Universitat Oberta de Catalunya

uoc.edu

# Semi-automatización del Proceso de Clasificación de Variantes en Cáncer Hereditario

**Laura Arnaldo Orts**
Máster universitario de Bioinformática y Bioestadística
Área 2: Subárea 9: Bioinformática y Análisis de Datos Ómicos

**Dr. Jose Luis Mosquera** *(director UOC)*
**Dra. Lídia Feliubadaló** *(co-directora IDIBELL)*

Martes, 5 de enero de 2021

## FICHA DEL TRABAJO FINAL

| | |
|---|---|
| **Título del trabajo:** | *Semi-automatización del proceso de clasificación de variantes en cáncer hereditario* |
| **Nombre del autor:** | *Laura Arnaldo Orts* |
| **Nombre del consultor/a:** | *Jose Luis Mosquera* |
| **Nombre del PRA:** | *Lídia Feliubadaló* |
| **Fecha de entrega (mm/aaaa):** | 01/2021 |
| **Titulación:** | *Máster universitario de Bioinformática y Bioestadística* |
| **Área del Trabajo Final:** | *TFM-Bioinformática y Bioestadística Área 2 aula 1* |
| **Idioma del trabajo:** | *Castellano* |
| **Palabras clave** | *Semi-automatización, clasificación de variantes, cáncer hereditario* |

**Resumen del Trabajo**

La clasificación de variantes en cáncer hereditario se basa en la aplicación de un conjunto de evidencias de patogenicidad y benignidad establecidas por organismos internacionales. Actualmente, se determina manualmente el cumplimiento de estas evidencias, siendo un trabajo largo y tedioso. En este proyecto se ha desarrollado una herramienta que semi-automatiza este proceso. Este recurso recoge la información de 6 fuentes de acceso libre y la utiliza para calcular un conjunto de 11 evidencias. La herramienta consta de un documento en *.xlsx*, llamado *Plantilla*, y tres programas desarrollados en R: (1) Programa1 calcula las evidencias automatizables y proporciona un veredicto provisional de clasificación que se recoge en el fichero de clasificación de la variante, creado en base al documento *Plantilla*, (2) ProgramaBateria es una extensión de Programa1 para un listado de variantes, y (3) una vez el usuario ha revisado y completado el fichero de clasificación emitido por alguno de los anteriores programas, Programa2 (re)calcula las evidencias y emite un veredicto de clasificación mejorado. Se ha semi-automatizado el proceso para variantes del tipo: *silent, missense, nonsense, frameshift,* pequeñas deleciones e inserciones *in-frame* y variantes que afectan en al codón de inicio. En el caso de variantes que afectan al *splicing*, se ha automatizado el cálculo de la evidencia más compleja. También se ha elaborado un manual que resume las características, funciones y modo de uso del programa, a modo de guía para el usuario y soporte a futuras actualizaciones.

**Abstract**

The classification of variants in hereditary cancer is based on the application of a set of pathogenicity and benignity pieces of evidence, established by international organisations. Currently, the fulfillment of these pieces of evidence is manually determined, which is a long and tedious task. In this project, a tool that semi-automates this process has been developed. This resource collects the information from 6 open-access sources and uses it to calculate a set of 11 pieces of evidence. The tool consists of a document in *.xlsx*, called *Plantilla,* and three programs developed in R: (1) Programa1 calculates automatable pieces of evidence and provides a provisional classification outcome that is saved in the variant classification file, based on the *Plantilla* document, (2) ProgramaBateria is an extension of Programa1 for a list of variants, and (3) once the user has reviewed and completed the classification file created by one of the both previous programs, Program2 (re)calculates the pieces of evidence and provides an improved classification outcome. The process has been semi-automatized for silent, missense, nonsense, frameshift, start-codon variants and in-frame small insertions and deletions. In addition, a document to summarise the tool capabilities, functions, and how to use it has been written. It is intended as a user guide, as well as to provide support for future updates.

# Índice

## Lista de figuras

## Lista de Tablas

# 1. Introducción

## 1.1 Contexto y justificación del Trabajo

En el diagnóstico de cáncer hereditario, a partir de muestras de DNA de sangre del paciente, se secuencia y analiza una batería de genes para identificar variantes que incrementen el riesgo de cáncer [1].

| | Benign | | Pathogenic | | | |
|---|---|---|---|---|---|---|
| | **Strong** | **Supporting** | **Supporting** | **Moderate** | **Strong** | **Very Strong** |
| **Population Data** | MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2* | | | Absent in population databases *PM2* | Prevalence in affecteds statistically increased over controls *PS4* | |
| **Computational And Predictive Data** | | Multiple lines of computational evidence suggest no impact on gene /gene product *BP4*<br>Missense in gene where only truncating cause disease *BP1*<br>Silent variant with non predicted splice impact *BP7* | Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3* | Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5*<br>Protein length changing variant *PM4* | Same amino acid change as an established pathogenic variant *PS1* | Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1* |
| **Functional Data** | Well-established functional studies show no deleterious effect *BS3* | | Missense in gene with low rate of benign missense variants and path. missenses common *PP2* | Mutational hot spot or well-studied functional domain without benign variation *PM1* | Well-established functional studies show a deleterious effect *PS3* | |
| **Segregation Data** | Non-segregation with disease *BS4* | | Co-segregation with disease in multiple affected family members *PP1* | Increased segregation data → | | |
| **De novo Data** | | | | *De novo* (without paternity & maternity confirmed) *PM6* | *De novo* (paternity & maternity confirmed) *PS2* | |
| **Allelic Data** | | Observed in *trans* with a dominant variant *BP2*<br>Observed in *cis* with a pathogenic variant *BP2* | | For recessive disorders, detected in *trans* with a pathogenic variant *PM3* | | |
| **Other Database** | | Reputable source w/out shared data = benign *BP6* | Reputable source = pathogenic *PP5* | | | |
| **Other Data** | | Found in case with an alternate cause *BP5* | Patient's phenotype or FH highly specific for gene *PP4* | | | |

**Figura 1.** Esquema de clasificación de las evidencias propuesto por Richards, 2015. La tabla muestra la organización de los criterios en función del tipo de evidencia (benigna/patogénica) y fuerza (*supporting, moderate, strong, very strong*). BA, *benign stand-alone*; BS, *benign strong*; BP, *benign supporting*; PVS, *pathogenic very strong*; PS, *pathogenic strong*; PM, *pathogenic moderate*; PP, *pathogenic supporting*.

Una vez identificadas las variantes en el DNA, se utiliza la combinación de una serie de evidencias basadas en las pautas determinadas por el *College of Medical Genetics and Genomics (ACMG) y la Association for Molecular Pathology (AMP) (Figura 1)* para clasificar estas variantes. Estas pueden ser clasificadas como patogénicas, probablemente patogénicas, benignas, probablemente benignas o de significado incierto [2]. De este modo se determina si pueden ser o no la causa de la elevada predisposición al cáncer.

Hasta el momento, todo este proceso se realiza manualmente, consultando distintos recursos (e.g. aplicaciones, bases de datos, webs, publicaciones científicas…). En base a la información recopilada, se determina si se cumplen o no las diferentes

evidencias para cada una de las posibles variantes causantes de la enfermedad. En conjunto, todas las evidencias suman un total de 28, tanto benignas como patogénicas, teniendo cada una de ellas un mayor o menor peso [2]. Llevar a mano esta tarea es un proceso lento, tedioso e ineficiente.

Con la implantación de la secuenciación de nueva generación y los robots pipeteadores en la mayoría de laboratorios, la clasificación de variantes se ha convertido en el cuello de botella del diagnóstico genético [3]. Agilizar este proceso permitiría clasificar de manera automática las variantes, reducir el tiempo de diagnóstico de los pacientes, y tratar de manera más eficiente a los pacientes.

En este proyecto se pretende automatizar la mayor parte del proceso que se realiza para declarar si las variantes pueden ser la causa o no de la predisposición al cáncer del paciente y su familia.

**NON-CANCER**

| Population | Allele Count | Allele Number | Number of Homozygotes | Allele Frequency | Al. Freq. 99% |
|---|---|---|---|---|---|
| European (non-Finnish) | 37 | 102426 | 0 | 0,000361236 | 0,00026941 |
| European (Finnish) | 0 | 21612 | 0 | 0 | 0 |
| Latino | 80 | 34244 | 0 | 0,002336176 | 0,00192378 |
| African | 1 | 14840 | 0 | 6,73854E-05 | 3,4564E-06 |
| South Asian | 10 | 30516 | 0 | 0,000327697 | 0,00017779 |
| East Asian | 0 | 17692 | 0 | 0 | 0 |
| Ashkenazy Jewish | 101 | 9566 | 1 | 0,010558227 | 0,00889161 |
| Other | 12 | 5594 | 0 | 0,002145156 | 0,00123779 |
| | | | | | |
| Male | 137 | 129048 | 0 | 0,00106162 | 0,00091697 |
| Female | 104 | 107442 | 1 | 0,000967964 | 0,00081731 |

**Figura 2.** Captura del resultado de una de las tablas generadas por el programa desarrollado en las prácticas, para la consulta de la variante c.3024G>A del gen *BRCA1* en la base de datos de gnomADv2.1.

En las prácticas del máster, se empezó a desarrollar el proyecto bajo la dirección de la Dra. Lídia Feliubadaló en el marco del programa de investigación CIBERONC de "Tumores de Tracto Digestivo", coordinado por Gabriel Capellà, en el Institut d'Investigació Biomèdica de Bellvitge (IDIBELL). Se programaron algunas de las evidencias automatizables que requieren información de la base de datos gnomAD (e.g. PM2, BA1, BS1 o BS2) [4]. Brevemente, primero, se descargó la base de datos, tanto de exomas como de genomas. Segundo, se filtraron las regiones de los genes asociadas a cáncer hereditario. Tercero, la información seleccionada se registró en un documento *Excel*, llamado Plantilla. Cuarto, se generó una consulta para encontrar las frecuencias de cualquier variante de los genes seleccionados. El resultado se presenta en una de las pestañas del documento Plantilla *(Figura 2)*. Quinto, esta información era utilizada por el programa para calcular las citadas evidencias.

**Tabla 1.** Reglas para combinar los criterios de clasificación de variantes de secuencia según las guías definidas por Richards, 2015.

| Classification | | Criteria |
|---|---|---|
| **Pathogenic** | I | (i) 1 Very strong (PVS1) AND<br>(a) ≥1 Strong (PS1–PS4) OR<br>(b) ≥2 Moderate (PM1–PM6) OR<br>(c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) OR<br>(d) ≥2 Supporting (PP1–PP5) |
| | II | ≥2 Strong (PS1–PS4) OR |
| | III | 1 Strong (PS1–PS4) AND<br>(a)≥3 Moderate (PM1–PM6) OR<br>(b)2 Moderate (PM1–PM6) AND ≥2 Supporting (PP1–PP5) OR<br>(c)1 Moderate (PM1–PM6) AND ≥4 supporting (PP1–PP5) |
| **Likely Pathogenic** | I | (i) 1 Very strong (PVS1) AND 1 moderate (PM1–PM6) OR<br>(ii) 1 Strong (PS1–PS4) AND 1–2 moderate (PM1–PM6) OR<br>(iii) 1 Strong (PS1–PS4) AND ≥2 supporting (PP1–PP5) OR |
| | II | ≥3 Moderate (PM1–PM6) OR |
| | III | (v) 2 Moderate (PM1–PM6) AND ≥2 supporting (PP1–PP5) OR |
| | IV | (vi) 1 Moderate (PM1–PM6) AND ≥4 supporting (PP1–PP5) |
| **Likely Benign** | I | (i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) |
| | II | (ii) ≥2 Supporting (BP1–BP7) |
| **Benign** | I | (i) 1 Stand-alone (BA1) OR |
| | II | (ii) ≥2 Strong (BS1–BS4) |

En este trabajo, se ha pretendido actualizar, mejorar y modificar la automatización de las evidencias programadas hasta el momento. Además, se ha implementado la automatización de las consultas necesarias en los diferentes recursos (e.g. bases de datos, webs,…), que se utilizan para declarar si la variante cumple o no las diferentes evidencias automatizables que quedaban pendientes. Estas declaraciones y las del resto de evidencias no automatizables son combinadas según las reglas dictadas por la ACMG/AMP *(Tabla 1)*. Al final del proceso se obtiene el veredicto de clasificación, junto con toda la información que se ha utilizado y en la que se ha basado el programa para definir cada una de las evidencias analizadas.

Se trata de una herramienta eminentemente práctica y de traslación casi inmediata a la práctica clínica. Al ser una herramienta para el diagnóstico, se ha validado mediante su uso en un set de variantes previamente clasificadas [1].

Por otro lado, de aquí en adelante, la herramienta debe mantenerse útil y actualizable con los nuevos conocimientos y versiones de fuentes de información. Por ello, se han

proporcionado las versiones de las bases de datos utilizadas y se ha enriquecido el código con notas que permitirán a un usuario con conocimientos de programación mantener la herramienta útil y actualizada, así como poder integrarla en la base de datos de variantes que se utiliza en el departamento (Pandora). Asimismo, se han generado unas instrucciones concisas para los usuarios sin conocimientos informáticos (*Manual del usuario*).

## 1.2 Objetivos del Trabajo

A continuación, se presentan los objetivos generales y específicos del TFM.

### 1.2.1. Objetivos generales

- Programación y cálculo de las evidencias que se puedan automatizar, y volcado en una plantilla en formato *Excel* desarrollada para este propósito.
- Elaboración de un nuevo programa para el cálculo de la clasificación de la variante, si las evidencias proporcionadas en el programa inicial son modificadas por el usuario.
- Documentación del código y de las fuentes de información, y creación de un manual de uso.

### 1.2.2. Objetivos específicos

**1.1.** Revisar las evidencias programadas hasta el momento.

1.1.1. Optimizar el código de las evidencias programadas.

1.1.2. Completar algunas de las evidencias programadas.

**1.2.** Programar las evidencias restantes que son automatizables.

1.2.1. Programar el código para definir la evidencia PVS1.

1.2.2. Programar el código para definir las evidencias PP3 y BP4.

1.2.3. Programar el código para definir la evidencia BP7.

1.2.4. Programar el código para definir la evidencia PM1.

1.2.5. Programar el código para la evidencia PM4.

**1.3.** Generar el documento en el que se recogerá la información y clasificación.

**2.1.** Generar un nuevo programa que se utilizará si el usuario modifica o añade alguna de las evidencias.

**3.1.** Documentar las fuentes de información y cómo se accede a ellas, para facilitar las actualizaciones y las referencias en caso de publicaciones. Dotar el código de comentarios que permitan entender mejor aquello que sea necesario.

**3.2.** Generar un *Manual del Usuario* explicando cómo debe proceder el usuario para utilizar el programa.

## 1.3 Enfoque y método seguido

Este proyecto se puede dividir en tres grandes fases o bloques. En la primera fase, se revisaron las evidencias ya programadas. Por un lado, se completaron algunas de las evidencias para las que ClinGen [5] ha establecido oficialmente umbrales específicos para ciertos genes. Se tomaron los valores establecidos para esos genes y se asignaron umbrales conservadores para el resto. Para las evidencias PP3 y BP4, se valoraron el resto de predictores de proteína que proporciona la base de datos dbNSFP [6]. Como resultado de la revisión bibliográfica, se decidió utilizar solamente REVEL, considerando que era un buen predictor, debido a que posee umbrales gen-específico y generales calibrados clínicamente para que definieran unas probabilidades de patogenicidad de 0.2 (BP4, evidencia de benignidad) y 0.8 (PP3, evidencia de patogenicidad) [7].

