Estudio de las variantes genéticas asociadas a la Miocardiopatía Dilatada Familiar

Lucía González Llorente
Máster de Bioinformática y Bioestadística
Bioinformática clínica

Nombre Director del Trabajo: Guerau Fernández Isern
Nombre Supervisor externo: Enrique Caso Peláez

08/01/2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	Estudio de las variantes genéticas asociadas a la Miocardiopatía Dilatada Familiar
Nombre del autor:	*Lucía González Llorente*
Nombre del consultor/a:	*Guerau Fernández Isern*
Nombre del PRA:	*Javier Luis Cánovas Izquierdo*
Fecha de entrega (mm/aaaa):	01/2020
Titulación:	*Máster universitario de Bioinformática y Bioestadística*
Área del Trabajo Final:	*Bioinformática clínica*
Idioma del trabajo:	*Castellano*
Palabras clave	Miocardiopatía Dilatada Familiar, Whole Exome Sequencing, predictores de patogenicidad.

Resumen del Trabajo (máximo 250 palabras):

La miocardiopatía dilatada aislada asintomática (MCD) se caracteriza por la dilatación y disfunción sistólica del ventrículo izquierdo, con la reducción de la fuerza de contracción del miocardio. Se considera que existe miocardiopatía dilatada familiar (MCDF) cuando dos o más miembros de una misma familia son diagnosticados de MCD idiopática.

Debido a la enorme heterogeneidad genética y fenotípica de MCDF, el conocimiento de las alteraciones moleculares de esta enfermedad es limitado, por lo que se necesita mejorar la estrategia actual de diagnóstico y tratamiento. Para ello, la utilización de las nuevas tecnologías, como la secuenciación de nueva generación (NGS) entre las que se incluye la secuenciación del exoma completo (WES), puede ayudar en la búsqueda nuevos marcadores moleculares y dianas terapéuticas.

El principal objetivo de este trabajo fue detectar los SNVs (*Single Nucleotide Variants*) que tienen una mayor relación con el desarrollo del MCDF. Para lograr este objetivo, se utilizaron diferentes herramientas bioinformáticas para el análisis de los datos obtenidos por WES de una

familia de 6 miembros, dos de los cuales están diagnosticados con cardiopatía familiar dilatada de origen idiopático.

Los resultados demuestran que el gen KCNJ12 juega un papel importante en la alteración de la estructura y conducción cardiaca, sugiriendo que la mutación P156L puede tener un papel patogénico en la miocardiopatía dilatada familiar. Mediante este trabajo se enfatiza la aplicación de WES en la identificación de mutaciones causales en esta enfermedad.

Abstract (in English, 250 words or less):

Nonsyndromic isolated Dilated Cardiomyopathy (DCM) is characterized by left ventricular enlargement and systolic dysfunction, a reduction in the myocardial force of contraction. When each of two or more closely related family members meet a formal diagnostic standard for idiopathic dilated cardiomyopathy diagnosis of familial dilated cardiomyopathy (FDCM) is made.

Due to the enormous genetic and phenotypic heterogenicity of FDCM, knowledge of the molecular alterations of this disease is limited, so it is necessary to improve the current strategy of diagnosis and treatment. To this end, the use of new technologies, such as new generation sequencing (NGS), including complete exome sequencing (WES), can help in the search for new molecular markers and therapeutic targets.

The main objective of this project was to detect the SNVs (Single Nucleotide Variants) that have a greater relationship with the development of the FDCM. To achieve this goal, different bioinformatics tools were used to analyze the data obtained by WES from a 6 members family, two of whom are diagnosed of idiopathic dilated cardiomyopathy.

Results demonstrates that KCNJ12 gene play an important role in the alteration of the structure and cardiac conduction, and suggest that the P156L mutation may have a pathogenic role in familial dilated cardiomyopathy. This study emphasizes the application of WES in the identification of causal mutations in this disease.

Índice

Lista de figuras

Lista de tablas

1. INTRODUCCIÓN

1.1 CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO

1.1.1. Descripción general

Para este trabajo se analizaron los datos de una familia de cinco miembros diagnosticados de miocardiopatía dilatada familiar (MCDF). Estos datos han sido obtenidos mediante la secuenciación del exoma completo (WES).

El posterior análisis bioinformático determinó la existencia de polimorfismos de nucleótido simple (SNPs), así como su capacidad predictora de la enfermedad mediante el uso de diferentes softwares.

Finalmente, se estudió la relación existente entre las variantes genéticas encontradas y las diferentes rutas funcionales de acuerdo a varias bases de datos, como Gene Ontology, KEGG u Open Targets Platform.

1.1.2. Justificación del TFM

La miocardiopatía dilatada aislada asintomática (MCD) se caracteriza por la dilatación y disfunción sistólica del ventrículo izquierdo, con la reducción de la fuerza de contracción del miocardio. Se ha descrito que la MD idiopática se presenta con una tasa anual de 5 a 8 casos por 100.000 personas [1]. Cuando se descartan todas las causas detectables excepto las genéticas, la MCD es denominada idiopática. Un estudio epidemiológico formal basado en la población, y realizado en el condado de Olmsted, Minnesota, de 1975 a 1984 estimó la prevalencia de miocardiopatía dilatada idiopática en 36.5: 100000 (~ 1: 2700) [2].

La MCD suele presentarse como: i) insuficiencia cardiaca con síntomas de congestión y / o reducción del gasto cardíaco, ii) arritmias y / o enfermedad del sistema de conducción cardiaco, y iii) enfermedad tromboembólica que incluye apoplejía o, iv) muerte súbita. Se estima que los eventos embólicos ocurren en el 4% de pacientes con miocardiopatía dilatada que tienen una fracción de eyección del ventrículo izquierdo (FEVI) ≤35% [2]. Además, el la incidencia de trombo del ventrículo izquierdo en pacientes con miocardiopatía dilatada y ritmo sinusal es del 13%, y el coágulo se localiza en el apéndice auricular izquierdo en el 68% de estos casos [3].

Por otro lado, se considera que existe MCDF cuando dos o más miembros de una misma familia son diagnosticados de MCD idiopática [4], siendo en gran parte una enfermedad de aparición

en adultos, aunque se ha demostrado una edad de inicio muy variable y penetrancia reducida [4-9]. La MCD idiopática es familiar entre el 20% y el 48% de los casos [10-13].

Aproximadamente el 50% de los casos de MCD tienen una etiología genética y hereditaria [14]. Más de 30 genes han sido asociados con MCDF [15], existiendo una considerable heterogeneidad genética [16], por lo que se necesita mejorar la estrategia actual de diagnóstico y tratamiento de esta enfermedad. Para ello, la utilización de las nuevas tecnologías, como la secuenciación de nueva generación (NGS), puede ayudar en la búsqueda nuevos marcadores moleculares y dianas terapéuticas.

El "método didesoxi" de Sanger [17] causó una revolución en la biología. Estos métodos llevaron a la secuenciación de genomas cada vez más grandes, culminando con el Proyecto Genoma Humano [18, 19]. Como siguiente paso, se emprendieron proyectos de secuenciación a gran escala para estudiar las variaciones de la secuencia humana. Sin embargo, para este tipo de proyectos, la secuenciación de Sanger era demasiado laboriosa y costosa. Por ello, el Instituto Nacional de Investigación del Genoma Humano (NHGRI) inició en 2004 un programa que aceleró el desarrollo de nuevos métodos. Así, las tecnologías de NGS han permitido la secuenciación del genoma completo de una persona con una importante reducción en el coste y tiempo empleado.

La tecnología NGS incluye la secuenciación del genoma completo (WGS), cuyo proceso es laborioso y costoso. Por esta razón, se han desarrollado técnicas de secuenciación dirigida, como los paneles de genes, o la secuenciación del exoma completo (WES).

El método de secuenciación del exoma ha conseguido eliminar la necesidad de elegir un subconjunto de genes para interrogar y enfocarse en los exones codificadores de proteínas, que supone un 1% del genoma [20], lo que supone un gran logro ya que se estima que el exoma contiene el 85% de las mutaciones causantes de enfermedades, así como muchos de los SNP (*Single-Nucleotide Polymorphism*) que predisponen a la enfermedad [21].

Para la miocardiopatía dilatada familiar existen pruebas diagnósticas para algunos genes que permiten detectar a pacientes afectados en estadios tempranos, aunque debido a su enorme heterogeneidad genética y fenotípica, el conocimiento de las alteraciones moleculares de esta enfermedad es limitado. Por ello, el análisis de la secuenciación del exoma completo puede ayudar a evidenciar la existencia de nuevas variantes genéticas asociadas a la enfermedad, permitiendo incluso identificar nuevos genes candidatos para el tratamiento de la enfermedad.

1.2 OBJETIVOS

Los objetivos del presente estudio son:

- Seleccionar posibles variantes genéticas en función de los patrones de herencia:
  - Análisis de la asociación con la enfermedad en función de los posibles modelos de herencia.

- Analizar el valor de diferentes softwares predictivos de patogenicidad.
  - Determinar que variantes son consideradas patogénicas mediante el uso de diferentes algoritmos basados en la conformación proteica. En este estudio se utilizarán SIFT [22], Polyphen-2 [23], Align-GVGD [24], Panther-PSEP [25], REVEL [26], PROVEAN [27] y FATHMM-XF [28] y Mutation Taster [29].

- Análisis de enriquecimiento de las rutas funcionales relacionadas con las variantes genéticas estudiadas:
  - De las variantes con mayor puntuación obtenidas en el objetivo dos, se analizará su relación con diferentes rutas biológicas mediante las bases de datos de Gene Ontology (GO), *KEGG* y *Open Targets Platform*.
  - Filtrado de las variantes seleccionadas por su implicación en las rutas biológicas más significativas.

