE.00000	The hotspot is th
chr19	19q13.1	RYR1	P21817	2163	2458	Arg	Arg	ENSP00000352608	38985204	38991296	imported informati	central core myopat	DOI:10.1002/humu.2205	The hotspot is th
chr19	19q13.1	RYR1	P21817	4163	4973	Arg	Pro	ENSP00000352608	39051957	39076781	imported informati	central core myopat	DOI:10.1213/ANE.00000	The hotspot is th
chr19	19q13.1	RYR1	P21817	2355	2355	Arg	Arg	ENSP00000352608	38990310	38990312	polymerase chain r	malignant hyperthe	DOI:10.1213/ANE.0b013	Not all the hotsp
chr19	19q13.1	RYR1	P21817	2354	2354	Val	Val	ENSP00000352608	38990307	38990309	polymerase chain r	malignant hyperthe	DOI:10.1213/ANE.0b013	Not all the hotsp
chr19	19q13.1	RYR1	P21817	34	614	Leu	Arg	ENSP00000352608	38931439	38948187	polymerase chain r	malignant hyperthe	DOI:10.1177/0310057X08	The hotspot invo
chr19	19q13.1	RYR1	P21817	4136	4973	Arg	Pro	ENSP00000352608	39051876	39076781	polymerase chain r	malignant hyperthe	DOI:10.1177/0310057X08	The hotspots in

chr19	19q13.1	RYR1	P21817	4583	4666	Val	Lys	ENSP00000352608	39062659	3,9E+07	polymerase c	malignant hy	DOI:10.1213/a	The hotspot is located ir
chr19	19q13.1	RYR1	P21817	4789	4837	Ser	Gln	ENSP00000352608	39070622	3,9E+07	polymerase c	malignant hy	DOI:10.1213/a	The hotspot is located ir
chr19	19q13.1	RYR1	P21817	4838	4882	Leu	Thr	ENSP00000352608	39071010	3,9E+07	polymerase c	malignant hy	DOI:10.1213/a	The hotspot is located ir
chr19	19q13.1	RYR1	P21817	4883	4983	Cys	Ile	ENSP00000352608	39075583	3,9E+07	polymerase c	malignant hy	DOI:10.1213/a	The hotspot is located ir
chr19	19q13.1	RYR1	P21817	16	55	Asp	Gln	ENSP00000352608	38931385	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1213/	The hotspot is exon 2 of
chr19	19q13.1	RYR1	P21817	481	526	Gly	Ala	ENSP00000352608	38945875	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1213/	The hotspot is exon 14 c
chr19	19q13.1	RYR1	P21817	2222	2266	Glu	Gly	ENSP00000352608	38987049	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1213/	The hotspot is exon 41 c
chr19	19q13.1	RYR1	P21817	2343	2405	Gly	His	ENSP00000352608	38989883	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1213/	The hotspot is exon 44 c
chr19	19q13.1	RYR1	P21817	4095	4208	Ala	Gln	ENSP00000352608	39051753	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1213/	The hotspot is exon 90 c
chr19	19q13.1	RYR1	P21817	4209	4479	Val	Gly	ENSP00000352608	39055599	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1213/	The hotspot is exon 91 c
chr19	19q13.1	RYR1	P21817	4883	4935	Cys	Gly	ENSP00000352608	39075583	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1213/	The hotspot is exon 102
chr19	19q13.1	RYR1	P21817	4667	4710	Val	Pro	ENSP00000352608	39063817	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1093/	The hotspot is exon 95 c
chr19	19q13.1	RYR1	P21817	4935	4957	Gly	Thr	ENSP00000352608	39075739	3,9E+07	DNA sequenc	malignant hy	DOI: 10.1093/	The hotspot is exon 103
chr19	19q13.1	RYR1	P21817	3916	4942	Ile	Gly	ENSP00000352608	39034043	3,9E+07	imported infc	RYR1-related	DOI:10.3233/J	The hotspot is in the reg
chr2	2p21	SOS1	Q07889	266	266	Thr	Thr	ENSP00000384675	39278351	3,9E+07	DNA sequenc	Noonan synd	DOI:10.1038/s41431-020-00708-6	
chr2	2p21	SOS1	Q07889	269	269	Met	Met	ENSP00000384675	39278342	3,9E+07	DNA sequenc	Noonan synd	DOI:10.1038/s41431-020-00708-6	
chr2	2p21	SOS1	Q07889	378	378	Thr	Thr	ENSP00000384675	39251219	3,9E+07	DNA sequenc	Noonan synd	DOI:10.1038/s41431-020-00708-6	
chr2	2q31	TTN	Q8WZ42	467	512	Val	Gln	ENSP00000343764	179658131	1,8E+08	chain termina	Primary dilat	DOI:10.1093/(	The hotspot is exon 9 of
chr2	2q31	TTN	Q8WZ42	2580	2580	Thr	Thr	ENSP00000343764	179637951	1,8E+08	3D structure c	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr2	2q31	TTN	Q8WZ42	2848	2848	Pro	Pro	ENSP00000343764	179634884	1,8E+08	3D structure c	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr2	2q31	TTN	Q8WZ42	2897	2897	Thr	Thr	ENSP00000343764	179634617	1,8E+08	3D structure c	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr2	2q31	TTN	Q8WZ42	2923	2923	Ile	Ile	ENSP00000343764	179634539	1,8E+08	3D structure c	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr2	2q31	TTN	Q8WZ42	2946	2946	Val	Val	ENSP00000343764	179634470	1,8E+08	3D structure c	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr2	2q31	TTN	Q8WZ42	2948	2948	Tyr	Tyr	ENSP00000343764	179634464	1,8E+08	3D structure c	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr2	2q31	TTN	Q8WZ42	2951	2951	Ile	Ile	ENSP00000343764	179634455	1,8E+08	3D structure c	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr2	2q31	TTN	Q8WZ42	2996	2996	Ala	Ala	ENSP00000343764	179633575	1,8E+08	chain termina	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr2	2q31	TTN	Q8WZ42	2297	2297	Gly	Gly	ENSP00000343764	179639100	1,8E+08	3D structure c	arrhythmoge	DOI:10.1080/07391102.2020.1768148	
chr3	3q22.3	MRAS	O14807	23	23	Gly	Gly	ENSP00000289104	138091792	1,4E+08	DNA sequenc	Noonan synd	DOI:10.1093/hmg/ddz108	
chr3	3q22.3	MRAS	O14807	68	68	Thr	Thr	ENSP00000289104	138116174	1,4E+08	DNA sequenc	Noonan synd	DOI:10.1093/hmg/ddz108	
chr3	3p21	SCN5A	Q14524	1340	1352	Val	Gly	ENSP00000328968	38601827	3,9E+07	imported infc	Brugada synd	DOI: 10.1152/ajpheart.00061.2021	
chr3	3p21	SCN5A	Q14524	1605	2016	Gly	Val	ENSP00000328968	38595770	3,9E+07	polymerase c	long QT syndı	DOI:10.1097/I	The hotspot is located ir
chr5	5q31.2	MYOT	Q9UBF9	1	119	Met	Asn	ENSP00000391185	137206341	1,4E+08	polymerase c	myofibrillar r	DOI:10.1212/(	The hotspot is located ir
chr6	6p24.3	DSP	P15924	250	604	Gln	Asp	ENSP00000369129	7563990	7571726	cultured cell	arrhythmoge	DOI:10.1083/jcb.201312110	doi: 10.10