Por otro lado, se optimizó el código para agilizar la ejecución del programa. Para las evidencias PS1 y PM5, se reescribió el código, dado que se cambiaba la fuente de obtención de la información. Se pasó de utilizar ClinVar Miner a utilizar el propio ClinVar, consiguiendo de este modo optimizar y agilizar el código (reducción de tiempo del 50%) [8] [9].

En la segunda fase, se procedió con la programación de las nuevas evidencias que quedaban por automatizar, comprobando conforme se iban finalizando que se ejecutaban y funcionaban de forma correcta, antes de proceder con las siguientes evidencias.

La primera de las evidencias que se ha programado ha sido PVS1. Esta fue programada para variantes *nonsense, frameshift,* con afectación del *splicing* y afectación del codón de inicio. Para las variantes *nonsense, frameshift* y de *splicing*, principalmente se utilizó como fuente de información Mutalyzer: Name Checker, ya que permite comprobar si se genera un codón *stop* prematuro [10]. En las que se ve afectado el codón de inicio, para empezar, se obtuvo la nomenclatura de Ensembl del transcrito canónico que proporciona dbNSFP para realizar la búsqueda y obtener la secuencia aminoacídica a través del UCSC Genome Browser [11] [12]. Finalmente, se utilizó la información obtenida de Simple ClinVar para determinar si hay variantes clasificadas como patogénicas antes de la segunda metionina [13].

Del archivo dbNSFP también se obtuvieron los valores de los algoritmos de conservación de nucleótidos (phyloP, phastCons y GERP++) para el cálculo de la evidencia BP7. PM4 se calcula solamente para pequeñas inserciones y deleciones *in-frame.* Por ello se comprobó volviendo a utilizar la herramienta Name Checker de Mutalyzer, que no se afecte la pauta de lectura y que no se genere ningún codón *stop* prematuro. Finalmente, se programó la evidencia PM1, utilizando la información de regiones críticas, dominios funcionales o *hot-spot* de aquellos genes en los que están definidas. También se programaron para todas las evidencias las correspondientes preguntas que el usuario deberá responder para que el Programa2 recalcule la clasificación de la variante.

Cabe destacar que durante la programación de la evidencia PVS1, se detectó que el archivo generado durante las prácticas con los datos de dbNSFP no era correcto. Brevemente, se observó que se había utilizado un genoma de referencia distinto del utilizado en el archivo *.BED* para filtrarlo. Con el fin de generar el nuevo archivo, se trabajó desde la línea de comandos de Linux, ya

que permitía procesar los archivos de dbNSFP de forma más rápida que R. Para mitigar el retraso en la planificación, se desarrolló paralelamente la evidencia PM4.

Paralelamente a las dos primeras fases del proyecto, se trabajó en la elaboración del documento que hace de Plantilla. Conforme se iban programando las diferentes evidencias y se sabía qué información se necesitaba para calcularlas, se fueron generando las diferentes pestañas y se les fue dando un formato adecuado para plasmar toda la información. La información para cada una de las variantes queda guardada en un documento independiente de clasificación con la fecha en la que se crea.

La tercera fase del trabajo consistió en el desarrollo de los otros dos programas (Programa2 y ProgramaBateria). En base a ciertas limitaciones que puede tener el Programa1 en el cálculo de algunas evidencias, se ha desarrollado el Programa2. Este recalcula la clasificación de la variante utilizando las aportaciones que el usuario podrá hacer en el documento de clasificación, después de ejecutar el Programa1. En función de los datos que muestre el documento de clasificación, las preguntas que le plantee y la información que pueda recoger en base a la literatura que le ha sugerido, recoge posibles cambios en las evidencias automatizadas y adición de nuevas evidencias por parte del usuario. También utiliza la información de las respuestas a las preguntas para (re)calcular evidencias (como las variantes de *splicing* para el cálculo de PVS1). Debido a su importancia, se empezó a programar desde el principio para que, fuera cual fuera el avance final del proyecto en la adición de nuevas automatizaciones, estuviera preparado para su uso.

Finalmente, se elaboró un nuevo programa (ProgramaBateria), que parte de un archivo de texto tabulado con diferentes variantes, para calcular y generar un documento con los datos y clasificación para cada una de ellas. Se ha utilizado este programa, para validar la clasificación de una lista de variantes previamente clasificadas manualmente. Así se ha asegurado el correcto funcionamiento del programa para su aplicación en diagnóstico.

## 1.4 Planificación del Trabajo

**Tareas**

Para objetivo 1.1.1 (optimizar PS1, PS1_Moderate y PM5):

- Para PS1, PS1_Moderate y PM5, investigar y programar cómo acceder y obtener la información directamente de ClinVar *(tiempo estimado: 5 días; tiempo empleado: 6 días)*.

- Revisar las evidencias programadas y saltar su cálculo cuando no apliquen según el tipo de variante para optimizar y agilizar el código *(tiempo estimado: 2 días; tiempo empleado: 1 día)*.

Para objetivo 1.1.2 (ampliar PP3/BP4, parte proteína):

- Definir los diferentes predictores de proteína que se utilizarán para PP3 y BP4, a partir de los que ofrece la base de datos de dbNSFP *(tiempo estimado: 2 días; tiempo empleado: 2 días)*.

- Investigar cómo conseguir la información de los predictores de proteína de la base de datos dbNSFP *(tiempo estimado: 2 horas; tiempo empleado: 1 día)*.

- Realizar una búsqueda bibliográfica para definir qué umbrales se utilizarán en cada caso *(tiempo estimado: 3 horas; tiempo empleado: 2 horas)*.

- Incluir umbrales gen-específicos para calcular las evidencias en aquellos genes en los que están definidos y adaptar el código *(tiempo estimado: 1 día; tiempo empleado: 2 días)*.

Para objetivo 1.2.1 (programar PVS1):

- Definir qué puntos de PVS1 son programables y automatizables *(tiempo estimado: 1 hora; tiempo empleado: 2 horas)*.

- Obtener las coordenadas de los exones de los transcritos canónicos de los genes *(tiempo estimado: 1 hora; tiempo empleado: 1 hora)*.

- Filtrar la base de datos y quedarse solamente con los genes utilizados en la batería de análisis para cáncer hereditario y con los transcritos canónicos de dichos genes *(tiempo estimado: 0.5 horas; tiempo empleado: 0.5 horas)*.

- Utilizar UCSC Genome Browser para obtener la secuencia aminoacídica de los genes de interés *(tiempo estimado: 2 horas; tiempo empleado: 1 día)*.

- Automatizar el uso de la herramienta Name Checker de Mutalyzer para saber si la variante genera o no un codón *stop (tiempo estimado: 2 días; tiempo empleado: 2 días)*.

- Programar el cálculo de PVS1 a partir de la información *(tiempo estimado: 7 días; tiempo empleado: 7 días)*.

Para objetivo 1.2.2 (completar PP3 y BP4; *splicing* + proteína):

- Definir un conjunto de cuestiones que el usuario responderá en la Plantilla y que proporcionarán la información necesaria para poder definir las evidencias PP3 y BP4 *(tiempo estimado: 1 día; tiempo empleado: 2 horas)*.

- Programar el cálculo PP3 y BP4 obteniendo la información directamente de la Plantilla *(tiempo estimado: 1 día; tiempo empleado: 1 hora)*.

Para objetivo 1.2.3 (programar BP7):

- Definir el predictor de conservación de nucleótido a utilizar para la evidencia BP7 *(tiempo estimado: 1 día; tiempo empleado: 3 horas)*.

- Investigar cómo obtener la información del predictor *(tiempo estimado: 2 días; tiempo empleado: 2 días)*.

- Programar el cálculo de BP7 a partir de la información *(tiempo estimado: 1 día; tiempo empleado: 1 día)*.

Para objetivo 1.2.4 (programar PM1):

- Incluir los umbrales/regiones en aquellos genes en los que están definidos *(tiempo estimado: 1 día; tiempo empleado: 1 día)*.

Para objetivo 1.2.5 (programar PM4):

- Definir la metodología para calcular la evidencia PM4: cómo comprobar que la deleción o inserción es *in-frame* y que se encuentra en una región no repetitiva *(tiempo estimado: 2 días; tiempo empleado: 2 días)*.

Para objetivo 1.3 (generar documento de clasificación):

- Generar un documento .xlsx que hará de Plantilla y que será rellenado por el programa *(tiempo estimado: 0.5 hora; tiempo empleado: 0.5 horas)*.

- Generar las diferentes pestañas en las que se irá recogiendo toda la información que se utiliza para determinar las diferentes evidencias *(tiempo estimado: 0.5 hora; tiempo empleado: 1 hora)*.

- Definir el formato de cada una de las pestañas para que recojan toda la información necesaria y sean lo más visuales posibles *(3 horas repartidas al final de cada evidencia, ya incluidas en el cálculo horario de cada una de ellas)*.

- Definir qué *inputs* se le proporcionarán al programa con la Plantilla para iniciar el proceso de clasificación de la variante *(tiempo estimado: 0.5 horas; tiempo empleado: 0.5 horas)*.

- Programar las diferentes vías para iniciar el análisis en función de los *inputs* que se le proporcionarán al programa relacionados con la variante que analizamos (nombre del gen, nombre de la variante, ...) *(tiempo estimado: 2 horas; tiempo empleado: 3 horas)*.

- Generar la pestaña donde se pedirá al usuario su *input* para aquellas evidencias que no sean automatizables y se podrá registrar la información en la que se basan los cálculos automatizados *(tiempo estimado: 1.5 horas; tiempo empleado: 3 horas)*.

- Generar la pestaña donde se mostrará el resumen del análisis y del veredicto *(tiempo estimado: 3 horas; tiempo empleado: 1 hora)*.

- Automatizar un párrafo modelo que se irá rellenando con la información usada para evaluar las evidencias y que podrá ser utilizado para rellenar el apartado de interpretación del informe de diagnóstico *(tiempo estimado: 2 días; tiempo empleado: 1 hora)*.

- Programar el código para que, en caso de proporcionarle una lista de variantes, se genere para cada una de ellas un documento *(tiempo estimado: 2 días; tiempo empleado: 2 días)*.

- Validar su correcta ejecución y resultado utilizando una lista de variantes ya clasificadas *(tiempo estimado: 1 día; tiempo empleado: 1 día)*.

Para objetivo 2.1.:

- Programar el código para obtener la información del documento en el que se resumen las evidencias calculadas y las añadidas por el usuario *(tiempo estimado: 2 horas; tiempo empleado: 3 horas)*.

- Programar el código para recalcular de clasificación de la variante a partir de los valores del documento *(tiempo estimado: 3 horas; tiempo empleado: 1 día)*.

- Programar el código para mostrar el nuevo veredicto de clasificación en el documento Plantilla y grabarlo como un archivo nuevo *(tiempo estimado: 1 hora; tiempo empleado: 0.5 hora)*.

Para objetivo 3.1 (documentar las fuentes y comentar el código):

- Dotar el código de comentarios que permitan entender mejor aquello que sea necesario *(tiempo estimado: 2 horas; tiempo empleado: 2 horas)*.

- Documentar las fuentes de información y cómo se accede a ellas, para facilitar actualizaciones, y las referencias en caso de publicaciones *(tiempo estimado: 1 hora; tiempo empleado: 1 hora)*.

Para objetivo 3.2 (generar *Manual del Usuario*):

- Redactar los pasos que debe realizar el usuario para hacer funcionar el programa y cómo debe de utilizarlo *(tiempo estimado: 1 día; tiempo empleado: 1 día)*.

- Definir el formato en el que se debe declarar la variante o los diferentes datos que se proporcionarán como *input* al programa *(tiempo estimado: 0.5 hora; tiempo empleado: 0.5 horas)*.

- Explicar para qué servirán cada uno de los documentos que vendrán con el programa *(tiempo estimado: 3 horas; tiempo empleado: 2 horas)*.

**1. Tareas realizadas no previstas en el plan de trabajo**

- Obtención de un listado de citaciones bibliográficas en las que aparece la variante a partir de ClinVar. Aunque no es una información utilizada específicamente para ninguna de las evidencias automatizadas, puede tener mucha relevancia a nivel informativo para el usuario, ya que le permitirá completar si lo necesita aquellas evidencias en las que se requiere una búsqueda bibliográfica (funcionales, clínicas, *de novo*, de cosegregación, etc.).

- Recálculo del intervalo de confianza de las frecuencias alélicas para las evidencias BA1 y BS1. El cálculo de las evidencias BA1 y BS1, utiliza el intervalo de confianza de las frecuencias alélicas que se obtienen de la muestra poblacional de gnomAD. Hasta el momento, este intervalo de confianza (IC) se estimaba en base al estadístico de Chi-cuadrado de la frecuencia alélica. Sin embargo, al revisar y verificar las evidencias programadas, se observó que este intervalo debía calcularse realizando una distribución de Poisson del recuento de alelos con esta variante, respecto al número total de alelos [14].

- Recálculo de la evidencia PM2. Hasta el momento para dar el veredicto del cumplimiento de la evidencia PM2, también se utilizaba el IC calculado de la frecuencia alélica. Sin embargo, al revisar la normativa general y específica de los genes, no se utiliza la frecuencia al 99%, sino directamente el valor de la frecuencia alélica. Además, para aquellos casos en que la variante no se haya detectado en una de las dos bases de datos (genomas o exomas) y, por tanto, no se halle la información completa del número de alelos secuenciados, se utilizará como alelos totales aquellos mostrados para la variante más cercana con información para las dos bases de datos.

- Programar los párrafos que aparecen en la columna de comentarios de la pestaña *"Evidence"* a modo de resumen informativo de la evidencia. Facilita al usuario la comprobación del correcto funcionamiento del programa al decidir el cumplimiento de las evidencias.

- Generar el documento de variantes patogénicas o probablemente patogénicas de la web de Simple ClinVar de todos los genes utilizados en el panel de diagnóstico. Permite calcular PVS1 para variantes que afectan al codón de inicio.

- Rehacer el archivo que contenía la información de la base de datos dbNSFP. El documento se había generado filtrándose por la posición del genoma de referencia hg38 en vez del hg19. Se utiliza para calcular las evidencias PVS1 (solo para variantes con afectación en el codón de inicio), PP3, BP4 y BP7.

- Cambio en el método de guardado de la información, duplicando el documento Plantilla y generando un nuevo documento con el nombre de la variante, el gen y la fecha en que se realiza el análisis (<SYMBOL_variant_yyyy-mm-dd.xlsx>).

## Hitos

1. Verificación del correcto funcionamiento de las diferentes evidencias ya programadas.
2. Realización del programa que recalcula el veredicto (Programa2).
3. Realización y verificación de la evidencia PVS1.
4. Realización y verificación de las evidencias PP3 y BP4.
5. Realización y verificación de la evidencia BP7.
6. Realización y verificación de la evidencia PM4.
7. Realización y verificación de la evidencia BS3 y PS3.
8. Verificación del correcto funcionamiento utilzando una batería de variantes.
9. Realización del Manual de uso del programa.