1.3 ENFOQUE Y MÉTODO A SEGUIR

El uso de la secuenciación del exoma completo (WES) utilizando tecnología de secuenciación NGS, permite un estudio más eficiente, obteniendo además una mayor probabilidad de generar un diagnóstico. Estas técnicas generan una enorme cantidad de información, que debe ser filtrada y manejada. Para ello, es necesario es uso de herramientas bioinformáticas que ayuden con este propósito.

Dado que el objetivo principal de este trabajo es detectar los SNVs (*Single Nucleotide Variants*) que tienen una mayor relación con el desarrollo del MCDF, se procederá al uso de diferentes herramientas que nos permitan encontrar las variantes genéticas que cumplan este objetivo.

Para ello, se analizarán las diferentes variantes encontradas en la familia a estudio, teniendo en cuenta el modelo de herencia y la frecuencia del alelo menos común. Para poder predecir el impacto biológico de una mutación, se han desarrollado numerosos predictores “in silico”.

Debido a que los algoritmos utilizados por cada predictor pueden variar, en el presente trabajo se utilizarán diferentes herramientas bioinformáticas. Así, mediante el registro de las puntaciones se analizarán, utilizando los datos obtenidos por WES, la posible patogenicidad de las variantes genéticas encontradas de una forma más efectiva.

Por último, se realizará el enriquecimiento de las variantes seleccionadas como patogénicas, analizando las rutas en las que pueden estar involucradas.

1.4 PLANIFICACIÓN

1.4.1. Tareas

Tareas objetivo 1 (7 días)

Filtrado por frecuencia de las variantes (8 días)

Análisis teniendo en cuenta los diferentes modelos de herencia (3 días)

Tareas objetivo 2 (31 días)

Estudios “in silico” de patogenicidad (13 días)

Filtrado por predicción (7 días)

Tareas objetivo 3 (14 días)

Exploración de las bases de datos (*Gene Ontology*, *KEGG* y *Open Targets Platform*). (13 días)

Determinación de las variantes genómicas implicadas en las rutas más significativas (9 días)

Elaboración de la memoria (12 días)

Preparación de la presentación (8 días)

1.4.2. Calendario

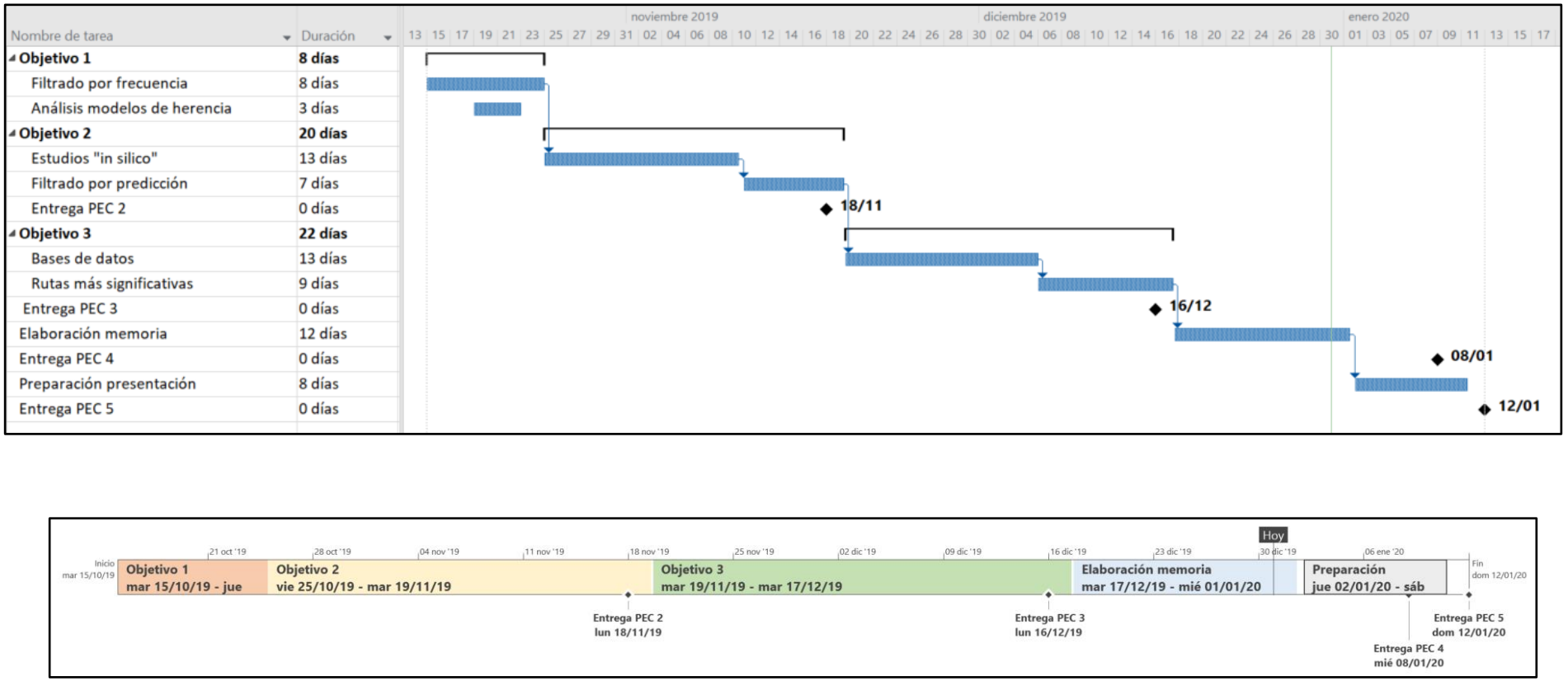


Figura 1. Calendario de planificación del TFM

1.4.3. Hitos

Para la correcta realización del proyecto dentro de los plazos previstos, deberán cumplirse los siguientes hitos:

- Desarrollo del trabajo Fase I (Entrega PEC 2): 18/11/2019:
  - Tareas objetivo 1
  - Tareas objetivo 2
- Desarrollo del trabajo Fase II (Entrega PEC 3): 16/12/2019
- Cierre de la memoria (Entrega PEC 4): 08/01/2020
- Elaboración de la presentación (Entrega PEC 5): 12/01/2020

1.5 BREVE SUMARIO DE PRODUCTOS OBTENIDOS

1. Número de variantes según la frecuencia del alelo menos común y tipo de herencia.
2. Código R para para filtrado de las variantes teniendo en cuenta los patrones de herencia.
3. Obtención de genes mutados con alta probabilidad de patogenicidad.
4. Obtención de genes mutados implicados en las rutas biológicas más relevantes.
5. Memoria final del trabajo, que detalla el trabajo realizado en las anteriores PECs.

1.6 BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA

La memoria del proyecto estará compuesta por las siguientes partes:

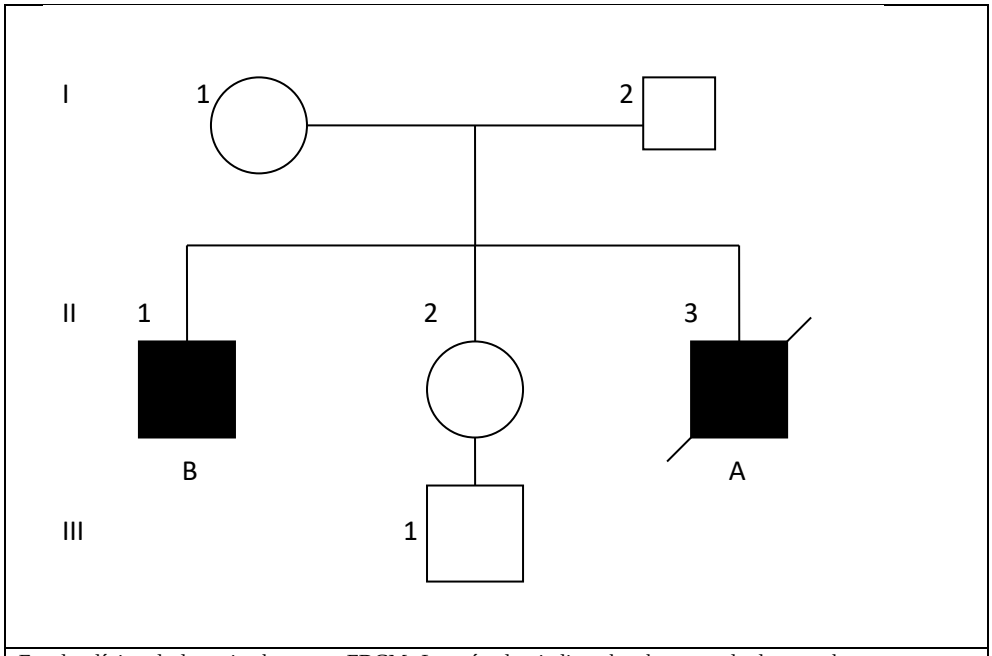
1. Material y métodos: explicación detallada de los datos y herramientas bioinformáticas utilizadas en el análisis. Pasos llevados a cabo para la elaboración del proyecto.
2. Resultados: descripción de los resultados obtenidos tras los diferentes análisis realizados.
3. Discusión: revisión bibliográfica y evaluación de los resultados.
4. Conclusiones: enumeración de las deducciones del proyecto de la forma más concreta posible.

2. RESTO DE CAPÍTULOS

2.1 MATERIAL Y MÉTODOS

2.1.1. Obtención de los datos

Para este trabajo, se analizarán los datos obtenidos de una familia de seis miembros. Dos de los familiares de primer grado están diagnosticados con cardiopatía familiar dilatada de origen idiopático.



Estado clínico de los miembros con FDCM. Los círculos indican hembras; cuadrados, machos; y un corte, fallecido. Los símbolos rellenos indican el estado afectado y los símbolos abiertos no afectados. A y B son probandos para DCM.

Figura 2. Árbol genealógico de la familia en estudio. I.1 Padre sano, I.2 Madre sana, II.1 Hijo diagnosticado, II.2 Hija sana, II.3 Hijo fallecido, III.1 Tercera generación sana.