chr6	6p24.3	DSP	P15924	1029	1793	Leu	Glu	ENSP00000369129	7579508	7581802	imported inf	arrhythmoge	DOI:10.1136/j	The hotsot is
chr6	6p24.3	DSP	P15924	1794	2871	Ala	His	ENSP00000369129	7582875	7586108	imported inf	arrhythmoge	DOI:10.1136/j	The hotsot is
chr7	7q34	BRAF	P15056	468	468	Phe	Phe	ENST00000288602	140481404	140481404	chain termina	Noonan synd	doi: 10.1186/s	This is a poss
chr7	7q36.3	DNAJB6	O75190	93	93	Phe	Phe	ENSP00000262177	157160108	157160110	DNA sequenc	autosomal dc	doi: 10.1111/ene.13598	
chr7	7q36.1	KCNH2	Q12809	618	618	Thr	Thr	ENSP00000262186	150648627	150648629	DNA sequenc	short QT sync	DOI:10.1016/j.jacep.2016.11	
chr7	7q36.1	KCNH2	Q12809	588	588	Asn	Asn	ENSP00000262186	150648717	150648719	DNA sequenc	short QT sync	DOI:10.1016/j.jacep.2016.11	
chr7	7q36.1	KCNH2	Q12809	618	618	Thr	Thr	ENSP00000262186	150648627	150648629	imported inf	short QT sync	DOI:10.1016/j.bbrc.2022.01.	
chr7	7q36.1	KCNH2	Q12809	26	105	Ser	Cys	ENSP00000262186	150674926	150656824	DNA sequenc	long QT syndı	DOI:10.1016/j	The hotspot i
chr7	7q36.1	KCNH2	Q12809	427	501	Tyr	Asp	ENSP00000262186	150649567	150649791	polymerase c	long QT syndı	DOI:10.1097/I	The hotspot i
chr7	7q36.1	KCNH2	Q12809	41	70	Val	His	ENSP00000262186	150671896	150671985	imported inf	long QT syndı	DOI:10.1042/I	The hotspots
chr8	8p23.1-p	GATA4	P43694	292	292	Cys	Cys	ENSP00000334458	11607710	11607712	DNA sequenc	congenital he	DOI:10.1186/	The hotspot,
chr9	9p13.1	GNE	Q9Y223	39	39	Arg	Arg	ENSP00000414760	36249236	36249238	DNA sequenc	GNE myopath	DOI:10.1186/s13023-022-02	
chr9	9p13.1	GNE	Q9Y223	207	207	Asp	Asp	ENSP00000414760	36236977	36236979	DNA sequenc	GNE myopath	DOI:10.1186/s	|Tested in Ch