### Calendario

En las figuras 3, 4, 5, 6, 7, 8, 9, 10 se muestra un Diagrama de Gantt con la temporalización de las tareas e hitos presentados en los apartados anteriores. En estas figuras se presentan tanto el cronograma que se planeó inicialmente y el cronograma final. Cuando ambos cronogramas empiezan a tener discordancia, se ha pasado a presentar el cronograma inicial en la parte superior de la imagen, de color rojo y con la inicial I al principio del nombre de la tarea.

.

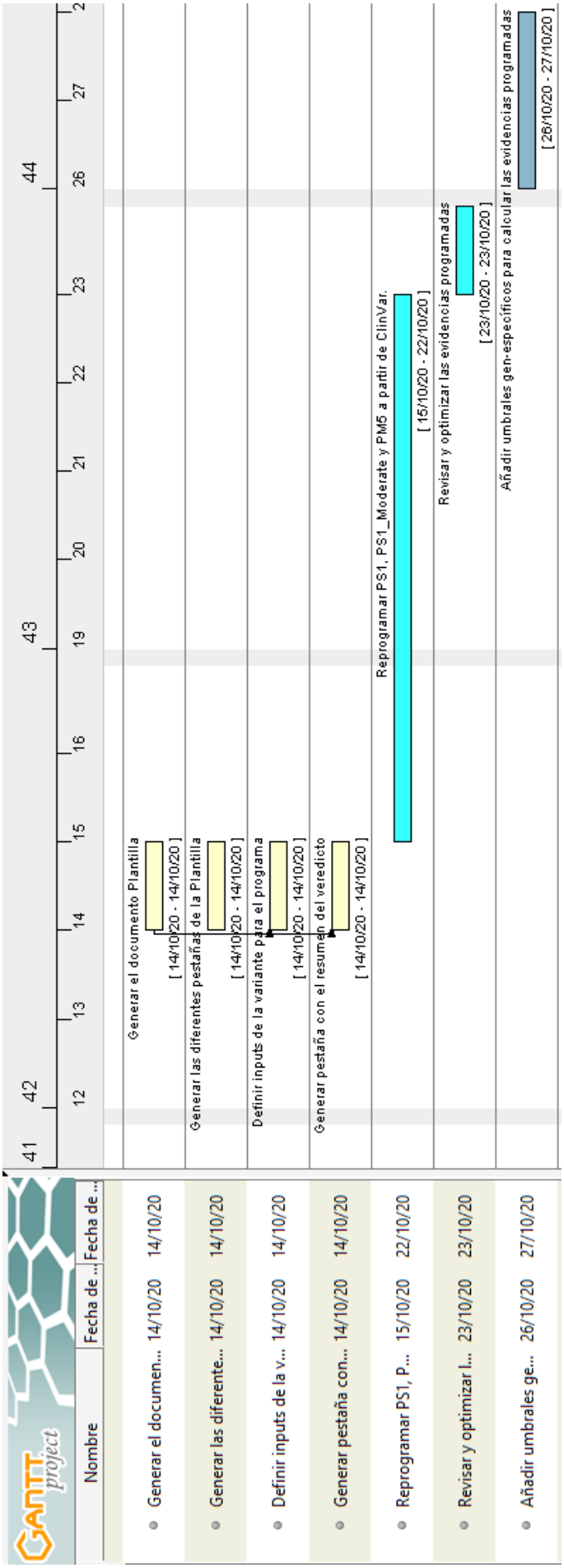



**Figura 3.** Diagrama de Gantt con las tareas que se completaron en las fechas previstas de la planificación inicial.

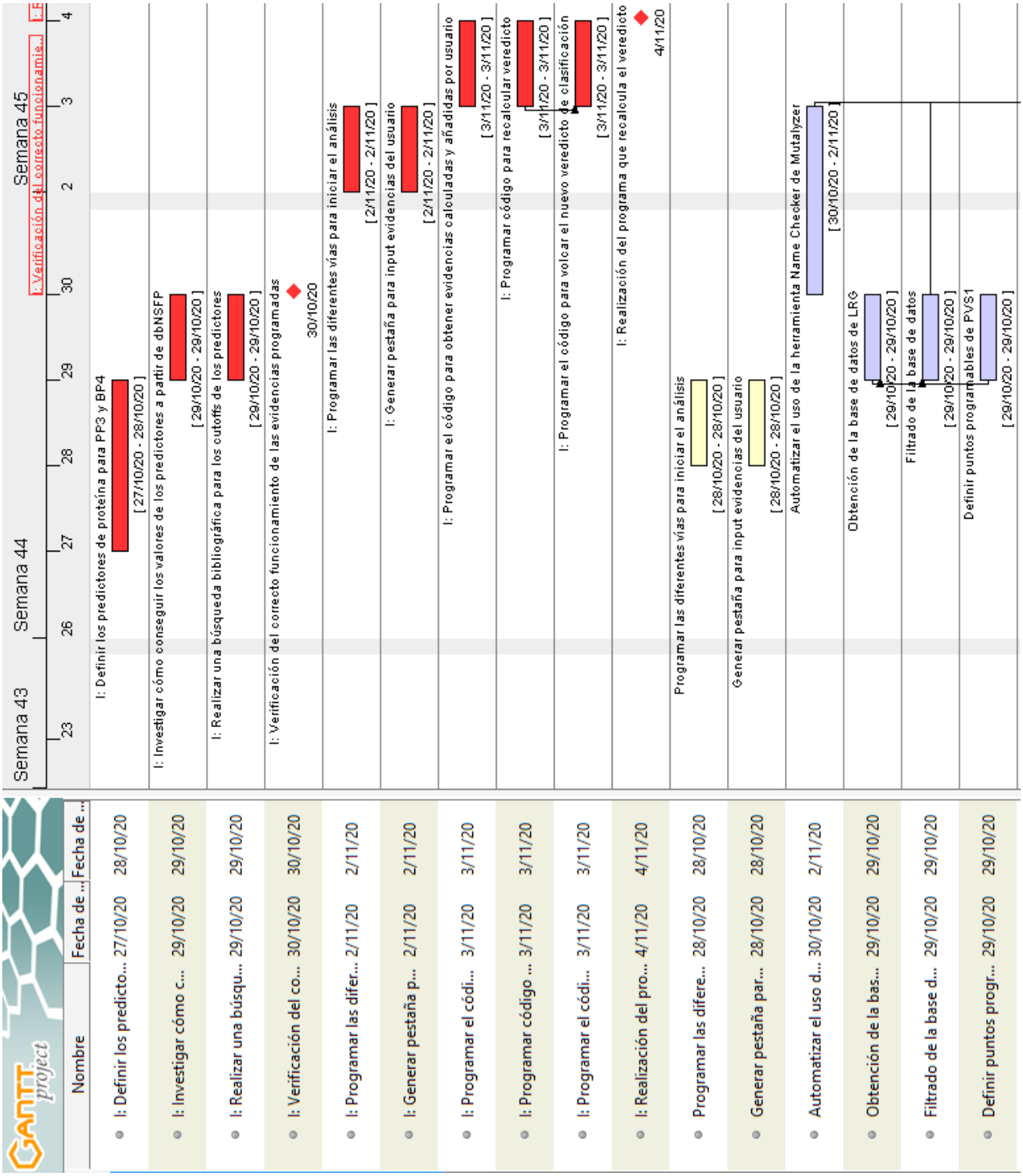



**Figura 4.** Diagrama de Gantt en la que se aprecian las primeras discordancias entre el cronograma establecido inicialmente (I, en rojo) y el cronograma realmente seguido.

**Figura 5.** Diagrama de Gantt en el que se observan las últimas tareas que debían tenerse realizadas y se han realizado hasta el cumplimiento de la primera fase de desarrollo

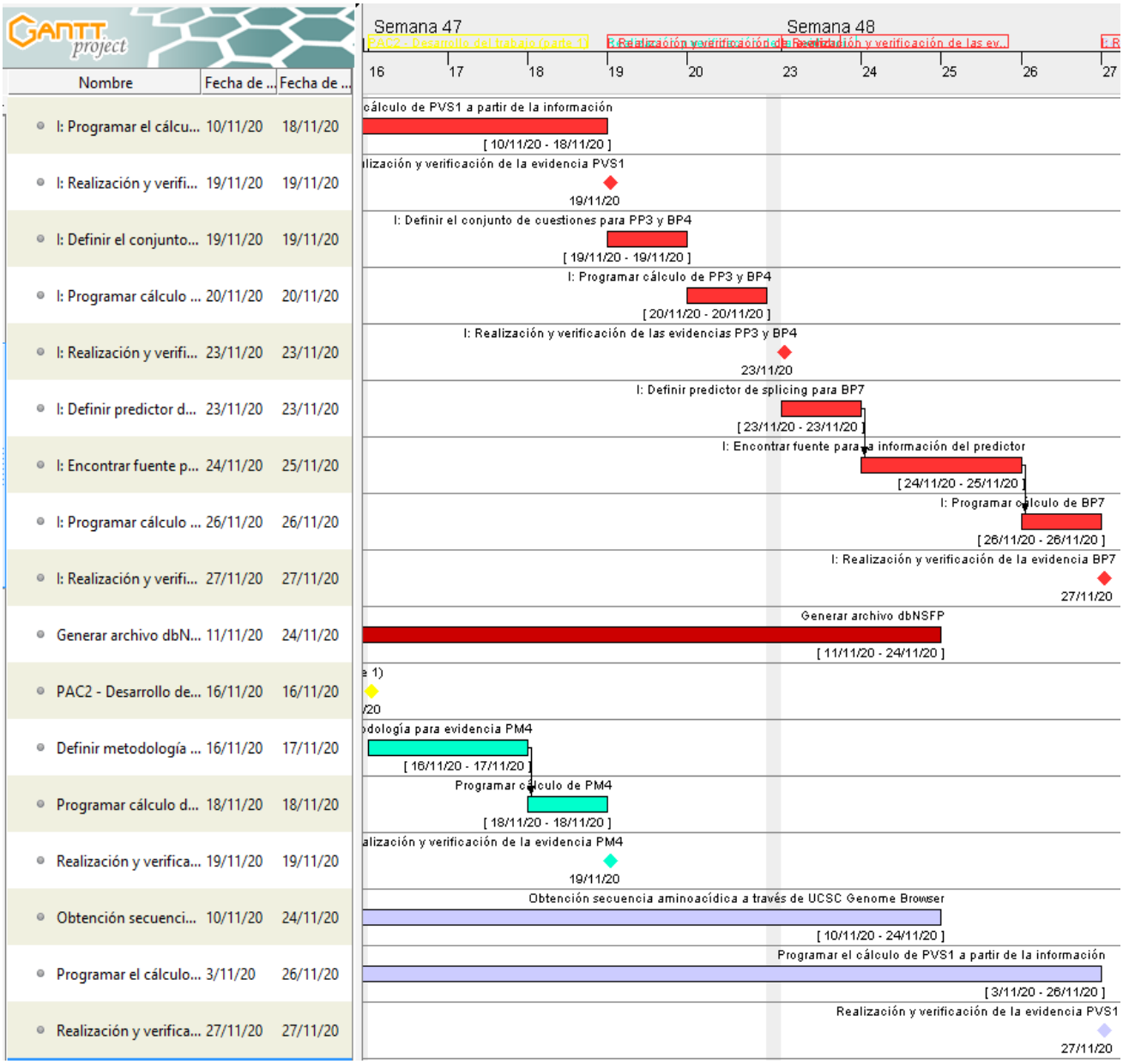



**Figura 6.** Diagrama de Gantt en el que se observa el tiempo que supuso la elaboración del archivo de la base de datos dbNSFP y cómo implico una segunda reestructuración del plan establecido.

Durante la primera fase de desarrollo *(figuras 3, 4, 5)*, se debían completar los hitos 1 y 2. Sin embargo, no se pudo finalizar ninguno de los dos. La desviación más importante fue no completar las evidencias PP3 y BP4. Este hecho se debió a la no-coincidencia en la disponibilidad horaria con diferentes especialistas del departamento, quienes eran clave para acabar de decidir qué predictores de proteína se debían utilizar para estas evidencias. Esto supuso que, para mitigar el problema, se decidiera ir avanzando en las tareas y objetivos que le seguían y no dependían de estas. Gracias a ello, se pudo avanzar más de lo esperado en el objetivo para programar la evidencia PVS1.

Durante la segunda fase *(figuras 6, 7, 8, 9)*, se debían completar los hitos 3, 4, 5, 6, 7 y 8. Sin embargo, se completaron los dos hitos de la primera fase, y los hitos 3, 4, 5 y 6. Se tuvo que modificar la realización de ciertos objetivos y tareas, porque se debió rehacer el archivo de la base de datos dbNSFP. Esto hizo que algunos objetivos no dependientes de esta tarea se llevasen a cabo

en paralelo a la construcción del archivo. Por contra, los objetivos y tareas dependientes (como las evidencias PP3 y BP4) se pospusieron.

Además, debido al tiempo que supuso la reelaboración de dicho archivo, se tuvo que descartar la realización del objetivo específico en el que se programaban las evidencias BS3 y PS3 por falta de tiempo. Por lo que no se ha podido completar el hito 7. La decisión de descartar este objetivo se debe a que solo iba a cumplir una función informativa para el usuario, proporcionándole un listado de citaciones que podría utilizar para responder a estas dos evidencias. Además, con la información que se consigue de ClinVar, para algunas evidencias ya se le proporciona el listado de citaciones que contiene esta web, cumpliendo esta función. También implicó que se tuvieran que posponer algunas tareas para después de la segunda fase de desarrollo, aunque supusiera tener que retrasar unos días la fecha prevista de la finalización del proyecto.