Para el estudio se utilizaron los datos obtenidos mediante la secuenciación del exoma completo (WES) de los familiares de primer grado (padres y hermanos) de los probandos (II.1 y II.3). No se obtuvieron datos de secuenciación para el individuo III.1.

La secuenciación fue llevada a cabo mediante el kit de Ion AmpliSeq Exoma (*ExoNIM, NIMGenetics*) de la tecnología de *Ion proton* (*Life Technologies Coorporation*). El flujo de trabajo bioinformático previo incluyó la eliminación de adaptadores de las secuencias, alineamiento de las secuencias al genoma de referencia e identificación mediante distintos algoritmos de SNVs en cada uno de los individuos.

2.1.2. Análisis

Utilizando los datos anteriores, se realizó el análisis de las SNVs según se ha descrito en el apartado de "Enfoque y método a seguir". En Anexos se puede encontrar el código de R utilizado en los diferentes apartados.

Filtrado por frecuencia de las variantes

Utilizando los datos anteriores, se utilizó la frecuencia de *Genome Aggregation Database* (gnomAD). Esta base de datos es un recurso desarrollado a nivel internacional, con el objetivo de agregar y armonizar los datos de secuenciación de una amplia variedad de proyectos a gran escala, y hacer que estos datos estén disponibles. Proporciona frecuencias alélicas (AF) con una significación estadística sólida.

Esta base muestra la frecuencia alélica para diferentes poblaciones, entre las que se encuentra la NFE (*Non-Finnish European*), la cual utilizaremos para el presente estudio. En estudios de enfermedades genéticas raras, las variantes comunes se consideran por encima de un umbral de AF de 0.01 [30, 31]. Por esta razón, para diferenciar entre las variantes raras y comunes en la población, se ha utilizado este umbral. Así, las variantes con una frecuencia menor de 0.01 serán consideradas raras, mientras las que tengan una frecuencia igual a superior serán consideradas comunes.

Análisis de los modelos de herencia

Para este estudio se han contemplado los siguientes modelos de herencia: variantes autosómicas dominantes heredadas de los parentales u originadas *de novo*, variantes autosómicas recesivas heredadas de los parentales o *de novo* y variantes ligadas a X heredadas o *de novo*. En el caso del modelo de herencia recesiva, las mutaciones pueden estar en posiciones distintas del mismo gen, uno en cada cromosoma (heterocigosis compuesta). Por esta razón, hemos diferenciado entre heterocigosis simple o compleja.

1. Variantes autosómicas dominantes heredadas de los parentales. Filtramos aquellas variantes que estén presentes en los hijos y al menos uno de los parentales.

2. Variantes autosómicas dominantes *de novo*. Filtramos aquellas variantes que están presentes en los hijos, pero no en los progenitores.

3. Variantes autosómicas recesivas:

- Heterocigosis compuesta: filtramos aquellos genes que tienen más de una variante. Para las variantes heredadas, se tienen en cuenta aquellas presentes en ambos parentales. En el caso de nuevas mutaciones, se considera que uno de los parentales tiene la variante, mientras la otra se ha originado *de novo*.

- Heterocigosis simple: filtramos las variantes teniendo en cuenta si han sido heredadas de ambos progenitores u originadas de novo. En el segundo caso, al igual que para la heterocigosis compuesta, se considera que uno de los parentales tiene la variante.

4. Variantes ligadas a X. Estas variantes afectan de forma diferente si el portador es hombre o mujer, debido a que los hombres estarán afectados sea dominante o recesiva, en cambio; las mujeres afectadas necesitan que la variante sea dominante o, de ser recesiva, esté en ambas copias. Por ello se analizará:

- Variantes dominantes: las heredadas implican que la variante esté al menos presente en uno de los progenitores, en el caso de los hombres se tendrán en cuenta aquellas variantes procedentes de la madre. Las variantes *de novo* están presentes en los hijos, pero no en los progenitores
- Variantes recesivas: en el caso de la hija la variante puede ser heredada de ambos progenitores o que se produzca *de novo*, en cuyo caso se considera que una de las variantes procede de uno de los parentales. Para los hombres se seguiría el mismo procedimiento que si fuese dominante.

Impacto biológico

Existen numerosos predictores *in silico* que ayudan a predecir la patogenicidad de una mutación determinada. Estas herramientas predicen el posible impacto de las sustituciones aminoacídicas en la estructura y función de las proteínas. Para llevar a cabo este apartado tendremos en cuenta aquellas variantes presentes en los dos hermanos afectados por MCDF.

En este estudio se han utilizado las siguientes herramientas:

1. SIFT (*Sorting Intolerant From Tolerant*) se basa en el grado de conservación de los residuos aminoacídicos que es modificado por la variante. La puntuación para cada posible sustitución de aminoácidos es la probabilidad de que la sustitución sea o no tolerada evolutivamente [22].

2. PolyPhen-2 (*Polymorphism Phenotyping*) está basado en un árbol de decisión, el cual tiene en cuenta consideraciones físicas y similitud de secuencias [23].

3. PROVEAN *(Protein Variation Effect Analyzer)* se basa en la homología de secuencias. Calcula la puntuación de la alineación entre una secuencia de proteínas y un gran número de variaciones de un solo locus de otra proteína se ha desarrollado previamente [27].

4. FATHMM-XF es un método basado en la asociación de la conservación evolutiva de las secuencias con puntuaciones específicas de la enfermedad [28].

5. REVEL (*Rare Exome Variant Ensemble Learner*) es un método que predice la patogenicidad de variantes sin sentido basándose en la combinación de las puntuaciones de 13 predictores: MutPred, FATHMM, VEST, PolyPhen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, y phastCons. En nuestro estudio se consideraron patogénicas las puntuaciones por encima de 0.5, que corresponden a una sensibilidad de 0.754 y especificidad de 0.891 [26].

6. Align-GVGD combina las características biofísicas de los aminoácidos y las alineaciones de múltiples secuencias de proteínas para analizar si las sustituciones sin sentido se encuentran en un espectro deletéreo a neutral. Está basado en la diferencia de Grantham [24]. La salida de este trabajo nos devuelve una clasificación cuyo valor indica si la variante a estudio es más o menos probable que interfiera con la función de la proteína. Así, estos valores van de class C65 (mayor probabilidad) a class C0 (menor probabilidad).

7. Panther-PSEP estima la probabilidad de que un SNP cause un impacto funcional en la proteína. Calcula el período de tiempo (en millones de años) que un aminoácido dado se ha conservado en el linaje que conduce a la proteína de interés. Cuanto más largo sea el tiempo de conservación, mayor será la probabilidad de impacto funcional [25].

8. Mutation Taster emplea el clasificador de Bayes para predecir el potencial de enfermedad de una alteración [29].

Nombre del predictor	Servidor Web
SIFT	https://sift.bii.a-star.edu.sg/www/SIFT_dbSNP.html
Polyphen-2	http://genetics.bwh.harvard.edu/pph2/bgi.shtml
PROVEAN	http://provean.jcvi.org/genome_submit_2.php?species=human
FATHMM-XF	http://fathmm.biocompute.org.uk/fathmm-xf/
REVEL	https://sites.google.com/site/revelgenomics/downloads
Align-GVGD	http://agvgd.hci.utah.edu/agvgd_input.php
Panther-PSEP	http://www.pantherdb.org/tools/csnpScoreForm.jsp?
Mutation Taster	http://www.mutationtaster.org/ChrPos.html

Tabla 1. Herramientas para el análisis *in silico.*

Enriquecimiento

El enriquecimiento de genes permite analizar, mediante el uso de bases de datos, los términos biológicos asociados a un gen o proteína. Esto nos permite obtener información más detallada sobre las relaciones funcionales y, por tanto; la posible implicación en rutas afectadas por MCDF. Para ello se analizaron las siguientes bases de datos: KEGG, *Gene Ontology* y *Open Targets Platform*

KEGG (*Kyoto Encyclopedia of Genes and Genomes*) es una base de datos usada para comprender las funciones y las características de elementos de alto nivel del sistema biológico, como la célula, el organismo y el ecosistema, a partir de información a nivel molecular, especialmente conjuntos de datos moleculares a gran escala generados por secuenciación del genoma y otras tecnologías experimentales de alto rendimiento [32]. Además de las rutas moleculares, gracias a esta base de datos podemos obtener información de las enfermedades relacionadas con un gen en concreto [33].

La base de conocimiento de Gene Ontology (GO) es la mayor fuente de información sobre las funciones de los genes. GO asigna un único identificador numérico a cada término, el cual a su vez se incluye dentro de una de las tres ontologías existentes: componente celular, función molecular o proceso biológico [34]. En el presente estudio analizaremos los procesos biológicos relacionados con las variantes seleccionadas.

Para estudiar más profundamente la asociación de los genes afectados por las variantes seleccionadas con distintas enfermedades, se obtuvieron los datos de la página web "*Open*

Targets Platform". Esta base es una integración de datos completa y robusta para el acceso y la visualización de posibles objetivos farmacológicos asociados con la enfermedad, mediante la puntuación de estas asociaciones basadas en la evidencia de 20 fuentes de datos. En nuestro análisis hemos analizado las enfermedades asociadas con un score ≥ 0.01.

2.2 RESULTADOS

2.2.1. Filtrado por frecuencia y tipo de herencia

Utilizando la base de datos de gnomAD de la población NFE, las variantes fueron filtradas por su frecuencia alélica. De todas las variantes obtenidas, el 95.6% está presente en esta base de datos.

Una vez filtradas las variantes, teniendo en cuenta el umbral de 0.01 descrito previamente, se realizó el análisis por el tipo de herencia. En las tablas podemos observar el número de variantes para cada condición.