4. CÓDIGO DE PYTHON DE VALIDACIÓN DE RESULTADOS

```
import pandas as pd

import os

import io

from io import StringIO

path = os.path.join("C:/Users/algz6/Desktop/TFG/proyecto_hotspots/validacion_python/datos_validacion/cardio_la_fe/","OGM-03-046_TSVC_variants_IonXpress_014.vcf")

def read_vcf(path):

    with open(path,'r') as f:

        lines= [l for l in f if not l.startswith('##')]

        return pd.read_csv(

            io.StringIO(''.join(lines)),

            dtype={'#CHROM':str,'POS':str},

            sep='\t'

            ).rename(columns={'#CHROM':'CHROM'}

df_vcf=read_vcf(path)

path_exc = os.path.join("C:/Users/algz6/Desktop/TFG/proyecto_hotspots/versiones_excel/bd_hotspots/","hotspots_revisado.xlsx")

def read_excel(path_exc):

    return pd.read_excel(path_exc, usecols=['CHROMOSOME', 'START GRCh37 ', 'END GRCh37'])

# Leer el archivo de Excel

df_excel = read_excel(path_exc)

df_excel["START GRCh37 "] = df_excel["START GRCh37 "].astype(float)

df_excel["END GRCh37"] = df_excel["END GRCh37"].astype(float)

df_vcf["POS"] = df_vcf["POS"].astype(float)

data = {'Chromosome':[], 'Position':[], 'REF':[], 'ALT':[]}

variants_in_hotspot = pd.DataFrame(data)

for i in range(len(df_vcf)):

    hotspot_evaluation = (df_excel["CHROMOSOME"] == df_vcf.iloc[i,0]) & (df_vcf.iloc[i,1] >= df_excel["START GRCh37 "]) & (df_vcf.iloc[i,1] <= df_excel["END GRCh37"])

    if True in hotspot_evaluation.values :
```

```
variant = {'Chromosome':df_vcf.iloc[i,0], 'Position':df_vcf.iloc[i,1], 'REF':df_vcf.iloc[i,3], 'ALT':df_vcf.iloc[i,4]}

variants_in_hotspot.loc[len(variants_in_hotspot)]= variant

variants_in_hotspot.to_excel('variants_in_hotstpot_OGM-03-046_TSVC_variants_IonXpress_014.xlsx')
```

5. RESULTADOS DE LA EVALUACIÓN DEL CASO DE USO

PACIENTE	GEN	VARIACIÓN CAUSANTE DE DIAGNÓSTICO	VARIACIÓN EN HOTSPOT	COMENTARIOS
1	KCNH2	H562Y	NO	Variación cercana al hotspot
2	TMEM43	S358L	NO	-
3	LMNA	c.1270delAfsX56	NO	-
4	KCNH2	c.1694C>T	NO	-
5	TTN	F13907L	NO	.
6	MYBPC3	F1246S	NO	Variación VUS[14]
7	MYH7	A797T	NO	Exoma patológico
8	MYBPC3	E542Q	NO	Variación patológica
9	TTN	P173863del	NO	Variación VUS
10	MYBPC3	W1007*	NO	-
11	KCNJ2	G215D	NO	-
12	LMNA	R335W	NO	-
13	MYH7	R1781C	NO	Variación probablemente patológica
14	-	-	-	No realizado
15	-	-	-	No realizado
16	-	-	-	No realizado

[14] **VUS.** *Variant Uncertain Significance*, variante de significado incierto

17	KCNH2			W568*			NO			-	
18	-			-			-			NGS sin mutaciones	
19	FLNC	DSG2	SCN5A	-	-	-	-	-	-	Variaciones VUS	
20	MYH7		TNNT2	I114T		R286C	NO		NO	Variación VUS	Variación VPP
21	-			-			-			No realizado	
22	DSP			R84*			NO			-	
23	RyR2			A3757D			NO			No concluyente, variación VUS	
24	-			-			-			No realizado	
25	-			-			-			NGS sin mutaciones	
26	DMD			M1360V			NO			Exoma con variación VUS	
27	TNNT2			R286H			NO			-	
28	KCNQ1			R243C			NO			-	
29	KCNQ1			F339S			NO			-	
30	LMNA			R335W			NO			-	
31	BRAF			D324H			NO			Variación VUS	
32	MYH7		HFE	R869C		H63D	NO		NO	-	
33	MYH7			E1356K			NO			-	
34	DES			-			-			NGS sin nada, variación VUS	
35	-			-			-			Antecedentes familiares, pero no se hace pruebas	
36	-			-			-			Antecedentes familiares, pero no cuadra fenotipo encontrado.	
37	LMNA			F237S			NO			-	

38			-			-	-	No realizado
39	DSP	DSG2	TTN	R425*	G863R	G7360R	NO	Variación asociada a DSG2 es VUS
40			TTN			I32934T*fs48	NO	-
41			CALR3			-	NO	Variación VUS
42			TTN			-	NO	Variación VUS