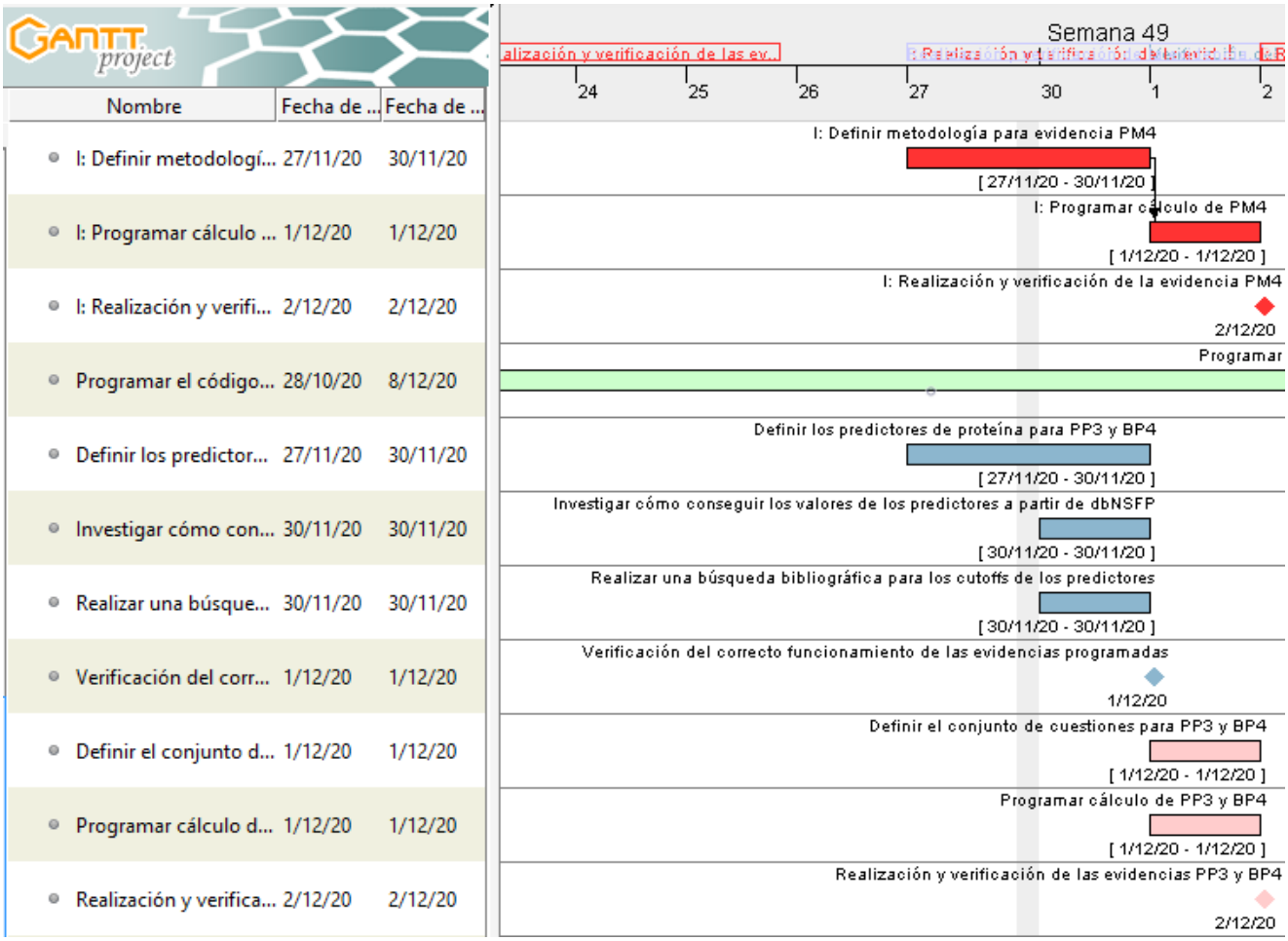



**Figura 7.** Captura de pantalla en el que se observa la realización de las evidencias PP3 y BP4 y, por lo tanto, la finalización del que era el primer hito.

**Figura 8.** Captura de pantalla en el que se observa el objetivo de las evidencias BS3 y PS3 en la planificación inicial, el cual se tuvo que descartar en este proyecto.

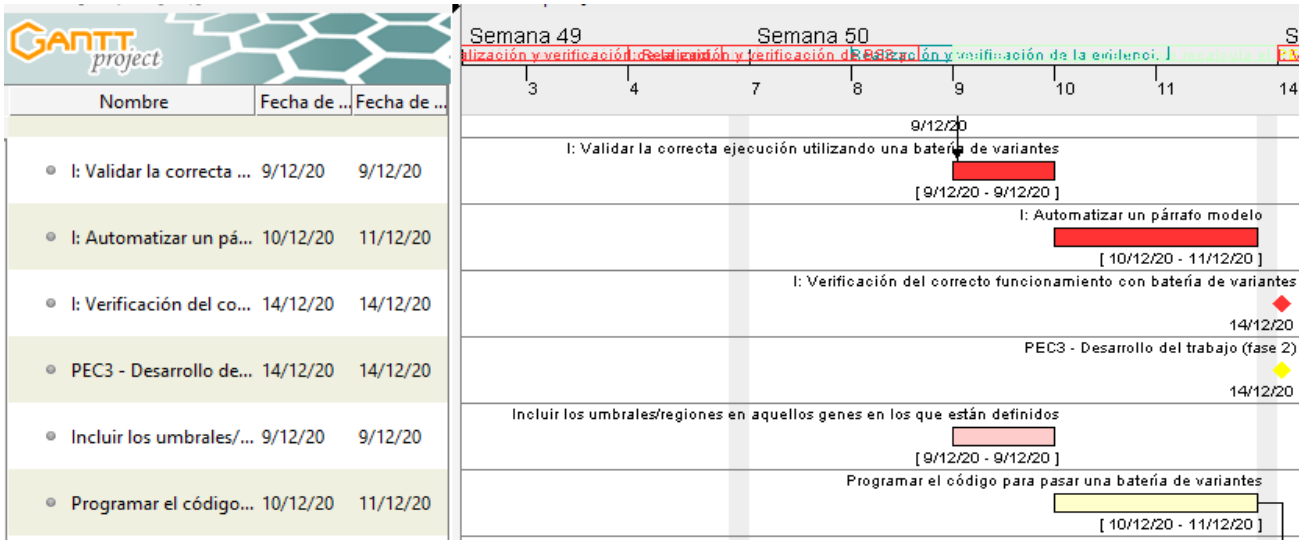



**Figura 9.** Captura de pantalla en el que se observan las últimas tareas que debían tenerse realizadas y se han realizado hasta el cumplimiento de la segunda fase de desarrollo.

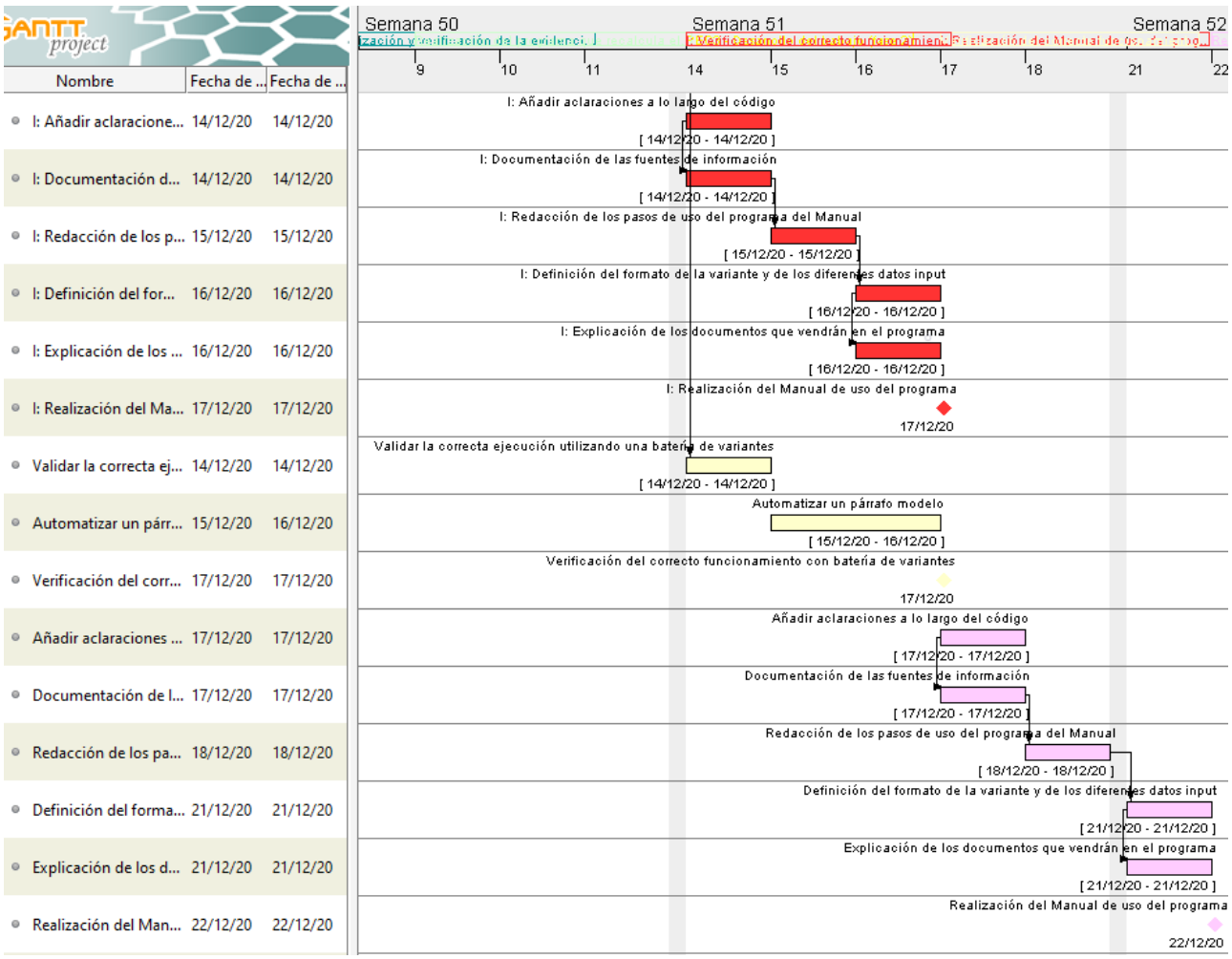



**Figura 10.** Captura de pantalla de los últimos días del proyecto, donde se observa la fecha prevista de finalización y la real.

Por otro lado, en ambas fases del desarrollo, se modificaron ciertas tareas en algunos objetivos, como las relacionadas con la creación del documento Plantilla y la programación del Programa2 que recalcula el veredicto después

del *input* del usuario, ya que se trataban de tareas que realmente dependían de cómo se fueran desarrollando el resto de evidencias que quedaban por programar. Por lo que se irían completando paralelamente a la realización de estas y al final del trabajo.

Finalmente, el hito 9 debía completarse en la siguiente semana de la entrega de la PEC3. Sin embargo, se decidió alargar unos días la fecha límite de la finalización del proyecto, para que durante esa semana y el lunes de la siguiente se completaran los hitos 8 y 9 y las tareas relacionadas con estos *(Figura 10)*.

En cuanto a las tareas realizadas que no estaban previstas, solamente se ha incluido la referente a la elaboración del archivo con la información de la base de datos dbSNFP *(Figura 5, 6)*, ya que fue la única que implicó una desviación real en la planificación original del trabajo.

## 1.5 Breve sumario de productos obtenidos

- **Plantilla.xlsx**: plantilla generada como documento *Excel*. En este archivo el programa procede a recoger toda la información que recolecta de las diferentes fuentes en relación a la variante, para poder dar respuesta al cumplimiento o no de las diferentes evidencias. También, en este archivo, se proporciona el veredicto de clasificación que el programa ha calculado en base al cumplimiento o no de las distintas evidencias.

- **Programa1.Rmd**: programa principal que procederá a calcular diferentes evidencias que están semi-automatizadas y dar un primer veredicto de clasificación de una variante. Está programado para calcular las evidencias PVS1 (para *nonsense, frameshift* y afectación del codón de inicio), PS1, PM1, PM2, PM4, PM5, PP3, BA1, BS1, BS2_Supporting, BP4 y BP7. Genera un documento de clasificación de la variante a partir de la Plantilla.

- **ProgramaBateria.Rmd**: programa alternativo al principal que a partir de un archivo proporcionado por el usuario en formato *.txt* con una lista de variantes, calcula la clasificación para cada una de ellas. Está programado para calcular las mismas evidencias que el Programa1.Rmd. Genera un documento .de clasificación para cada una de las variantes a partir de la Plantilla.

- **Programa2.Rmd**: programa secundario que procederá a recalcular algunas evidencias y el veredicto final de clasificación de la variante. Este programa solamente se corre una vez se ha corrido el Programa1.Rmd o el ProgramaBateria.Rmd. Se le proporciona el archivo de clasificación generado con uno de los dos programas anteriores y, con la información que proporciona el usuario, calcula PVS1 para variantes en las que se ve afectado el *splicing*, recalcula algunas evidencias, y da un veredicto de clasificación mejorado.

- **Manual del usuario.pdf**: documento con toda la información necesaria para el uso del programa para la semi-automatización de variantes en genes de cáncer hereditario.

- **Materiales suplementarios**: archivos procedentes de bases de datos que utiliza el programa para calcular las distintas evidencias. Todos ellos están filtrados por las regiones de los genes que se estudian en el diagnóstico del

cáncer hereditario (excepto el archivo con la información de la matriz BLOSUM62 [15]). Los archivos son:

- genoma.xlsx y exoma.xlsx: contienen la información de gnomAD de la base de datos de genomas y la de exomas, respectivamente.
- archivos con la estructura cromosomaXX_ord_filtrat.xlsx: existe uno para cada cromosoma (excepto del cromosoma 20 y 21) y contienen la información de la base de datos de dbNSFP.
- LRG_genes.txt: contiene la información de la base de datos de LRG de los transcritos canónicos de los genes [16].
- Start_Codon.xlsx: contiene la información de las variantes patogénicas o probablemente patogénicas con 2 o más estrellas según ClinVar que se encuentran en Simple ClinVar.
- BLOSUM62_probabilities.csv: contiene las puntuaciones de sustitución de la matriz BLOSUM62.

## 1.6 Breve descripción de los otros capítulos de la memoria

- **Capítulo 2**: se fundamentan los conceptos biológicos relacionados con el campo de la clasificación de variantes claves para el seguimiento del trabajo, como son los tipos de variantes que existen, la clasificación de variantes (ACMG/AMP), y los genes que se estudian en cáncer hereditario.

- **Capítulo 3**: se detallan los materiales que se han utilizado para la realización del proyecto, bases de datos, herramientas web y otros recursos, y los métodos aplicados.

- **Capítulo 4**: se presentan los resultados obtenidos en el proyecto: los archivos con la información de bases de datos que se utilizan en el programa, la creación de la Plantilla, el funcionamiento del Programa1, del ProgramaBateria y del Programa2, y la creación del Manual del usuario.

- **Capítulo 5**: se enumeran las conclusiones que se han obtenido, se realiza un análisis crítico del trabajo realizado y se plantean las posibles líneas futuras para el desarrollo de la herramienta en la clasificación de variantes para el diagnóstico de cáncer hereditario.

- **Capítulo 6**: se listan los acrónimos técnicos necesarios para la comprensión del texto.

- **Capítulo 7**: se enumeran las diferentes fuentes bibliográficas consultadas para la realización de este proyecto.

- **Capítulo 8**: se recoge aquella información que debido a su extensión o su relevancia no se ha incluido en el trabajo principal.

# 2. Contexto biológico

## Tipos de variantes

Se conocen diferentes tipos de variantes genéticas. Las más comunes son las sustituciones. Se tratan de variantes que han sufrido un cambio a nivel de un único nucleótido. Dentro de las sustituciones, existen las variantes *silent*, las *missense* y las *nonsense*. En el caso de las *silent*, se tratan de variantes en las que el cambio nucleotídico genera un codón que codifica para el mismo aminoácido. De modo que existe un cambio a nivel de secuencia de DNA, pero no a nivel de secuencia proteica. En cuanto a las *missense*, se trata de variantes en las que el cambio nucleotídico genera un codón que codifica para un aminoácido distinto. En este caso, se genera un cambio tanto a nivel de secuencia de DNA como de proteína, pero no se afecta la longitud de la proteína. Finalmente, las variantes *nonsense* son variantes en donde el cambio nucleotídico genera un codón *stop* prematuro, provocando que el resto de aminoácidos que le seguirán no se codifiquen ni formen parte de la proteína, es decir, generando una proteína truncada [17].