Variantes comunes

Herencia autosómica dominante

	II-1	II-2	II-3
Autosómica dominante heredada	40146	40941	40924
Autosómica dominante de novo	444	504	459

Tabla 2. Herencia autosómica dominante variantes comunes.

Herencia autosómica recesiva

	Heterocigosis simple			Heterocigosis compleja		
	II-1	II-2	II-3	II-1	II-2	II-3
Autosómica recesiva heredada	22884	23488	23526	35768	36446	36501
Autosómica recesiva de novo	17261	17451	17397	445	507	472

Tabla 3. Herencia autosómica recesiva variantes comunes.

Herencia ligada a gen X

En este caso diferenciamos el caso de las mujeres (la hermana II-2), de los hombres (hermanos II-1 y II-3)

	Dominante		Recesiva	
	Heredada	de novo	Heredada	de novo
II-2	334	9	110	224

Tabla 4. Herencia ligada gen X hija variantes comunes.

	II-1	II-3
Ligada a X heredada	217	5
Ligada a X de novo	220	5

Tabla 5. Herencia ligada gen X hijos variantes comunes.

Variantes raras

Herencia autosómica dominante

	II-1	II-2	II-3
Autosómica dominante heredada	4204	4288	4182
Autosómica dominante de novo	4342	4386	4268

Tabla 6. Herencia autosómica dominante variantes raras.

Herencia autosómica recesiva

	Heterocigosis simple			Heterocigosis compleja		
	II-1	II-2	II-3	II-1	II-2	II-3
Autosómica recesiva heredada	3045	3101	3107	1680	1724	1678
Autosómica recesiva de novo	1159	1187	1075	544	548	485

Tabla 7. Herencia autosómica recesiva variantes raras.

Herencia ligada a gen X

Al igual que en el caso anterior, diferenciamos el análisis en hombres o mujeres.

	Dominante		Recesiva	
	Heredada	de novo	Heredada	de novo
II-2	100	101	39	61

Tabla 8. Herencia ligada gen X hija variantes raras.

	II-1	II-3
Ligada a X heredada	44	67
Ligada a X de novo	47	69

Tabla 9. Herencia ligada gen X hijos variantes raras.

2.2.2. Selección de variantes a estudio

Para llevar a cabo los siguientes pasos del flujo de trabajo previsto, se decidió analizar aquellas variantes compartidas entre los dos individuos afectados, es decir II-1 y II-3. La razón de esta decisión es conocer que variantes pueden generan el fenotipo de la enfermedad. En la figura 3 se observan las variantes que tienen en común ambos hermanos.

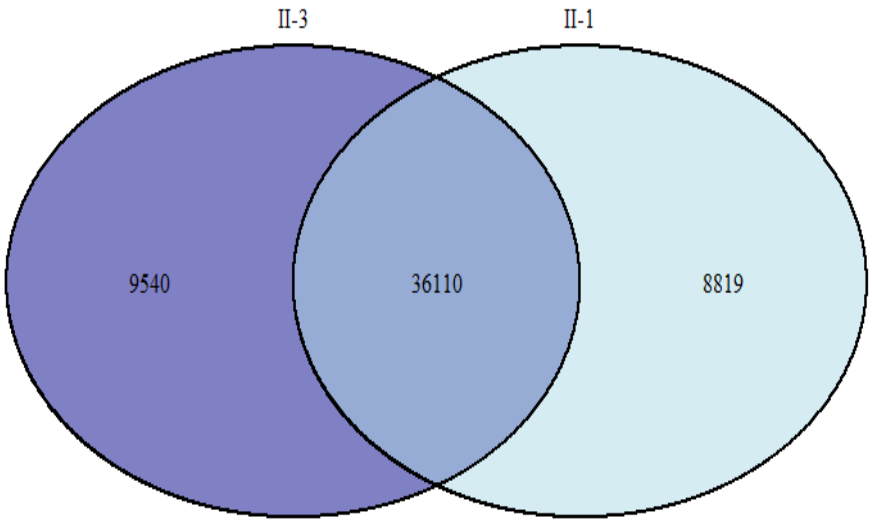


Figura 3. Diagrama de Venn de la descendencia afectada. III.1 Hijo diagnosticado y II.3 Hijo fallecido.

En la siguiente figura se muestra el número de variantes comunes entre estos individuos afectos y la hermana sana (II-2), que como se puede observar es un alto porcentaje.

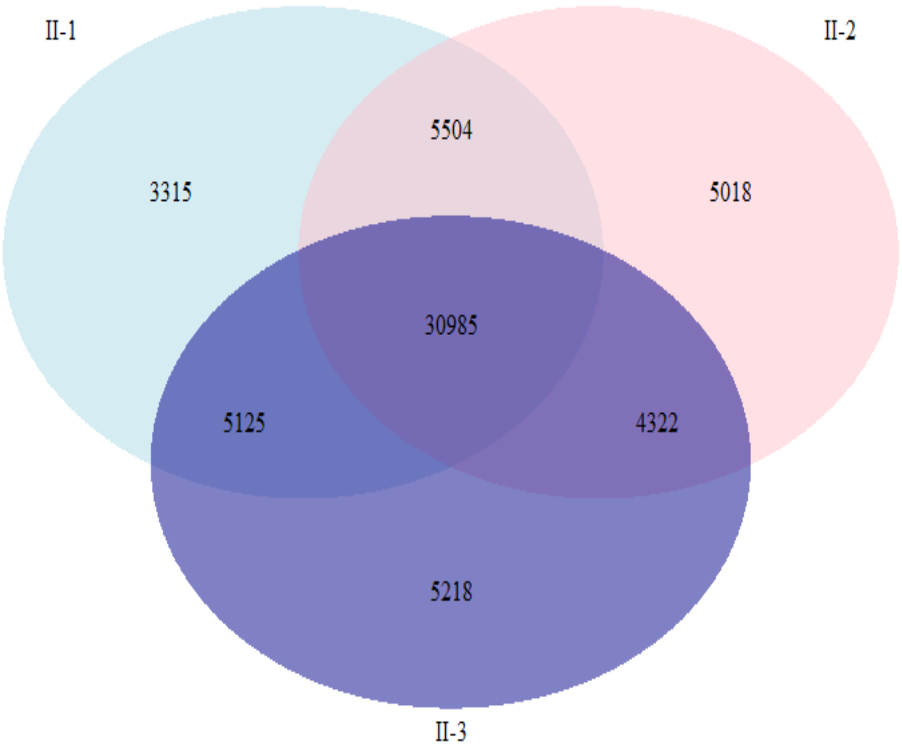


Figura 4. Diagrama de Venn de la descendencia. III.1 Hijo diagnosticado, II.2 Hija sana y II.3 Hijo fallecido.

2.2.3. Estudios "in silico" de patogenicidad

Como ya se ha descrito previamente, se analizó la predicción del daño o perjuicio de los SNPs usando tanto el enfoque basado en la homología de la secuencia y de la estructura como en la identificación del fenotipo de enfermedad asociado con SNPs.

Tras analizar los diferentes logaritmos utilizados, se decidió que las variantes que fueran consideradas dañinas o perjudiciales por todos los algoritmos fueran analizadas en las siguientes fases del estudio. El hecho de que todos los predictores utilizados consideren patogénicas las variantes le da una mayor fuerza estadística al análisis.

Debemos tener en cuenta una excepción en la variante rs13312995, cuyo resultado con Panther-PSES fue *Not scored*: *No PANTHER family for input sequence*. Esto significa que la secuencia de entrada no coincide con ninguna secuencia para la que hay una familia de genes en PANTHER. O la reconstrucción ancestral de la familia de genes coincidentes en PANTHER no tuvo éxito.

Independientemente de este resultado, los otros siete predictores la consideraron patogénica y por ello se decidió tenerla en consideración para el posterior análisis.

Las siguientes tablas muestran los resultados tanto para las variantes comunes como para las raras.

Variantes comunes

	rs1664022	rs1714864	rs13312995
Sustitución	R25L	P156L	R830W
Gen	PDE4DIP	KCNJ12	AHI1
SIFT	DELETERIOUS	DELETERIOUS	DELETERIOUS
Polyphen-2	damaging	damaging	damaging
PROVEAN	Deleterious	Deleterious	Deleterious
FATHMM-XF	pathogenic	pathogenic	pathogenic
REVEL	0.696	0.787	0.501
Panther-PSEP	probably damaging	probably damaging	Not scored
Mutation Taster	Disease causing	Disease causing	Disease causing
Align-GVGD	Class C65	Class C65	Class C65

Tabla 10. Variantes comunes seleccionadas tras los estudios *in silico*.

Variantes raras

	rs34595252
Sustitución	D658G
Gen	WDR36
SIFT	DELETERIOUS
Polyphen-2	damaging
PROVEAN	Deleterious
FATHMM-XF	pathogenic
REVEL	0.926
Panther-PSEP	probably damaging
Mutation Taster	Disease causing
Align-GVGD	Class C65

Tabla 11. Variante rara seleccionada tras los estudios *in silico*.

2.2.4. Enriquecimiento

De las variantes seleccionadas tras el estudio *in silico*, se realizó el enriquecimiento de las rutas funcionales, independientemente de que las variantes fuesen comunes o raras en la población. Además, se analizaron las enfermedades asociadas a los genes asociados a las variantes a través de la página web "*Open Targets Platform*".

Las rutas encontradas para cada base son las siguientes:

Gene Ontology

La siguiente tabla muestra las rutas metabólicas enriquecidas relacionadas con los procesos biológicos referidos a musculatura y corazón.

SNP ID	GEN	GO ID	DESCRIPCIÓN
rs1714864	KCNJ12	GO:0006936	muscle contraction
		GO:0061337	cardiac conduction
		GO:0008016	regulation of heart contraction
rs13312995	AHI1	GO:0001947	heart looping

Tabla 12. Procesos biológicos enriquecidos mediante *Gene Ontology*.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

Cómo ya se mencionó previamente en material y métodos, KEGG nos permite obtener información tanto de las rutas moleculares, como de las enfermedades relacionadas con un gen en concreto.