También se pueden encontrar otro tipo de variantes puntuales que afectan a la longitud de la secuencia nucleotídica. Estas son las pequeñas deleciones e inserciones (INDELs). En el caso de las deleciones, como su nombre indica, se tratan de mutaciones en las que ha habido una eliminación de nucleótidos, cortando la secuencia nucleotídica. Mientras que las inserciones son variantes en las que ha habido una adición de nucleótidos, alargando la secuencia. Hay un tipo especial de inserciones llamadas duplicaciones, en las cuales los nucleótidos insertados corresponden a la duplicación de uno o un conjunto de nucleótidos de la secuencia adyacente. Dentro de ambos tipos, se pueden encontrar inserciones o deleciones *in-frame* e inserciones o deleciones *out of frame*. Las que se generan *in-frame* son inserciones o deleciones en las que el número de nucleótidos añadidos o delecionados son tres o múltiplo de tres, haciendo que no se altere la pauta de lectura. En cambio, en las que son *out of frame* el número de nucleótidos añadidos o delecionados no son tres ni múltiplo de tres, provocando una alteración de la pauta de lectura y, en muchos casos, la generación de un codón *stop* prematuro *(Figura 11)*. Este tipo de mutaciones que acaba generando el codón *stop* prematuro son las llamadas *frameshift* [17].

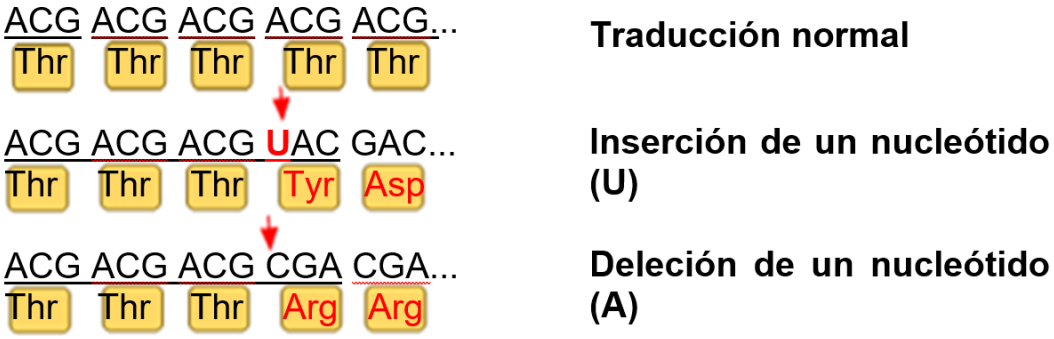



**Figura 11.** Ejemplo de cómo se ve afectada la traducción de las proteínas al alterar la pauta de lectura por una pequeña inserción o deleción.

Otro tipo de variante que existen son las variantes de *splicing*. Este tipo de variantes afectan al procesado normal del pre-mRNA, provocando que no se incorpore algún exón en el mRNA final o se incorpore alguna región intrónica. Las variantes de *splicing* canónicas son aquellas que afectan directamente a las regiones de reconocimiento de *splicing* ($\pm$1, $\pm$2) que se encuentran al inicio y final de cada exón. Pero el *splicing* también puede verse afectado por variantes en otras regiones del pre-mRNA que acaban provocando una alteración de este. La alteración del *splicing* puede acabar provocando la retención total o parcial de un intrón en la secuencia (*intron retention*) o la pérdida total o parcial de un exón (*exon skipping*). Debido a la adición o deleción de estos es posible que se vea alterada también la pauta de lectura y acabe generando un codón *stop* prematuro, o en el caso de la retención de un intrón incorpore la combinación de un codón *stop (Figura 12)*. De modo que en estos casos además se está generando una *frameshift* [18].

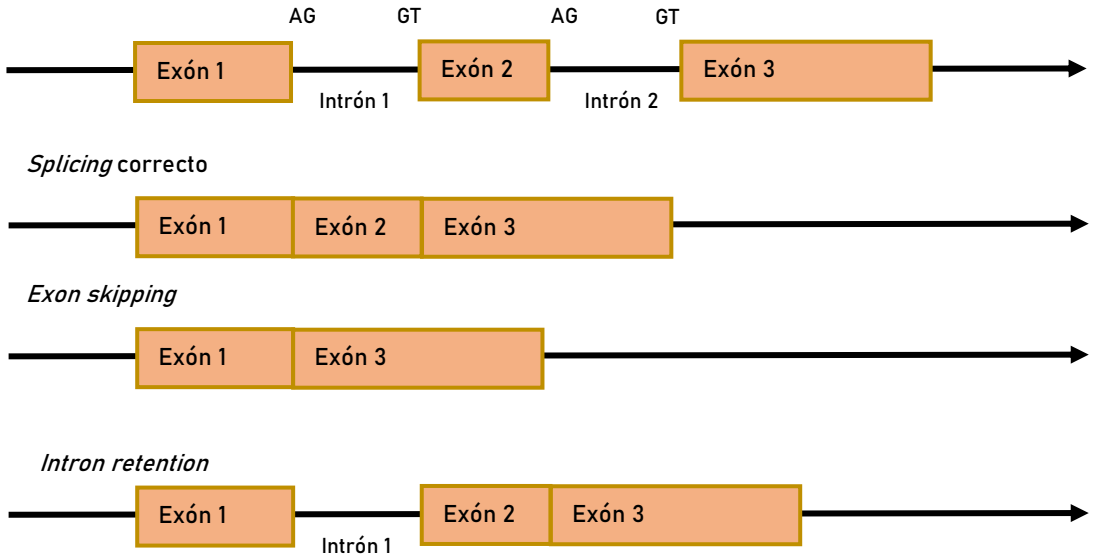



**Figura 12.** Ejemplo del mecanismo de *splicing*, donde se observa su funcionamiento cuando se realiza correctamente, cuando se deleciona un exón y cuando se añade un intrón.

Finalmente, se pueden producir afectaciones en el codón de inicio. Este tipo de variantes acaban afectando a la primera metionina de la secuencia que hace de codón de inicio, provocando que la maquinaria de transcripción del DNA empiece a transcribir a partir de la siguiente metionina que encuentre en la secuencia. Esto acaba provocando una deleción de los aminoácidos que se encontrarían antes de esta segunda metionina. La primera metionina puede verse afectada o bien por una mutación puntual directamente en cualquiera de los tres nucleótidos que forman el codón de inicio, o bien por una deleción que se produzca en esta región [19].

Otro tipo de variantes son las variantes intrónicas, que son las que se producirían en los intrones de los genes. Por falta de tiempo, se ha decidido no incorporarlas en este trabajo. También existen como variantes los grandes reordenamientos, que consisten en deleciones o duplicaciones de 1 o más exones. Este tipo de variantes, pese a formar parte del algoritmo de decisión de [20], también quedaron fuera de los objetivos de este trabajo. Ello se debe a que, por su naturaleza, aunque hay métodos para identificarlos, difícilmente se logran caracterizar completamente. Suelen ignorarse los puntos de ruptura, situados en zonas no secuenciadas con librerías que sólo cubren exones y regiones adyacentes, como las que se suele usar en diagnóstico de cáncer

hereditario. Esta limitación de conocimiento limita a su vez la capacidad de nombrarlas inequívocamente y de reconocerlas como iguales o no a otras descritas en la literatura y bases de datos. Todo ello limita su clasificación automática y por ello se decidió clasificarlas manualmente [21].

## Clasificación de variantes (ACMG/AMP)

En 2013, el American College of Medical Genetics and Genomics (ACMG) junto con la Asociación de Patología Molecular (AMP) se unieron para formular unas reglas o criterios para la clasificación de variantes en la mayoría de genes responsables de enfermedades hereditarias [2]. Hasta el momento no existían unas reglas generalizadas y la metodología que utilizaban los distintos laboratorios para determinar la clasificación de las variantes recaía sobre la interpretación del propio laboratorio. Esto dificultaba la comparación entre variantes, o que diferentes laboratorios clasificaran de forma distinta la misma variante.

En 2013, se establecieron unas terminologías estándar específicas de clasificación de variantes en: patogénicas, probablemente patogénicas, significado incierto, probablemente benignas y benignas. Además, describieron un proceso para clasificar las variantes en estas cinco categorías basado en unos criterios que utilizan diferentes tipos de evidencia (datos de población, datos computacionales y predictivos, datos funcionales, datos de segregación, datos *de novo*, datos alélicos) *(Figura 1)* [2]. De modo que algunos de estos criterios reflejan características que es más probable que cumplan variantes benignas, mientras que otros serán más específicos de variantes patogénicas. Por otro lado, los criterios quedan subclasificados en función del peso que tenga la evidencia. Es decir, hay ciertos criterios que tienen un peso superior, porque se trata de características que están muy vinculadas a patogenicidad o a benignidad; mientras que hay otras que la vinculación no es tan directa y se les da un peso proporcional a esta.

Más adelante ClinGen, que pertenecen al National Institutes of Health (NIH), observaron que estas reglas establecidas eran poco detalladas y específicas. Por ello, crearon comités para genes o enfermedades con el objetivo de poder desarrollar unas guías más específicas [5].

## Genes analizados en cáncer hereditario

Actualmente en el laboratorio de diagnóstico de cáncer hereditario del Institut d'investigació Biomèdica de Bellvitge (IDIBELL) en el que se ha realizado el TFM, se analiza un panel propio de 135 genes implicados en cáncer hereditario. Estos genes se pueden encontrar en el artículo Feliubadaló, 2017.

En el laboratorio, tras secuenciar los 135 genes, clasifican las variantes de aquellos genes relacionados con el fenotipo del paciente [23] y/o de su familia, y se informa de todas las variantes no benignas y de su interpretación.

Para la clasificación de las variantes, uno de los especialistas se encarga de clasificarlas y, posteriormente, otro se encarga de verificar su correcta clasificación. En promedio, este proceso toma unas dos horas y media de trabajo de un licenciado en ciencias de la salud para cada variante. Finalmente, una vez se encuentra la variante clasificada y comprobada, proceden a introducir la información de su clasificación en la base de datos del

departamento Pandora. Las variantes con clasificaciones anteriores a un año son reclasificadas tras la búsqueda de información actualizada.

### Tipos de predictores

Existen diferentes tipos de predictores, pero concretamente para el cálculo de algunas evidencias se utilizan de tres tipos diferentes. Estos son: predictores de proteína, predictores de *splicing* y algoritmos de conservación de nucleótidos [2].

Los predictores de proteína puntúan las variantes según su capacidad de alterar la función de la proteína, principalmente en función de la conservación filogenética del codón alterado y el cambio que produce en él. Para ello, normalmente estos son entrenados utilizando variantes ya clasificadas [7].

En cambio, los predictores de *splicing* evalúan la posibilidad de que se vea alterado el *splicing*, en función del reconocimiento de secuencias consenso de *splicing* en la región de referencia o alterada [24].

Finalmente, los algoritmos de conservación de nucleótidos evalúan la conservación del nucleótido afectado a partir de alineamientos del genoma humano con el de diversas especies.

# 3. Materiales y métodos

## 3.1 Materiales

### Bases de datos

- The Genome Aggregation Database (**gnomAD**) [4]: base de datos desarrollada por un conjunto de investigadores internacionales, que tienen como objetivo el poder unificar la información proporcionada por diferentes proyectos de secuenciación de exomas y genomas. Toda esta información se encuentra disponible para toda la comunidad científica y se encuentra en contenido descargable para los genomas de referencia GRCh37 (hg19) y GRCh38.

  En este proyecto se procedió a descargar la base de datos del genoma GRCh37 en su versión 2 (concretamente la versión 2.1.1), ya que es el genoma de referencia que utiliza el laboratorio. Esta contiene la información de 125.748 secuencias de exomas y 15.708 secuencias del genoma completo de individuos no relacionados secuenciados como parte de estudios genéticos poblacionales y algunos específicos de una enfermedad. El contenido descargable de exomas y genomas de la base de datos fue posteriormente filtrado, utilizando un archivo *.BED* con las posiciones de inicio y final de las regiones codificantes más 20pb (pares de bases) adyacentes de los genes que se analizan en el laboratorio.

  Esto se realizó durante el desarrollo de las prácticas. Primero se realizó desde el terminal de LINUX con el comando `VariantsToTable` de la herramienta `GATK` (versión 4.0.0.0), para pasar el archivo *.vcf* a archivo *.table* [25][26]. De modo que se le proporcionaba el archivo *.vcf* y *.BED* como *input,* se le indicaba las columnas de interés del *.vcf* y se generaba el archivo *.tabla* filtrado como *output*. En este caso, se seleccionaron los subconjuntos de pacientes *non-cancer* y *non-neuro* con las diferentes subpoblaciones: *European (non-Finnish), European (Finnish), Latino, African, South Asian* (solamente en la de exomas), *East Asian, Ashkenazi Jewish, Other*, y los datos para *male* y *female*. De cada una de ellas, se obtuvieron los valores del recuento de alelos con la variante (*Allele Count*, AC), número de alelos totales (*Allele Number*, AN), número de individuos homocigotos para la variante (*Number of Homozygotes*, nhomalt), y la frecuencia alélica de la variante (*Allele Frequency*, AF).

  Los archivos filtrados se guardaron en dos archivos de *Excel*, uno conteniendo la información de la base de datos de genomas filtrada *(*genoma.xlsx) y el otro con la información de la base de datos de exomas filtrada (exoma.xlsx).

- Locus Reference Genomic (**LRG**) [16]: base de datos que contiene la secuencia de referencia estable para variantes que han sido reportadas. Entre la información que se puede obtener de ella, se encuentra las nomenclaturas de los transcritos y proteínas de las variantes, así como las posiciones de inicio y de final de los exones de los distintos transcritos.

- Database of Human Nonsynonymous SNVs and their Functional Predictions (**dbNSFP**) [6]: base de datos desarrollada para la predicción funcional y la

anotación de todas las variantes puntuales no sinónimas (*non-synonymous single-nucleotide variants,* nsSNVs). Entre la información que se puede encontrar en ella hay: información referente a la variante, como la posición que ocupa en tres genomas de referencia distintos (hg19, hg18 y hg38), cromosoma en el que se encuentra, nucleótido y aminoácido de referencia y alternativo, símbolo del gen, anotación del cDNA y de la proteína…; información de gnomAD; información de ClinVar; y sobretodo información referente a diferentes valores de predictores de proteína y diferentes valores de algoritmos de conservación de nucleótidos. Entre los predictores de proteína se encuentran: SIFT, Polyphen2, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST4, MetaSVM, MetaLR, M-CAP, REVEL, MutPred, MVP, MPC, PrimateAI, DEOGEN, Aloft, CADD, DANN, fathmm_MKL, fathmm_XF, Eigen, GenoCanyon, GM12878, H1-hESC, HUVEC, LINSIGHT. Y entre los algoritmos de conservación se encuentran: GERP++, phyloP, phastCons, SiPhy.

- **ClinVar** [9]: portal web en el que se recoge información relacionada con variantes genéticas y su implicación en la salud humana. Entre la información que proporciona de interés para los criterios, se encuentra la clasificación de la variante (benigna, probablemente benigna, significado incierto, probablemente patogénica, patogénica) según los laboratorios remitentes y el *review status*, el cual proporciona el nivel de confianza que respalda la clasificación. El *review status* viene representado por estrellas de 0 a 4 significando en cada caso: 0 estrellas es que nadie ha proporcionado una interpretación; 1 estrella significa que un único remitente ha proporcionado una interpretación (*single submitter*) o que hay múltiples remitentes que han proporcionado la interpretación, pero en conflicto; 2 estrellas implica que dos o más remitentes han proporcionado una interpretación y coinciden; 3 estrellas son los casos en que la ha proporcionado un *expert panel*; y 4 estrellas cuando la proporciona una *practice guideline*.

- **Simple ClinVar** [13]: se trata de otro portal web cuyo objetivo es proporcionar estadísticas de genes y enfermedades. Para ello utilizan todas las variantes genéticas que se encuentran disponibles en ClinVar. A través de ella, permite seleccionar aquella información que te interesa (gen, significado clínico de las variantes, *review status*) y descargar todas las variantes que coincidan con tu filtrado.

### Herramientas web

- **Mutalyzer (*Name Checker*)** [10]: página web que permite generar o comprobar la nomenclatura de una variante respecto a varias referencias disponibles siguiendo la normativa de la Human Genome Variation Society (HGVS). Posee distintos tipos de herramientas para la comprobación, entre ellas la herramienta *Name Checker*. Esta herramienta parte de la descripción completa de la variante y comprueba que sea correcta. Sin embargo, su principal utilidad en este proyecto es que, además de comprobar la nomenclatura correcta de la variante, proporciona diferente información, entre ella si se acaba generando o no un *stop* prematuro en la secuencia.

## Otros recursos

- **BLOSUM62** [15]: matriz de conservación que se utiliza para comparar la variante que se está clasificando con las variantes que se encuentran en ClinVar que afectan al mismo codón y no son sinónimas. En este caso, esta matriz se encuentra guardada en un archivo *Excel* (BLOSUM62_probabilities.csv).

- **Documento Plantilla inicial**: documento en formato *.xlsx* desarrollado durante las prácticas, en el que se recogía la información de las evidencias programadas hasta entonces. Constaba de una pestaña en el que se almacenaría la información resumida de las evidencias que se cumplían y el veredicto de clasificación (aunque su rellenado no estaba programado todavía), una pestaña en la que se mostraba toda la información obtenida de gnomAD, una pestaña en donde recogía la información de ClinVar Miner en referencia a nuestra variante, una pestaña modelo utilizada para mostrar la información de las distintas variantes descritas en el mismo codón, extraída de ClinVar Miner, una pestaña en donde se recogería la información de los predictores de proteína (solamente se tenía REVEL), una pestaña en donde se rellenarían las evidencias que se fueran cumpliendo, y una pestaña en la que se haría el recuento de evidencias por tipo (no programada). ClinVar Miner era el portal web utilizado previamente en lugar de ClinVar.

- **Criterios ACMG/AMP** [2]: ACMG/AMP desarrolló un sistema de clasificación basado en un conjunto de evidencias. En total formuló los criterios para 28 evidencias diferentes, de las cuales 16 son patogénicas y 12 son benignas. Dentro de los dos subgrupos se dividen en función de la fuerza de la evidencia. De las patogénicas, hay 1 "*very strong*" (PVS1), 4 "*strong*" (PS1, PS2, PS3, PS4), 6 "*moderate*" (PM1, PM2, PM3, PM4, PM5, PM6) y 5 "*supporting*" (PP1, PP2, PP3, PP4, PP5). De las benignas, hay 1 "*stand-alone*" (BA1), 4 "*strong*" (BS1, BS2, BS3, BS4), y 7 "*supporting*" (BP1, BP2, BP3, BP4, BP5, BP6, BP7).

  De cada una de las evidencias, existen unas reglas para determinar su cumplimiento o no. Estas reglas están establecidas con unas formulaciones y umbrales a nivel general. Sin embargo, para 5 genes se han publicado reglas específicas para algunas evidencias. Estos genes son *CDH1, PTEN, ATM, CHEK2* y *TP53* [1][27][28][29][30]. Además, a pesar de tener una fuerza intrínseca descrita más arriba, se le puede dar otra fuerza (mayor o menor) si el usuario lo considera.

  Las evidencias PM2, BA1, BS1 dependen de la frecuencia alélica en la población. BS2 depende del número de homocigotos. De modo que estas cuatro evidencias se rellenarán con la información obtenida de gnomAD.

  La evidencia PVS1 depende de la generación de un codón prematuro en la secuencia y del desencadenamiento o no de NMD (*nonsense-mediated decay*). De modo que se utilizará la información proporcionada por Mutalyzer.

  Las evidencias PS1 y PM5 dependerán de la información de ClinVar. Para PS1 se utilizarán aquellas variantes *missense* que produzcan el mismo cambio de codón, mientras que para PM5 se utilizarán aquellas que

produzcan cambio a otro aminoácido evolutivamente más conservativo según la matriz BLOSUM62.

La evidencia PM1 se declara cuando la variante cae en regiones que son críticas para alguna funcionalidad del gen. Solamente son programables aquellos genes en los que estas regiones están bien establecidas.

La evidencia PM4 aplica para deleciones o inserciones pequeñas que se encuentran *in-frame*, especialmente si afectan a regiones críticas.

Las evidencias PP3 y BP4 dependen del resultado que determinen los predictores de proteína en relación a la patogenicidad de la variante. Para algunos genes se utilizan umbrales gen-específicos.

Finalmente, la evidencia BP7 depende del resultado que determinen los algoritmos de conservación de nucleótidos en relación a la conservación de la posición en la que se encuentra la variante de estudio.

El resto se tratan de evidencias no automatizables o que, debido al tiempo del que se disponía, no se podían incluir en este proyecto.

- **Árbol de decisión para PVS1**: se basa en el árbol de decisión que se recoge en el artículo Tayoun, 2018 *(Figura 13)*. Se detalla cómo calcular PVS1 en función del tipo de variante del que se parte (*nonsense, frameshift,* grandes reordenamientos (inserciones o deleciones), afectación del *splicing*, afectación del codón de inicio).

## 3.2 Métodos

- **Generación del archivo LRG_genes.txt**: con el fin de conocer las posiciones de inicio y final de los exones pertenecientes a los transcritos canónicos de los genes para la evidencia PVS1, se generó un archivo con la información del archivo descargado de la web LRG con el genoma de referencia GRCh37. Este archivo contiene información de la nomenclatura de los transcritos de cada gen en LRG, el símbolo del gen, el cromosoma en el que se encuentran, la cadena que lo codifica, las posiciones de inicio y final del transcrito, todas las coordenadas de inicio y final de los exones, la nomenclatura de la proteína en LRG, y las posiciones de inicio y final de la CDS. De este archivo solamente se seleccionaron las columnas de la nomenclatura del transcrito, el símbolo del gen, las posiciones de inicio y final de los exones y la nomenclatura de la proteína. Además, se seleccionaron solamente los genes que se estudian en cáncer hereditario.

- **Obtención archivo Start_Codon.xlsx**: se ha construido un archivo con la información de las variantes patogénicas o probablemente patogénicas de los diferentes genes descritos en ClinVar. Para ello se utilizó la web Simple ClinVar. En esta web se realizó la búsqueda de cada uno de los genes del panel. Se filtraron aquellas variantes con un significado clínico de patogénicas y probablemente patogénicas, y que además tuvieran un *review status* de *criteria provided multiple submitter no conflicts, reviewed by expert panel* y *practice guideline*, ya que en estos casos era cuando se le daba un peso de 2, 3 y 4 estrellas en ClinVar. Posteriormente, se descargó el archivo con ellas. Finalmente, se realizó la unión de los distintos archivos generados de todos los genes, utilizando el comando `cat`

desde la línea de comandos de Linux. El archivo Start_codon.txt fue pasado a *.xlsx* desde *Excel.*

- **Validación de variantes clasificadas**: se han utilizado un total de 40 variantes ya clasificadas en el departamento para el gen *ATM* [1]. Entre ellas hay variantes *silent, missense, nonsense, frameshift* (inserciones y deleciones) y una *missense* con afectación del *splicing*. No se ha escogido ninguna variante de *splicing* canónica, ya que este tipo de variante se produce en las posiciones +/-1 y +/-2, correspondiendo a regiones intrónicas. Por lo que, aunque se tiene programado el cálculo de PVS1 si se le proporciona el nombre de la variante en RNA, el código no está programado para calcular el resto de evidencias en estas variantes.

  Se generó un archivo de texto (*.txt*) con la información de las distintas variantes que se validarían. Este archivo se encuentra entre los materiales suplementarios como variants_validacio.txt, al igual que los archivos generados por el programa de todas las variantes en *.xlsx*.

- **R** (versión 3.6.1) [31]: lenguaje de programación estadístico de software libre utilizado principalmente para el cálculo computacional estadístico y la elaboración de gráficos, entre otras funciones. Entre ellas se encuentra el paquete `htmltab`, que permite obtener tablas de *links* web, o el paquete `RPostgreSQL` el cual permite trabajar con órdenes muy similares a las que se utiliza en lenguaje MySQL [32][33]. Gracias a ello permite trabajar desde R con la base de datos del departamento Pandora, para obtener información referente a la variante.

  La decisión de trabajar con lenguaje de programación R fue principalmente para poder generar un programa en formato Rmarkdown y facilitar la lectura al potencial usuario [34].

  RMarkdown (versión 2.1) permite generar archivos en el que utilizando la programación mediante *chunks* puede leer otros lenguajes, entre ellos R, y generar como *output* un archivo *html* o *pdf* fáciles de leer. Se encuentra disponible para su uso en el programa RStudio (versión 1.1.463). La razón principal por la que se trabajará con RMarkdown es porque permite tener de forma ordenada los diferentes apartados de los que consta el código, separados por títulos, y permite esconder aquella información con la que no se esté trabajando. Esto permite facilitar la lectura, localizar errores dentro del código y modificarlos cuando sea necesario.

- **Excel**: programa de hoja de cálculo que cuenta con diferentes herramientas para el cálculo, la elaboración de gráficos, generar tablas dinámicas e incluso un lenguaje de programación para aplicaciones (Visual Basic). Gracias al paquete `XLConnect` de R permite fácilmente mostrar la información en un documento *Excel*, de forma ordenada (por pestañas) y permite editarlo sin modificar el formato que se ha prediseñado [35]. Además, cuenta con una interfaz a la que la mayoría de potenciales usuarios están familiarizados, siendo muy visual y fácil de manejar. Esto se puede observar en la disposición de la información por pestañas, permitiendo tenerla de forma ordenada. En este caso, tener la información que el programa utiliza para cada evidencia en una pestaña.

  Además, la estructuración de una de sus pestañas ("*Evidence*") está pensada para facilitar el trabajo del usuario. Primero, solamente se le

proporcionan los criterios del gen que está analizando, para poder consultar fácil y rápidamente. Segundo, el programa genera un comentario, proporcionando la información principal que ha permitido declarar el cumplimiento de las diferentes evidencias. Gracias a ello, podrá verificar fácilmente el correcto funcionamiento del programa. Tercero, incorpora las columnas para declarar *met/unmet* y el peso del resto de evidencias no automatizables. Finalmente, el propio programa le genera una serie de cuestiones que el usuario deberá responder para que el Programa2 complete la clasificación con el cálculo o recálculo de algunas otras evidencias. Todo en conjunto, permite que el usuario solo interactúe con esta pestaña para verificar el funcionamiento y ejecutar el Programa2. Sin olvidar que, si la información del comentario no fuera suficiente, quisiera comprobar algún dato o usar la información bibliográfica o de variantes similares para añadir evidencias de clasificación, tiene a su disposición el resto de pestañas. Por todo lo expuesto, trabajar con este formato facilita realmente el trabajo del usuario.

Además, gracias al paquete `XLConnect` de R, plasmar la información en las diferentes pestañas es relativamente fácil.

- **Línea de comandos de Linux** (Ubuntu 18.0) [36]: se basa en un lenguaje de *scripting* que procede del *sh* utilizado en Unix, que es el *Bash*, permitiendo ejecutar *scripts,* ejecutar binarios, entre otras tareas. Además, permite poder procesar archivos grandes (seleccionar determinadas columnas, ordenarlos, filtrarlos por un valor en una columna, …) en un tiempo relativamente pequeño. Previamente se utilizó para el procesado de los archivos *.vcf*. En este proyecto se ha utilizado principalmente para trabajar con los archivos de la base de datos dbNSFP.

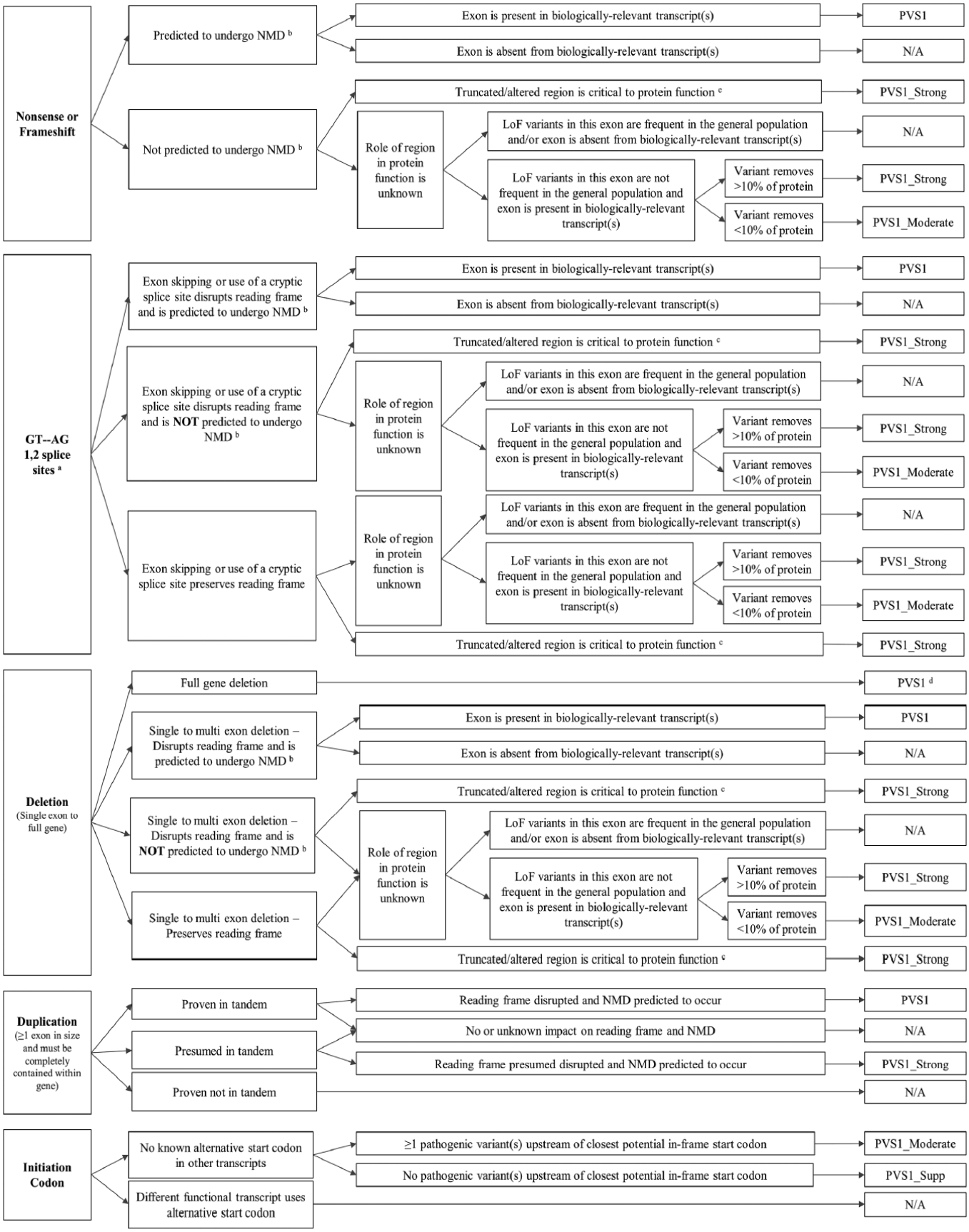



**Figura 13.** Árbol de decisión de PVS1 propuesto en el artículo de Tayoun, 2018.

# 4. Resultados

## 4.1 Obtención de archivos procedentes de bases de datos

Para generar el archivo **LRG_genes.txt** con la información necesaria para la evidencia PVS1, se ha desarrollado el código que se facilita en el Anexo 8.1. Se seleccionó manualmente el set mínimo de transcritos con exones implicados en la clínica, a partir de la lista facilitada por el laboratorio. Aunque en la mayoría de genes se trataba sólo de 1 (transcrito t1), en otros genes se tuvo que incluir el t2 u otros. El archivo resultante contiene la siguiente información: nomenclatura en LRG de los transcritos canónicos de los genes, el símbolo del gen, las posiciones de inicio y final de los exones, y la nomenclatura de la proteína.