KEGG Pathways

KEGG Pathways
KCNJ12 (rs1714864)
Organismal Systems
Nervous system
04725 Cholinergic synapse
Organismal Systems
Endocrine system
04921 Oxytocin signaling pathway
WDR36 (rs34595252)
Genetic Information Processing
Translation
03008 Ribosome biogenesis in eukaryotes

Tabla 13. Rutas metabólicas enriquecidas mediante KEGG.

KEGG Diseases

KEGG Diseases
WDR36 (rs34595252)
Nervous system diseases
Eye disease
00612 Primary open angle glaucoma
AHI1 (rs13312995)
Congenital malformations
Congenital malformations of the nervous system
00530 Joubert syndrome

Tabla 14. Enfermedades relacionadas mediante KEGG.

Open Targets Platform

En este caso, se tuvieron en cuenta las rutas con un score mayor o igual a 0.01 relacionadas con musculatura y corazón. A continuación, se muestran las tablas con la información obtenida.

PDE4DIP (rs1664022)		
Disease	**Score**	**Therapeutic area**
angiosarcoma	0.28	neoplastic, precancerous and hyperplastic disease | vascular disease
hemangioblastoma	0.17	nervous system disease | neoplastic, precancerous and hyperplastic disease | vascular disease
adrenal cortex carcinoma	0.17	neoplastic, precancerous and hyperplastic disease | urinary system disease | vascular disease | endocrine system disease
hypertrophic cardiomyopathy	0.03	musculoskeletal or connective tissue disease | heart disease
stroke	0.02	nervous system disease | vascular disease

Tabla 15. Enfermedades relacionadas con PDE4DIP mediante *Open Targets Platform*.

KCNJ12 (rs1714864)		
Disease	**Score**	**Therapeutic area**
dilated cardiomyopathy	0.04	musculoskeletal or connective tissue disease | heart disease
Familial dilated cardiomyopathy	0.02	genetic, familial or congenital disease | musculoskeletal or connective tissue disease | heart disease
cardiac arrhythmia	0.01	heart disease

Tabla 16. Enfermedades relacionadas con KCNJ12 mediante *Open Targets Platform*.

AHI1 (rs13312995)		
Disease	**Score**	**Therapeutic area**
Microcephaly - facio-cardio-skeletal syndrome, Hadziselimovic type	0.19	genetic, familial or congenital disease | heart disease
peripheral arterial disease	0.10	vascular disease
Ataxia-telangiectasia	0.04	metabolic or nutritional disease | nervous system disease | neoplastic, precancerous and hyperplastic disease | genetic, familial or congenital disease | urinary system disease | integumentary system disease | psychiatric disorder | visual system disease | immune system disease | vascular disease | reproductive system or breast disease | endocrine system disease
sleep apnea	0.03	nervous system disease | integumentary system disease | psychiatric disorder | heart disease | respiratory or thoracic disease
cerebrovascular disorder	0.01	nervous system disease | vascular disease
Schizencephaly	0.01	nervous system disease | genetic, familial or congenital disease | vascular disease
rheumatic heart disease	0.01	heart disease

Tabla 17. Enfermedades relacionadas con AHI1 mediante *Open Targets Platform*.

2.3 DISCUSIÓN

Como se mencionó anteriormente se analizaron, mediante varios predictores de patogenicidad, las variantes compartidas entre los dos hermanos afectos, ya que ambos individuos han desarrollado miocardiopatía dilatada familiar.

Los predictores más descritos y utilizados son SIFT, Align-GVGD, Panther y PolyPhen-2. Existen varias diferencias entre ellos a la hora de analizar la patogenicidad de las variantes. Así, el método SIFT es una herramienta basada en la homología de secuencia, a través de la alineación de secuencia múltiple de proteínas (PMSA). Predice las variantes en la secuencia de consulta como "neutral" o "perjudicial" utilizando probabilidades normalizadas calculadas a partir de la alineación de secuencia múltiple de entrada [35]. La alineación construida por SIFT contiene secuencias homólogas con una medida de conservación media de 3.0 donde la conservación está representada por el contenido de la información [36] para minimizar los errores falsos positivos y falsos negativos. Las variantes en una posición con probabilidades normalizadas inferiores a 0,05 se predicen deletéreas y se pronostican neutrales con una probabilidad mayor o igual a 0,05. Un segundo algoritmo de análisis que utiliza matrices de puntuación (similar al enfoque SIFT) está integrado en la base de datos PANTHER [37]. La misión principal de la base de datos PANTHER es organizar los genes en familias y subfamilias y clasificarlos de acuerdo con la función inferida. Gran parte de la organización lograda por esta base de datos se basa en hacer PMSA en un gran número de subfamilias y familias de genes. Sin embargo, una limitación importante es que los PMSA de PANTHER generalmente cubren solo las porciones más conservadas de genes, limitando la fracción de sustituciones a las que se puede aplicar.

Por otro lado, Align-GVGD predice las variantes en la secuencia de consulta basándose en una combinación de la variación de Grantham (GV), que mide la cantidad de variación evolutiva bioquímica observada en una posición particular en la alineación basándose en un PMSA, y la desviación de Grantham (GD), que mide la diferencia bioquímica entre referencia y aminoácido codificado por la variante [38]. El clasificador original utiliza un conjunto de cinco criterios basados en GV y GD que clasifica las variantes como "neutrales", "no clasificadas" o "perjudiciales" [38]. Por ejemplo, en el caso extremo de GV = 0, la alineación se conserva completamente en esa posición y cualquier otra variante se considerará perjudicial. El nuevo clasificador proporciona grados ordenados que van desde los más patógenos a los menos probables[39]. Una característica importante es que el sitio web Align-GVGD alberga alineamientos de secuencias múltiples de proteínas curadas para varios genes importantes de susceptibilidad al cáncer, como BRCA1 y TP53.

Un enfoque algo independiente para el análisis de las sustituciones depende en gran medida de las consideraciones estructurales de las proteínas, a veces junto con algoritmos de aprendizaje automático. Los clasificadores basados en reglas o árboles de decisión se centran en una serie de características anotadas en las posiciones de aminoácidos individuales de una proteína humana. PolyPhen-2 (Polymorphism Phenotyping) es el clasificador más conocido de esta familia y predice las variantes como "benignas", "posiblemente dañinas" o "probablemente dañinas" en base a ocho características predictivas basadas en secuencia y tres basadas en estructura que fueron seleccionadas por un algoritmo codicioso iterativo. Otra característica útil es que el algoritmo calcula una probabilidad posterior de Bayes de que una mutación dada sea perjudicial [23]. La versión basada en la web selecciona secuencias homólogas usando un algoritmo de agrupamiento y luego construye y refina la alineación produciendo una alineación que contiene tanto ortólogos como parálogos que pueden o no ser de longitud completa, lo que produce una mayor amplitud de secuencias, pero una profundidad reducida en comparación con la alineación Align-GVGD. Los autores sostienen que esto conduce a predicciones más precisas porque la mayoría de las variantes perjudiciales afectan la estructura de la proteína en comparación con la función específica de la proteína [23].

En términos generales, la precisión de los predictores varía del 65 al 80% cuando se analizan variantes patogénicas [40]. Por ello, se decidieron incorporar otros cuatro algoritmos a los previamente mencionados (PROVEAN, REVEL, FATHMM-XF y MutationTaster), obteniendo así un mayor abanico para poder decidir qué predictores se utilizarían finalmente. De hecho, en los últimos años, se han llevado a cabo numerosos estudios para analizar que predictores, individualmente o en conjunto, son más efectivos [39, 41-50].

Así, un estudio sobre la predicción para las mutaciones CFTR (regulador de la conductancia transmembrana de la fibrosis quística) mostró que las puntuaciones de PANTHER tienen una correlación estadística global significativa con el espectro de gravedad de la enfermedad asociada con mutaciones en el gen CFTR. En contraste, las puntuaciones derivadas de PolyPhen y SIFT solo mostraron diferencias significativas entre las variantes que causan fibrosis quística y las que no lo hacen [41].

Otro estudio en cáncer de colon hereditario, mostró que Align-GVGD tenía mejor especificidad que la sensibilidad y, por lo tanto, una predicción relativamente baja de falsos positivos de patógenos. Por otro lado, SIFT y PolyPhen mostraron mejor sensibilidad que especificidad y, por lo tanto, predicción relativamente baja de falsos negativos de neutral.

En este mismo estudio se mostró que la especificidad de SIFT y PolyPhen fueron similares (81%), pero la sensibilidad de SIFT fue mucho mejor que la de PolyPhen (82% frente a 58%). En particular, SIFT creó las alineaciones y los resultados podrían haber sido mejores si se hubieran ejecutado con PMSA optimizados y curados [42].

El desarrollo de un nuevo clasificador denominado PredictSNP2 incluyó a FATHMM, siendo el predictor, entre los seis analizados, con mejor precisión para diferentes tipos de mutaciones [46].

Por otro lado, en una investigación realizada para evaluar la fiabilidad in silico, se probaron cuatro herramientas de predicción (Align-GVGD, SIFT, PolyPhen-2, MutationTaster) para analizar 236 variantes BRCA1 / 2 que previamente habían sido clasificadas por comités de expertos. Así, PolyPhen-2 logró los valores más bajos de sensibilidad, especificidad, precisión y Coeficiente de correlación de Matthews (MCC). Align-GVGD logró los valores más altos de especificidad, precisión y MCC, pero SIFT y MutationTaster lo superaron en cuanto a su sensibilidad [47]. Este estudio muestra que las consecuencias clínicas no deben basarse únicamente en pronósticos *in silico*, ya que las predicciones pueden ser bastante pobres.

Por ello, y para mejorar el rendimiento de predicción, se han estudiado varias combinaciones de los predictores existentes. De hecho, el clasificador PredictSNP incluye, entre otras herramientas, los predictores SIFT, PANTHER y PolyPhen-2, mostrando un mejor rendimiento de predicción en conjunto [44].

También se ha observado que las predicciones combinadas de cinco algoritmos de uso común (Polyphen, SIFT, CADD, PROVEAN y MutationTaster) tienen mayor concordancia frente a los 18 estudiados [45]. Además, SIFT y PROVEAN mostraron ser la mejor combinación de herramientas *in silico* para las variantes del gen KCNH2, relacionado con el síndrome de QT largo [50].

Cómo podemos observar tras analizar algunos de los estudios llevados a cabo con diferentes algoritmos, es difícil deducir con cuál se obtendría la mejor predicción, siendo por tanto la combinación de varios la mejor opción. Dado que uno de los objetivos del presente proyecto es obtener una buena predicción de las variantes estudiadas, se utilizaron los ocho algoritmos anteriormente descritos con el fin de obtener un rendimiento mejorado de la predicción.

De entre las variantes analizadas, cinco fueron consideradas dañinas o perjudiciales por todos los predictores analizados, y por tanto fueron seleccionadas. En la siguiente tabla se muestra la información sobre dichos SNVs, con los transcritos considerados como patogénicos por los predictores analizados.

SNP rsID	Sustitución	Tipo de herencia	Gen	Transcritos
rs1664022	R25L	Recesiva hereditaria	PDE4DIP	ENSP00000358353
				ENSP00000358355
				ENSP00000358357
				ENSP00000358360
				ENSP00000358363
				ENSP00000435483
rs34595252	D658G	Dominante hereditaria	WDR36	ENSP00000423067
				ENSP00000424628
rs1714864	P156L	Recesiva hereditaria	KCNJ12	ENSP00000328150
rs13312995	R830W	Dominante hereditaria	AHI1	ENSP00000265602
				ENSP00000322478
				ENSP00000356774
				ENSP00000388650

Tabla 18. Variantes seleccionadas tras el estudio *in silico*.

El gen PDE4DIP, en el cual se encuentra la primera variante seleccionada, no tiene ninguna ruta o enfermedad asociada a las bases KEGG o *Gene Ontology*. En cambio, a través de *Open Targets Platform* (OTP), podemos observar que este gen está asociado a varios procesos neoplásicos, así como a enfermedades vasculares y cardiacas (Tabla 18). De hecho, el producto de este gen, la proteína Miomegalina, ha sido descrito como un importante regulador de la contractibilidad cardíaca, pudiendo ser un factor causante de la cardiomiopatía hipertrófica [51]. Además, se ha observado, una relación entre este gen y un alto riesgo de accidente cerebrovascular isquémico [52].

La base de datos de OTP y *Gene Ontology* no ha mostrado ningún resultado para el gen WDR36 con los filtros utilizados, aun así, se ha podido realizar el enriquecimiento a través de la base KEGG, que muestra una asociación de este gen con glaucoma y con la biogénesis mitocondrial en eucariotas.

Respecto al gen AHI1, las bases de datos analizadas muestran que está relacionado con la torsión cardiaca y el síndrome de Joubert, el cual afecta a nivel multiorgánico y puede implicar defectos cardiacos, aunque no es común [53]. Por otro lado, la base OTP muestra una relación de este gen con numerosas enfermedades vasculares, entre ellas enfermedad reumática cardiaca.

Por último, la base de *Gene Ontology* muestra la relación del gen KCNJ12 con rutas de regulación cardiaca y muscular. Además, KEGG asocia este gen con la señalización de Oxitocina y la sinapsis Colinérgica, estando ambas rutas relacionadas con procesos cardiacos. La oxitocina ha sido descrita por sus propiedades cardioprotectoras [54, 55], y se ha observado que la actividad colinérgica es un importante regulador de la remodelación cardíaca y que mejorar la señalización colinérgica en pacientes humanos con enfermedad cardíaca puede reducir la morbilidad y la mortalidad [56].

Además, *Open Targets Platform* muestra la asociación de este gen con la cardiomiopatía dilatada, cardiomiopatía dilatada familiar y arritmias cardiacas. Un estudio previo realizado en una familia con MCDF mediante WES, ha descrito que mutaciones en el gen KCNJ12 puede ser una causa de esta enfermedad [57].

En nuestro estudio hemos visto que la variante del gen KCNJ12 presenta una herencia recesiva hereditaria (tabla 21). Esta variante está presente en todos los miembros de la familia, pero sólo dos de ellos han desarrollado la enfermedad. Esto podría explicarse debido a la recesividad del SNV, siendo estos individuos (II-1 y II-2) los únicos que tengan las dos copias de la variante, es decir; han heredado esta mutación de ambos parentales.

3. CONCLUSIONES

En este trabajo, se puede concluir que el SNV que presenta una mayor relación con el desarrollo de la enfermedad es la variante del gen KCNJ12. Este gen juega un papel importante en la alteración de la estructura y conducción cardiaca, y nuestro estudio muestra que la mutación P156L parece ser patogénica.

Además, tras estudiar los diferentes predictores, se decidió incluir a los ocho utilizados en el análisis de las variantes, ya que la combinación de estos nos permite obtener un rendimiento de predicción significativamente mejorado.

Se han cumplido todos los objetivos propuestos en el plan de trabajo, aunque se han llevado a cabo varios cambios en la planificación inicial para mejorar el resultado final, como la modificación en el tipo de filtrado por frecuencia de las variantes, la adicción de un nuevo predictor y de una nueva base para el enriquecimiento de las variantes seleccionadas.

Una de las limitaciones del presente trabajo es que los estudios realizados sobre una sola familia, tienen la desventaja de que los resultados no pueden ser generalizables, ya que no están basados en estudios sistemáticos. Sin embargo, este tipo de análisis ayuda a la identificación de nuevas enfermedades, a la detección de nuevas dianas terapéuticas, y a la identificación de raras manifestaciones de una enfermedad. Por otro lado, la falta del análisis de CNV (variación en el número de copias) sería otra limitación del estudio, pero no en la valoración de las predicciones de patogenicidad.

Para completar este estudio, considero que sería importante tener una confirmación de la mutación candidata del gen KCNJ12, que se podría llevar a cabo con diferentes técnicas como la secuenciación de Sanger. Además de la confirmación, un siguiente paso podría ser el estudio de la funcionalidad de esta mutación para poder conocer su implicación en el fenotipo de la miocardiopatía dilatada familiar.

Por otro lado, otro punto a tener en cuenta sería el análisis de las combinaciones de los predictores, ya que considero que las herramientas web de SIFT, PolyPhen-2 y PROVEAN presentan una gran facilidad de manejo respecto al resto, permitiendo una gran cantidad de formatos de identificación de las variantes y la inserción de varias al mismo tiempo. Esto supone una ventaja y sería un punto a tener en cuenta en futuros proyectos.

4. GLOSARIO

Miocardiopatía dilatada (MCD): enfermedad del músculo cardiaco que se caracteriza por la dilatación y disfunción sistólica del ventrículo izquierdo, con la reducción de la fuerza de contracción del miocardio.

Miocardiopatía dilatada familiar (MCDF): cuando dos o más miembros de una misma familia son diagnosticados con MCD idiopática.

NGS: acrónimo del inglés *Next Generation Sequencing* (secuenciación de nueva generación).

WES: acrónimo del inglés *Whole Exome Sequencing.* Técnica genómica para secuenciar todas las regiones de genes que codifican proteínas de un genoma (también conocido como exoma).

SNP: acrónimo del inglés Single Nucleotide Plymorphism (Polimorfismo de nucleótido simple).

SNV: acrónimo del inglés Single Nucleotide Variant (Variante de nucleótido simple). A diferencia de los SNPs carece de limitaciones de frecuencia.

genomAD: acrónimo del inglés *Genome Aggregation Database.* Base de datos internacional que agrega datos de secuenciación a gran escala.

Frecuencia alélica: proporción que se observa de un alelo específico respecto al conjunto de los que pueden ocupar un locus determinado en la población.

Predictores de patogenicidad: algoritmos genéticos que predicen el posible impacto de las sustituciones aminoacídicas en la estructura y función de las proteínas.

KEGG: acrónimo de *Kyoto Encyclopedia of Genes and Genomes.* Colección de bases de datos en línea de genomas, rutas enzimáticas, y químicos biológicos.

GO: acrónimo en inglés de *Gene Ontology***.** El proyecto de ontología génica provee un vocabulario controlado que describe el gen y los atributos del producto génico en cualquier organismo.

OTP: acrónimo en inglés de *Open Targets Platform.* Base de datos para el acceso y la visualización de posibles objetivos farmacológicos asociados con la enfermedad.

PMSA: acrónimo del inglés *Protein Multiple Sequence Alignment* (alineación de secuencia múltiple de proteínas).