Desde la línea de comandos de Linux, se generaron varios archivos filtrados a partir de **dbNSFP**. En el Anexo 8.2. se muestra el procedimiento seguido para obtener los archivos finales. Se obtuvo un archivo para cada cromosoma, a excepción de los cromosomas 20 y 21, ya que no existe ningún gen en cáncer hereditario que se encuentre en estos cromosomas. Los archivos resultantes contienen los predictores de proteína y algoritmos de conservación, y la nomenclatura de Ensembl. Estos archivos son necesarios para las evidencias BP4, PP3, BP7 y variantes en el inicio de codón para el cálculo de PVS1.

El archivo **Start_codon**.xlsx contiene la información de las variantes patogénicas y probablemente patogénicas descritas en ClinVar para los genes de cáncer hereditario. Contiene la información del *link* a la web de ClinVar, el símbolo del gen, el tipo de variante, la consecuencia, el significado clínico, el *review status*, el listado fenotípico, el nombre de la variante, el aminoácido de referencia y el alternativo, la posición aminoacídica, y el valor del predictor CADD. Este archivo se utiliza para obtener las variantes patogénicas o probablemente patogénicas que se encuentran antes de la segunda metionina de la secuencia. Permite calcular la evidencia PVS1 para variantes en las que se ve afectado el codón de inicio.

## 4.2 Generación de Plantilla

Partiendo de la Plantilla desarrollada en las prácticas (ver apartado 3.1), se han creado nuevas pestañas, y se ha modificado las ya existentes para proporcionar más información al usuario. La nueva versión se puede consultar en el documento Plantilla.xlsx aportado como material suplementario. Consta de 12 pestañas *(Figura 14)* y se detallan a continuación.

- Pestaña 1 (*Classification Summary*): recoge un resumen de las evidencias que se cumplen, la suma de las diferentes evidencias según su fuerza, el veredicto de clasificación y un párrafo resumen del por qué se cumplen las evidencias.

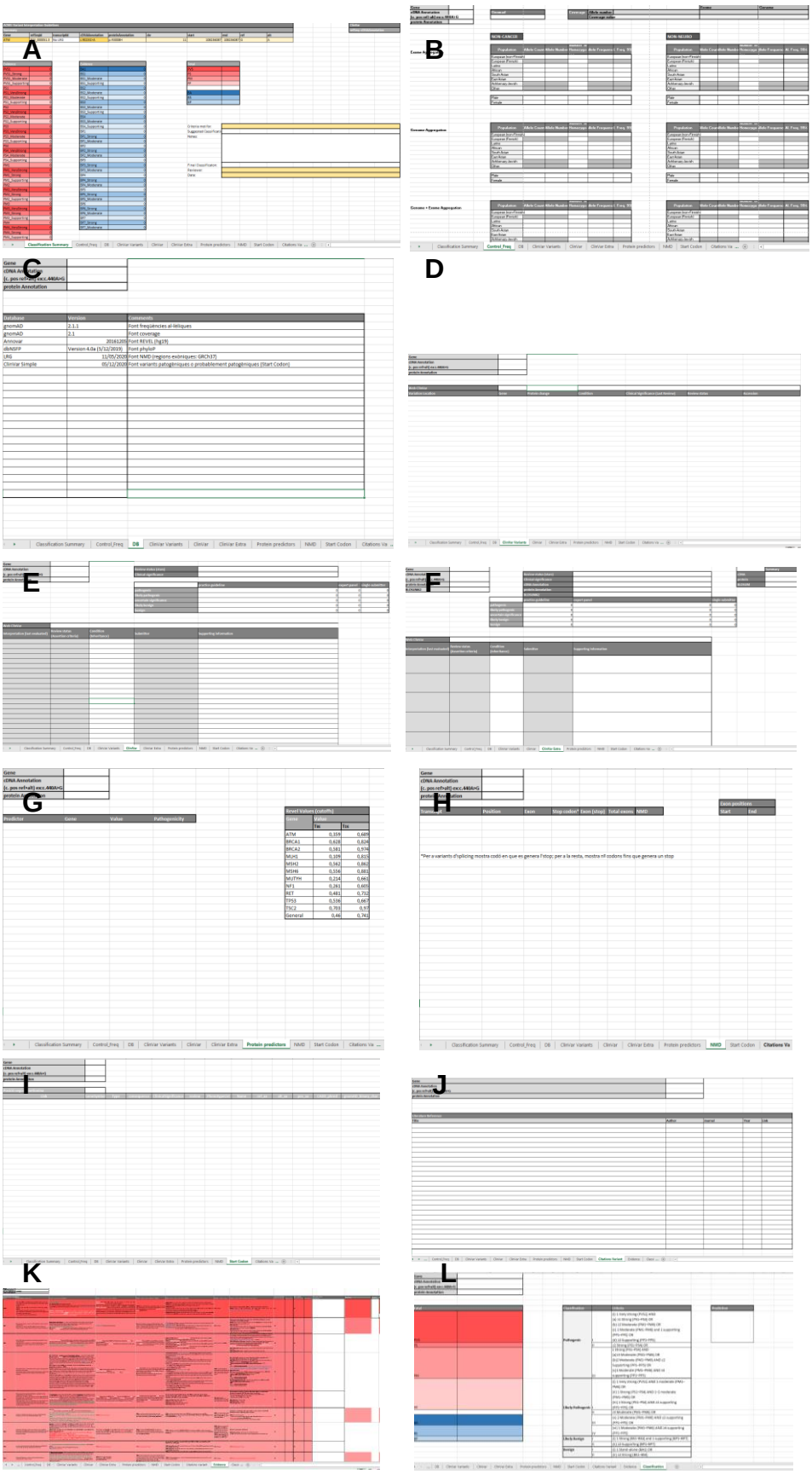



**Figura 14.** Visión general de las diferentes pestañas que forman el documento Plantilla. A) Visión general de la pestaña *Classification Summary*; B) Visión general de la pestaña *Control_Freq*; C) Visión general de la pestaña *DB*; D) Visión general de la pestaña *ClinVar Variants*; E) Visión general de la pestaña *ClinVar*; F) Visión general de la pestaña *ClinVar Extra*; G) Visión general de la pestaña *Protein predictors*; H) Visión general de la pestaña *NMD*; I) Visión general de la pestaña *Start Codon*; J) Visión general de la pestaña *Citations Variant*; K) Visión general de la pestaña *Evidence*; L) Visión general de la pestaña *Classification.*

Esta pestaña se encontraba en la Plantilla inicial. A continuación, se procederá a explicar las mejoras o novedades que se han introducido:

- Actualmente los colores utilizados para distinguir evidencias patogénicas y benignas son rojo y azul, respectivamente. Se han incluido en diferentes tonalidades según la fuerza de la evidencia *(Figura 15)*. Los colores actuales pueden ser distinguibles por usuarios que padezcan daltonismo.

- Se ha creado una casilla para cada una de las fuerzas posibles (*very strong, strong, moderate, supporting*) para todas las evidencias *(Figura 15).*

| Evidence | |
|---|---|
| PVS1 | 0 |
| PVS1_Strong | 0 |
| PVS1_Moderate | 0 |
| PVS1_Supporting | 0 |
| PS1 | 0 |
| PS1_VeryStrong | 0 |
| PS1_Moderate | 0 |
| PS1_Supporting | 0 |
| PS2 | 0 |
| PS2_VeryStrong | 0 |
| PS2_Moderate | 0 |
| PS2_Supporting | 0 |
| PS3 | 0 |
| PS3_VeryStrong | 0 |
| PS3_Moderate | 0 |
| PS3_Supporting | 0 |
| PS4 | 0 |
| PS4_VeryStrong | 0 |
| PS4_Moderate | 0 |
| PS4_Supporting | 0 |
| PM1 | 0 |
| PM1_VeryStrong | 0 |
| PM1_Strong | 0 |

| Evidence | |
|---|---|
| BA1 | 1 |
| BS1 | 0 |
| BS1_Moderate | 0 |
| BS1_Supporting | 0 |
| BS2 | 0 |
| BS2_Moderate | 0 |
| BS2_Supporting | 0 |
| BS3 | 0 |
| BS3_Moderate | 0 |
| BS3_Supporting | 0 |
| BS4 | 0 |
| BS3_Moderate | 0 |
| BS4_Supporting | 0 |
| BP1 | 0 |
| BP1_Strong | 0 |
| BP1_Moderate | 0 |
| BP2 | 0 |
| BP2_Strong | 0 |
| BP2_Moderate | 0 |
| BP3 | 0 |
| BP3_Strong | 0 |
| BP3_Moderate | 0 |
| BP4 | 1 |

**Figura 15.** Fragmento de las tablas de la pestaña *"Classification Summary"* en la que se muestra un resumen de las evidencias que se han cumplido. Con 1 se indican las evidencias que se cumplen y con 0 las que no.

- Pestaña 2 (*Control_Freq*): muestra los datos que se obtienen de gnomAD. En caso de no encontrarse la variante en alguna de las bases de datos (genomas y/o exomas), mostrará el promedio de alelos analizados de las variantes que se encuentren justo antes y justo después, y a la distancia en pb (pares de base) a la que se encuentra de ellas.

- Pestaña 3 (*DB*): en ella se recoge las versiones de las diferentes bases de datos de los archivos que se utilizan.

- Pestaña 4 (*ClinVar Variants*): proporciona una tabla con las diferentes variantes que se encuentran en ClinVar que se dan en el mismo codón que la variante. Incluye la URL de la página de ClinVar, la localización de la

variante, el símbolo del gen, el cambio proteico, las condiciones (enfermedades) en la que se han detectado, el significado clínico, el *review status*, y el número de acceso a ClinVar de cada una de las variantes.

- Pestaña 5 (*ClinVar*): recoge la información de la variante que proporciona ClinVar. Esta fue desarrollada durante las prácticas. Contiene la URL de la web de ClinVar, la clasificación de la variante, el *review status*, las condiciones (enfermedades) en la que se han detectado, el remitente, información extra, el recuento según el remitente y la clasificación de la variante, y el número de estrellas.

- Pestaña 6 (*ClinVar Extra*): sirve de molde para rellenar la información de ClinVar del resto de variantes que se dan en el mismo codón. Se creará una pestaña por cada variante que exista. También fue desarrollada durante las prácticas. Contiene la misma información que la pestaña ClinVar referente a las variantes que se den en el mismo codón. Además, incluye las puntuaciones de sustitución de BLOSUM62 de la variante y de la variante que se está clasificando, y un resumen de la comparativa de ambas variantes (del cDNA, de la proteína y del valor de BLOSUM62).

- Pestaña 7 (*Protein predictors*): se recogen los valores de los diferentes predictores de proteína y algoritmos de conservación de nucleótidos utilizados (REVEL, y phyloP, phastCons y/o GERP++).

- Pestaña 8 (*NMD*): se muestra la información que se utiliza en el cálculo de PVS1 para determinar si se da o no NMD. Se muestra la nomenclatura del transcrito canónico de LRG del gen, la posición de la variante, el exón en el que se encuentra, el número de codones desde la variante hasta que se genera el codón *stop*, el exón en el que se genera el codón *stop*, el número total de exones del transcrito, si se forma o no NMD, y las coordenadas de inicio y final de cada uno de los exones del transcrito.

- Pestaña 9 (*Start Codon*): se recoge el listado de variantes patogénicas o probablemente patogénicas que se encuentran en Simple ClinVar con 2 o más estrellas y que se dan antes de la segunda metionina. Entre la información que se proporciona de estas variantes está la URL a la web de ClinVar, el símbolo del gen, el tipo de variante (deleción, inserción, sustitución...), la consecuencia que tiene (codón *stop*, *missense*, *frameshift*…), el significado clínico, el *review status*, el tipo de cáncer en el que se encuentra, la nomenclatura del transcrito, el aminoácido de referencia y el alternativo, la posición del aminoácido (número de codón), el valor del CADD_phred (predictor) y el `gnomAD_binary_char` (indica si la variante se encuentra en la base de datos de gnomAD o no). También muestra la posición que ocupa la segunda metionina.

- Pestaña 10 (*Citations Variant*): se muestra el listado de citas bibliográficas que proporciona ClinVar sobre la variante. En esta se recoge el título del artículo, el autor/es, la revista, el año y el *link* al PMID.

- Pestaña 11 (*Evidence*): se proporcionan las evidencias que el programa habrá determinado que se cumplen, la fuerza que les da y un breve comentario sobre el por qué le da ese veredicto. En esta pestaña el usuario también podrá determinar el cumplimiento o no y el peso del resto de evidencias que el programa no habrá podido calcular, o si no está de acuerdo con alguna de las calculadas modificarle su cumplimiento o peso.

Además, el programa podrá proporcionar un conjunto de cuestiones que el usuario deberá responder para que calcule o recalcule alguna evidencia.

La siguiente pestaña también estaba presente en el documento inicial y ha sufrido cambios. Algunos de ellos son:

- Se utiliza el mismo código de colores que en la pestaña "*Classification Summary*".