5. BIBLIOGRAFÍA

1. Dec GW, Fuster V: **Idiopathic dilated cardiomyopathy**. *N Engl J Med* 1994, **331**(23):1564-1575.
2. Codd MB, Sugrue DD, Gersh BJ, Melton LJ, 3rd: **Epidemiology of idiopathic dilated and hypertrophic cardiomyopathy. A population-based study in Olmsted County, Minnesota, 1975-1984**. *Circulation* 1989, **80**(3):564-572.
3. Arimura T, Hayashi YK, Murakami T, Oya Y, Funabe S, Arikawa-Hirasawa E, Hattori N, Nishino I, Kimura A: **Mutational analysis of fukutin gene in dilated cardiomyopathy and hypertrophic cardiomyopathy**. *Circ J* 2009, **73**(1):158-161.
4. Burkett EL, Hershberger RE: **Clinical and genetic issues in familial dilated cardiomyopathy**. *J Am Coll Cardiol* 2005, **45**(7):969-981.
5. Sivasankaran S, Sharland GK, Simpson JM: **Dilated cardiomyopathy presenting during fetal life**. *Cardiol Young* 2005, **15**(4):409-416.
6. Judge DP: **Use of genetics in the clinical evaluation of cardiomyopathy**. *JAMA* 2009, **302**(22):2471-2476.
7. Dellefave L, McNally EM: **The genetics of dilated cardiomyopathy**. *Curr Opin Cardiol* 2010, **25**(3):198-204.
8. Hershberger RE, Morales A, Siegfried JD: **Clinical and genetic issues in dilated cardiomyopathy: a review for genetics professionals**. *Genet Med* 2010, **12**(11):655-667.
9. Hershberger RE, Siegfried JD: **Update 2011: clinical and genetic issues in familial dilated cardiomyopathy**. *J Am Coll Cardiol* 2011, **57**(16):1641-1649.
10. Ackerman MJ, Priori SG, Willems S, Berul C, Brugada R, Calkins H, Camm AJ, Ellinor PT, Gollob M, Hamilton R *et al*: **HRS/EHRA expert consensus statement on the state of genetic testing for the channelopathies and cardiomyopathies this document was developed as a partnership between the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA)**. *Heart Rhythm* 2011, **8**(8):1308-1339.
11. Hershberger RE, Parks SB, Kushner JD, Li D, Ludwigsen S, Jakobs P, Nauman D, Burgess D, Partain J, Litt M: **Coding sequence mutations identified in MYH7, TNNT2, SCN5A, CSRP3, LBD3, and TCAP from 313 patients with familial or idiopathic dilated cardiomyopathy**. *Clin Transl Sci* 2008, **1**(1):21-26.
12. Hershberger RE, Norton N, Morales A, Li D, Siegfried JD, Gonzalez-Quintana J: **Coding sequence rare variants identified in MYBPC3, MYH6, TPM1, TNNC1, and TNNI3 from 312 patients with familial or idiopathic dilated cardiomyopathy**. *Circ Cardiovasc Genet* 2010, **3**(2):155-161.

13. Herman DS, Lam L, Taylor MR, Wang L, Teekakirikul P, Christodoulou D, Conner L, DePalma SR, McDonough B, Sparks E *et al*: **Truncations of titin causing dilated cardiomyopathy**. *N Engl J Med* 2012, **366**(7):619-628.
14. Piran S, Liu P, Morales A, Hershberger RE: **Where genome meets phenome: rationale for integrating genetic and protein biomarkers in the diagnosis and management of dilated cardiomyopathy and heart failure**. *J Am Coll Cardiol* 2012, **60**(4):283-289.
15. Hershberger RE, Hedges DJ, Morales A: **Dilated cardiomyopathy: the complexity of a diverse genetic architecture**. *Nat Rev Cardiol* 2013, **10**(9):531-547.
16. McNally EM, Golbus JR, Puckelwartz MJ: **Genetic mutations and mechanisms in dilated cardiomyopathy**. *J Clin Invest* 2013, **123**(1):19-26.
17. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci U S A* 1977, **74**(12):5463-5467.
18. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.
19. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
20. Teer JK, Mullikin JC: **Exome sequencing: the sweet spot before whole genomes**. *Hum Mol Genet* 2010, **19**(R2):R145-151.
21. Rabbani B, Tekin M, Mahdieh N: **The promise of whole-exome sequencing in medical genetics**. *J Hum Genet* 2014, **59**(1):5-15.
22. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nat Protoc* 2009, **4**(7):1073-1081.
23. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.
24. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A: **Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral**. *J Med Genet* 2006, **43**(4):295-305.
25. Tang H, Thomas PD: **PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation**. *Bioinformatics* 2016, **32**(14):2230-2232.

26. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D *et al*: **REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants**. *Am J Hum Genet* 2016, **99**(4):877-885.
27. Choi Y, Chan AP: **PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels**. *Bioinformatics* 2015, **31**(16):2745-2747.
28. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C: **FATHMM-XF: accurate prediction of pathogenic point mutations via extended features**. *Bioinformatics* 2018, **34**(3):511-513.
29. Schwarz JM, Cooper DN, Schuelke M, Seelow D: **MutationTaster2: mutation prediction for the deep-sequencing age**. *Nat Methods* 2014, **11**(4):361-362.
30. Kobayashi Y, Yang S, Nykamp K, Garcia J, Lincoln SE, Topper SE: **Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation**. *Genome Med* 2017, **9**(1):13.
31. Maffucci P, Bigio B, Rapaport F, Cobat A, Borghesi A, Lopez M, Patin E, Bolze A, Shang L, Bendavid M *et al*: **Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis**. *Proc Natl Acad Sci U S A* 2019, **116**(3):950-959.
32. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Res* 2016, **44**(D1):D457-462.
33. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: **KEGG: new perspectives on genomes, pathways, diseases and drugs**. *Nucleic Acids Res* 2017, **45**(D1):D353-D361.
34. Gene Ontology C: **The Gene Ontology project in 2008**. *Nucleic Acids Res* 2008, **36**(Database issue):D440-444.
35. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions**. *Genome Res* 2001, **11**(5):863-874.
36. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: **Information content of binding sites on nucleotide sequences**. *J Mol Biol* 1986, **188**(3):415-431.
37. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function**. *Genome Res* 2003, **13**(9):2129-2141.
38. Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV: **Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods**. *Nucleic Acids Res* 2006, **34**(5):1317-1325.

39. Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB, Group IUGVW: **In silico analysis of missense substitutions using sequence-alignment based methods**. *Hum Mutat* 2008, **29**(11):1327-1336.

40. Michels M, Matte U, Fraga LR, Mancuso ACB, Ligabue-Braun R, Berneira EFR, Siebert M, Sanseverino MTV: **Determining the pathogenicity of CFTR missense variants: Multiple comparisons of in silico predictors and variant annotation databases**. *Genet Mol Biol* 2019, **42**(3):560-570.

41. Dorfman R, Nalpathamkalam T, Taylor C, Gonska T, Keenan K, Yuan XW, Corey M, Tsui LC, Zielenski J, Durie P: **Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene?** *Clin Genet* 2010, **77**(5):464-473.

42. Barnetson RA, Cartwright N, van Vliet A, Haq N, Drew K, Farrington S, Williams N, Warner J, Campbell H, Porteous ME *et al*: **Classification of ambiguous mutations in DNA mismatch repair genes identified in a population-based study of colorectal cancer**. *Hum Mutat* 2008, **29**(3):367-374.

43. Hicks S, Wheeler DA, Plon SE, Kimmel M: **Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed**. *Hum Mutat* 2011, **32**(6):661-668.

44. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J: **PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations**. *PLoS Comput Biol* 2014, **10**(1):e1003440.

45. Ghosh R, Oak N, Plon SE: **Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines**. *Genome Biol* 2017, **18**(1):225.

46. Bendl J, Musil M, Stourac J, Zendulka J, Damborsky J, Brezovsky J: **PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions**. *PLoS Comput Biol* 2016, **12**(5):e1004962.

47. Ernst C, Hahnen E, Engel C, Nothnagel M, Weber J, Schmutzler RK, Hauke J: **Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics**. *BMC Med Genomics* 2018, **11**(1):35.

48. Holland KD, Bouley TM, Horn PS: **Comparison and optimization of in silico algorithms for predicting the pathogenicity of sodium channel variants in epilepsy**. *Epilepsia* 2017, **58**(7):1190-1198.

49. Azevedo L, Mort M, Costa AC, Silva RM, Quelhas D, Amorim A, Cooper DN: **Improving the in silico assessment of pathogenicity for compensated variants**. *Eur J Hum Genet* 2016, **25**(1):2-7.

50. Leong IU, Stuckey A, Lai D, Skinner JR, Love DR: **Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations**. *BMC Med Genet* 2015, **16**:34.

51. Uys GM, Ramburan A, Loos B, Kinnear CJ, Korkie LJ, Mouton J, Riedemann J, Moolman-Smook JC: **Myomegalin is a novel A-kinase anchoring protein involved in the phosphorylation of cardiac myosin binding protein C**. *BMC Cell Biol* 2011, **12**:18.

52. Auer PL, Nalls M, Meschia JF, Worrall BB, Longstreth WT, Jr., Seshadri S, Kooperberg C, Burger KM, Carlson CS, Carty CL *et al*: **Rare and Coding Region Genetic Variants Associated With Risk of Ischemic Stroke: The NHLBI Exome Sequence Project**. *JAMA Neurol* 2015, **72**(7):781-788.

53. Brancati F, Dallapiccola B, Valente EM: **Joubert Syndrome and related disorders**. *Orphanet J Rare Dis* 2010, **5**:20.

54. Alizadeh AM, Mirzabeglo P: **Is oxytocin a therapeutic factor for ischemic heart disease?** *Peptides* 2013, **45**:66-72.

55. Gutkowska J, Jankowski M: **Oxytocin revisited: its role in cardiovascular regulation**. *J Neuroendocrinol* 2012, **24**(4):599-608.

56. Roy A, Guatimosim S, Prado VF, Gros R, Prado MA: **Cholinergic activity as a new target in diseases of the heart**. *Mol Med* 2015, **20**:527-537.

57. Yuan HX, Yan K, Hou DY, Zhang ZY, Wang H, Wang X, Zhang J, Xu XR, Liang YH, Zhao WS *et al*: **Whole exome sequencing identifies a KCNJ12 mutation as a cause of familial dilated cardiomyopathy**. *Medicine (Baltimore)* 2017, **96**(33):e7727.

6. ANEXOS

A continuación, se muestra el código de R utilizado para el análisis de los datos. Por simplicidad, en este caso mostramos es script para uno de los individuos, aunque se han realizado los mismos pasos para los cinco miembros de la familia.

FRECUENCIA ALÉLICA

Instalamos los paquetes necesarios para analizar la frecuencia alélica por gnomAD:

```
BiocManager::install("MafDb.gnomAD.r3.0.GRCh38")
require(MafDb.gnomAD.r3.0.GRCh38)
require(GenomicScores)

BiocManager::install("SNPlocs.Hsapiens.dbSNP149.GRCh38")
require(SNPlocs.Hsapiens.dbSNP149.GRCh38)
```

Analizamos la frecuencia alélica en uno de los individuos (ejemplo hermana)

```
ls("package:MafDb.gnomAD.r3.0.GRCh38")
mafdb <- MafDb.gnomAD.r3.0.GRCh38
snps <- hermana$dbsnp

rng <- snpsById(SNPlocs.Hsapiens.dbSNP149.GRCh38, ids=snps, ifnotfound="drop")
rng
frecuencia <- gscores(mafdb, rng, pop= c("AF_nfe", "AF"))
frecuencia <- as.data.frame(frecuencia)
frecuencia <- na.omit(frecuencia)
```

Nos quedamos con las columnas que nos interesan:

```
frecuencia <- frecuencia[, c("RefSNP_id", "AF_nfe", "AF")]
colnames(frecuencia)[1] <- "dbsnp"
```

Lo unimos con la base inicial:

```
Hermana <- merge(hermana, frecuencia, by="dbsnp")
hermana$dbsnp <- NULL
```

Filtramos las variantes por la frecuencia alélica:

```
frecuencia_comunes <- data2[which(data2$AF_nfe>0.01 | data2$AF_nfe == 0.01),]

frecuencia_raros <- data2[which(data2$AF_nfe<0.01),]
```

HERENCIA

El código será utilizado para la hermana y variantes comunes. Hay que tener en cuenta que este análisis se realizó con las variantes comunes y raras por separado.

Herencia autosómica dominante

Heredada:

En este ejemplo, unimos los datos de los parentales con los de la hija (padre y madre por separado), para ver los que tienen en común. Tenemos en cuenta las variantes presentes en la madre, en el padre o en ambos.

```
padrehija <- merge(hija,padre, by = "locus")
madrehija <- merge(hija,madre, by = "locus")
```

Ahora unimos ambas comparaciones.

```
Dominantes_heredados <- merge (padrehija, madrehija, all.x = T, all.y = T)
```

Mutaciones *de novo*:

Nos quedamos con aquellas variantes que no están en el padre ni en la madre:

```
Dominantes_de_novo <- hija$locus[!(hija$locus %in% Dominantes_heredados $locus)]
```

Herencia autosómica recesiva- heterocigosis simple

Heredada:

Unimos los datos de los parentales con los de la hija (padre y madre por separado), para ver los que tienen en común.

```
padrehija <- merge(hija,padre, by = "locus")
madrehija <- merge(hija,madre, by = "locus")
```

Ahora unimos ambas comparaciones, pero al hacerlo filtramos aquellas variantes que están presentes en ambos parentales.

```
Recesivos_heredados <- merge (padrehija, madrehija, by="locus")
```

Mutaciones *de novo:*

Debido a que es muy raro que se produzcan mutaciones en las dos copias, consideramos que una de las variantes ya está presente en uno de los dos parentales.

Filtramos las variantes que están presentes sólo en uno de los parentales

```
data <- madre$locus[!(madre $locus %in% padre $locus)]
data2 <- padre$locus[!(padre $locus %in% madre $locus)]
```

Unimos ambas bases:

```
Recesivos_de_novo <- merge (data, data2, all.x = T, all.y = T)
```

Herencia autosómica recesiva- heterocigosis compuesta

Heredada:

Filtramos por los genes que tienen más de una variante, es decir; los que, ya que estos son los que pueden ser heterocigotos compuestos.

```
hija_hetcompuesta <- hija[hija$gene %in% hija$gene[duplicated(hija$gene)],]
```

Seleccionamos en las bases del padre y de la madre estas variantes, y las unimos:

```
padhija_hetcompuesto <- padre[padre$locus[(padre$locus %in% hija_ hetcompuesta $locus)],]
madhija_ hetcompuesto <-  madre[madre$locus[(madre$locus %in% hija_hetcompuesta $locus)],]

madypad <- merge (padhija_hetcompuesto, madhija_ hetcompuesto, all.x = T, all.y = T)
```

Generamos una base con las variantes comunes entre padre y madre:

```
Común_madypad <- merge (padhijo, madhijo, by= "locus")
```

Unimos todas, obteniendo las variantes, provengan de la madre, del padre o de ambos:

```
Variant_def<- merge (hija_hetcompuesta, Común_madypad, all.x = T, all.y = T)
```

Finalmente, nos quedamos con los genes que tienen duplicados, ya que los que tienen una sola variante por gene no podrían tener herencia por heterocigosis compleja, ya que o sólo hay una variante en uno de los parentales, o bien, la variante es la misma en ambos.

```
Recesivos_heredados_compuesto <- variant_def[variant_def$gene %in% variant_def$gene[duplicated(variant_def$gene)],]
```

Mutaciones *de novo:*

Buscamos las variantes que están presentes en la hija de la base del apartado anterior, y no en los progenitores.

```
Recesivos_heredados__de_novo <- hija_hetcompuesta $locus[!(hija_hetcompuesta $locus %in% Recesivos_heredados_compuesto $locus)]
```

Herencia ligada al cromosoma X dominante:

Este caso se aplicaría a herencia dominante en la hermana, y a los hijos, ya que los hombres estarán afectados sea dominante o recesiva.

Filtramos por aquellos locus que se encuentran en el cromosoma X:

```
hija_X <- hija[which(hija$locus %like% "chrX"), ]
padre_X <- padre[which(padre$locus %like% "chrX"), ]
madre_X <- madre[which(madre$locus %like% "chrX"), ]
```

Heredada:

Unimos los datos de los parentales con los de la hija (padre y madre por separado), para ver los que tienen en común.

```
padrehija <- merge(hija_X,padre_X, by = "locus")
madrehija <- merge(hija_X,madre_X, by = "locus")
```

Ahora unimos ambas comparaciones:

```
Dominantes_heredados_X <- merge (padrehija, madrehija, all.x = T, all.y = T)
```

Mutaciones *de novo*:

Nos quedamos con aquellas variantes que no están en el padre ni en la madre:

```
Dominantes_de_novo _X <- hija_X $locus[!(hija_X $locus %in% Dominantes_heredados_X $locus)]
```

Herencia ligada al cromosoma X recesiva:

Heredada:

Unimos los datos de los parentales con los de la hija (padre y madre por separado), para ver los que tienen en común.

```
padrehija_X <- merge(hija_X,padre_X, by = "locus")
madrehija_X <- merge(hija_X,madre_X, by = "locus")
```

Ahora unimos ambas comparaciones, pero al hacerlo filtramos aquellas variantes que están presentes en ambos parentales.

```
Recesivos_heredados_X <- merge (padrehija_X, madrehija_X, by="locus")
```

Mutaciones *de novo:*

Debido a que es muy raro que se produzcan mutaciones en las dos copias, consideramos que una de las variantes ya está presente en uno de los dos parentales.

Filtramos las variantes que están presentes sólo en uno de los parentales

```
data <- madre_X $locus[!(madre_X$locus %in% padre_X$locus)]
data2 <- padre_X $locus[!(padre_X$locus %in% madre_X$locus)]
```

Unimos ambas bases:

```
Recesivos_de_novo_X <- merge (data, data2, all.x = T, all.y = T)
```

ENRIQUECIMIENTO

Gene Ontology

```
Install.packages("biomaRt")
library(biomaRt)
```

Definimos el objeto biomart:

```
mart <- useMart(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
```

Realizamos la consulta:

```
gene_ontology <- getBM(attributes = c("hgnc_symbol", "ensembl_gene_id", "protein_id","go_id", "namespace_1003", "name_1006"), filters = "hgnc_symbol", values = c("PDE4DIP", "KCNJ12", "WDR36", "AHI1"), mart = mart)
```

Seleccionamos de proceso biologicos aquellos relacionados con el músculo o el corazón:

```
Biological_process <- gene_ontology[which(gene_ontology$namespace_1003 == "biological_process"),]
install.packages("data.table")
library(data.table)
heart_muscle <- Biological_process[which(Biological_process$name_1006 %like% "heart" | Biological_process$name_1006 %like% "muscle" | Biological_process$name_1006 %like% "cardiac"), ]
```

KEGG Pathways

```
BiocManager::install("KEGG.db")
library(org.Hs.eg.db)
sym = c("PDE4DIP", "KCNJ12", "WDR36", "AHI1")
```

Obtenemos las IDs de Entrez gene asociadas a los símbolos de genes:

```
EG_IDs = mget(sym, revmap(org.Hs.egSYMBOL),ifnotfound=NA)
```

Buscamos toda la información disponible en la base KEGG para cada uno de los genes asociados a las variantes seleccionadas:

```
BiocManager::install("KEGGREST")
require(KEGGREST)

Kegg_PDE4DIP <- keggGet("hsa:9659")
Kegg_KCNJ12 <- keggGet("hsa:3768")
Kegg_WDR36 <- keggGet("hsa:134430")
Kegg_AHI1 <- keggGet("hsa:54806")
```

DIAGRAMA DE VENN

Sólo los hermanos:

```
grid.newpage()
draw.pairwise.venn(area1 = 44929, area2 = 45650, cross.area = 36110, category = c("II-1",
    "II-3"), fill = c("light blue", "dark blue"), alpha = rep(0.5, 2), cat.pos = c(0,
    0), cat.dist = rep(0.025, 2), scaled = FALSE)
```

Los hermanos y la hermana:

```
grid.newpage()
draw.triple.venn(area1 = 44929, area2 = 45829, area3 = 45650, n12 = 36489, n23 = 35307, n13 = 36110, n123
= 30985, category = c("II-1", "II-2", "II-3"), lty = "blank",
    fill = c("light blue", "pink", "dark blue"))
```