- Se han ampliado las columnas que presenta la tabla. Conserva la columna con los códigos de evidencia y la columna con el criterio original de la evidencia según ACMG. Se ha añadido una columna con el criterio general (para la mayoría de genes) actualizado de la evidencia. También se ha añadido otra columna para cada uno de los genes que tienen criterios o umbrales específicos (*CDH1, PTEN, ATM, CHEK2* y *TP53*) con el detalle del criterio. Se han añadido dos columnas más que serán rellenadas por el programa *(Figura 16)*. En la primera, pondrá *met* o *unmet* en función de si se cumple o no la evidencia, NC si la evidencia no se ha calculado o NA si no aplica a ese gen. En la segunda se especificará la fuerza con la que se aplicará la evidencia en caso de ser *met* (*very strong, strong, moderate* o *supporting*). Cuenta con dos columnas más que podrán ser rellenadas por el usuario si no está de acuerdo con el resultado de algunas de las evidencias o si añade alguna que no se haya calculado. Como las que rellena el programa, la primera de estas columnas pondrá *met, unmet,* NC o NA, y la segunda columna dará la fuerza de la evidencia. Incorpora otra columna de comentarios en la que el programa resumirá, si se cumple una evidencia, la razón por la se cumple. Finalmente, cuenta con un conjunto de columnas al final en las que el programa, a veces, mostrará una serie de preguntas que el usuario deberá responder antes de ejecutar el segundo programa.

| Program analisys | | User analysis | | Program analisys |
|---|---|---|---|---|
| Analyzed (met/unmet) | Strenght evidence | Analyzed (met/unmet) | Strenght evidence | Comments |
| met | very strong | | | Es genera un codó stop a 3 codons des del codó on es dóna la variant, per tant, a l'exó 62. Es dóna NMD. |

**Figura 16.** Fragmento de la tabla de la pestaña *"Evidence"* en el que se observan las columnas añadidas que rellenará el programa y el usuario.

- Pestaña 12 (*Classification*): recoge el recuento de las diferentes evidencias en función de su fuerza y el veredicto de clasificación.

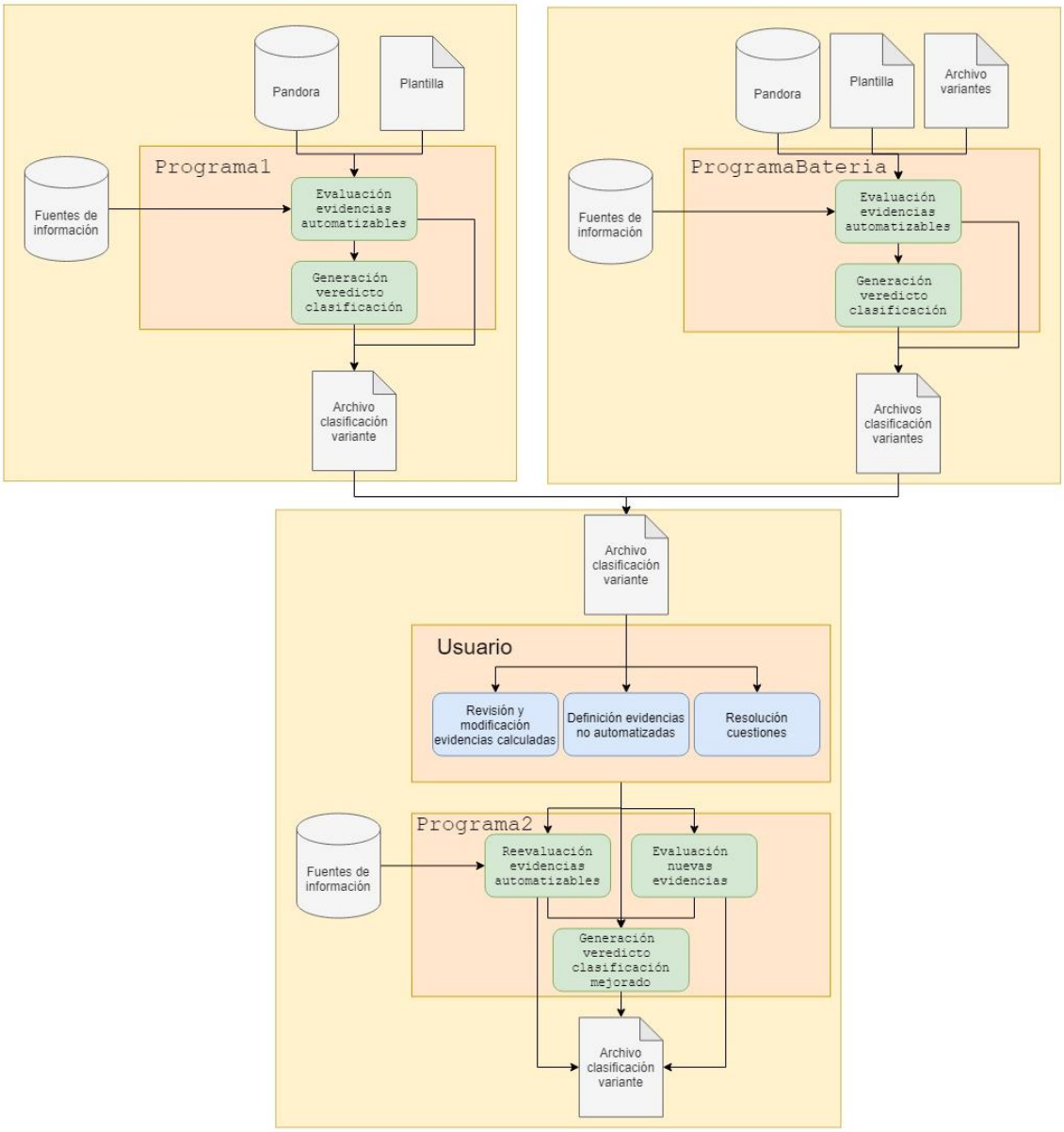



Figura 17. Organigrama general del funcionamiento de los tres programas: Programa1, ProgramaBateria y Programa2. El documento generado con la clasificación de la variante del Programa1 o de las variantes del ProgramaBateria es utilizado como *input* en el Programa2.

## 4.3 Funcionamiento del Programa1

El Programa1 lleva a cabo un primer cálculo de todas aquellas evidencias que son automatizables y proporciona un primer veredicto de clasificación *(Figura 17)*. En algunos casos, según las evidencias que haya podido calcular, el veredicto ya será Patogénica o Benigna, por lo cual habrá sido suficiente con ejecutar este programa y no hará falta utilizar el Programa2, a menos que la variante afecte el *splicing*. Se trata de un documento en formato RMarkdown y se considera como el programa principal. Todo el código utilizado se puede consultar en el documento Programa1.Rmd que se aporta como material suplementario. Además, los documentos de todas las variantes clasificadas que se utilizan como modelo para mostrar los resultados también se proporcionan como material suplementario.

Antes de realizar cualquier volcado de información, el programa genera el documento de clasificación como copia del documento Plantilla.xlsx y lo renombra con el símbolo del gen, la anotación de la variante, y la fecha en la que se genera (<SYMBOL_variant_yyyy-mm-dd.xlsx>). Será en este documento de clasificación en el que realmente realiza el registro de la información.

En los siguientes subapartados se presenta su desarrollo y funcionamiento en función de las evidencias que se vayan a calcular.

**Obtención de datos**

Inicialmente, el programa lee la primera pestaña del documento Plantilla.xlsx. De esta pestaña (*Classification Summary*), captura la información referente a la variante de estudio que ha indicado previamente el usuario. El código utilizado para realizar esto se encuentra en el apartado *Obtención de datos* del documento Programa1.Rmd.

En esta pestaña se encuentra la tabla que se observa en la figura 18. Como mínimo el usuario debe proporcionar la información de la columna *Gene* y *cDNAAnnotation* (resaltadas en color amarillo oscuro). Opcionalmente el usuario puede completar las columnas restantes (resaltadas en color amarillo claro). La columna transcriptId (en color blanco) es completada por el propio Programa 1 cuando se analicen variantes en las que se deba calcular PVS1 y proporcionará el nombre del transcrito en LRG.

| ACMG | Variant | Interpretation | Guidelines | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Summary | | | | | | | | | |
| **Gene** | **refSeqId** | **transcriptId** | **cDNAAnnotation** | **proteinAnnotation** | **chr** | **start** | **end** | **ref** | **alt** |
| ATM | NM_000051.3 | | c.162T>C | p.Y54= | 11 | 108098592 | 108098592 | T | C |

**Figura 18.** Captura de pantalla de la tabla mostrada en la pestaña *"Classification Summary" con la información de la variante que se va a analizar.*

El programa intenta leer todas las columnas. Si solamente se le proporciona la información de las columnas *Gene* o *cDNAAnnotation*, este procede a obtener el resto de la información de la variante conectándose a la base de datos del departamento de Pandora. En caso de que la variante se encuentre en Pandora, obtiene el resto de la información y prosigue con el programa. Si no encuentra la información, emite un mensaje de error, indicando que se requiere el resto de la información, para ejecutar el programa.

**Evidencias PM2, BA1, BS1 (dependen de la frecuencia alélica)**

Las evidencias PM2, BA1 y BS1 son calculadas para todos los tipos de variantes que se han programado (*silent, missense, nonsense, frameshift,* afectación del codón de inicio). Para su cálculo, el programa obtiene información de la frecuencia de la variante en la población general a partir de los archivos genoma.xlsx y exoma.xlsx *(Figura 19)*. El código utilizado para realizar esto se encuentra en los apartados *Obtención de datos, Exoma+genoma, Overall frequency, Cálculo frecuencia alélica CI 99%,* y *Evidence (subapartados PM2, BA1, BS1)* del documento Programa1.Rmd.

A continuación, el programa procede a buscar en ambos archivos si la variante que se está analizando se encuentra o no presente. Para ello busca la fila de

los archivos en la que coincide el cromosoma, la posición de inicio -1, el nucleótido de referencia y el alternativo. Si se encuentra en ambos o en alguno de los dos significa que la variante se encuentra en la base de datos de gnomAD, sino significará que no está presente.

En caso de no encontrarse en gnomAD, se comprueba que la región en gnomAD esté bien cubierta (suficientes individuos correctamente secuenciados). Para ello busca las variantes justamente anterior y posterior que haya respecto a la posición de la variante de estudio. De estas dos posiciones, interesa el número de alelos totales (AN) que se analizan y la distancia en el genoma respecto a la variante. Esto se realiza cuando la variante no se encuentre en ninguna de las dos bases de datos, o cuando falte en una de ellas.

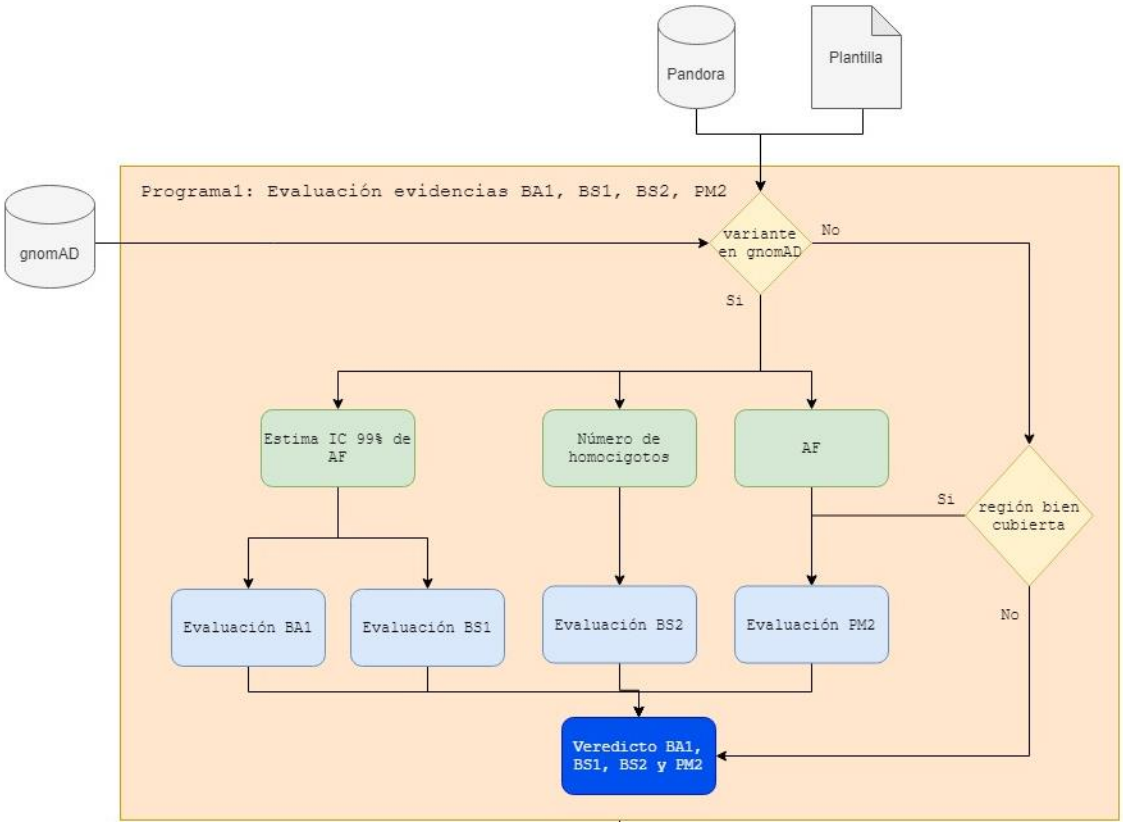



**Figura 19.** Ejemplo de organigrama que muestra la lógica detrás del Programa1 para el cálculo de las evidencias BA1, BS1, PM2 y BS2, las cuales dependen de la información de gnomAD.

En caso de encontrarse en gnomAD, procede a obtener la información de las diferentes columnas. De este archivo coge el recuento de alelos (AC), los alelos totales analizados (AN) y el número de individuos homocigotos para la variante (nhomalt) y la frecuencia alélica (AF) de las diferentes poblaciones (*European (non-Finnish), European (Finnish), Latino, African, South Asian* (solo en exomas), *East Asian, Ashkenazi Jewish, Other*, y los datos *male* y *female*), tanto del subconjunto de individuos *non-cancer* como *non-neuro*.

Si esta información la obtiene de genomas y exomas, unifica los datos de ambas en una única tabla. Si solamente se encuentra en una de ellas, utiliza el AN calculado de la otra base de datos para calcular la AF. También unifica las diferentes poblaciones, calculando los AC, AN, nhomalt y AF totales.

Finalmente, calcula las AF al 99% de las diferentes poblaciones y total (excepto en los genes *TP53* y *CDH1* que calcula al 99,99%). Para su cálculo, estima el intervalo de confianza, basándose en la distribución de Poisson. Por ello, utiliza la función `PoissonCI` del paquete `DescTools` [37]. El programa toma el valor del límite inferior del intervalo para ser lo más conservador posible.

Toda la información que obtiene de gnomAD la recoge en la pestaña "*Control_Freq*" de la Plantilla. En la figura 20 se puede observar cómo quedaría rellenada una de las tablas que se encuentra en esta pestaña.

Después de obtener toda la información, procede a calcular las diferentes evidencias. Para BA1, utiliza la AF al 99% de algunas de las poblaciones (*European (non-Finnish), Latino, African, South Asian* y *East Asian*). Para el gen *CDH1*, se requiere que se hayan detectado 5 o más alelos con la variante y que la frecuencia sea superior a 0.002 en alguna de las subpoblaciones. Para *PTEN* también se necesitan los 5 alelos o más, pero con una frecuencia superior a 0.01. Para *ATM* solamente se exige que la frecuencia sea superior a 0.005. Para *TP53*, se vuelve a exigir como mínimo los 5 alelos, pero con una frecuencia igual o superior a 0.001. Finalmente, para el resto de genes, se exige una frecuencia superior a 0.01.

| Population | Allele Count | Allele Number | Number of Homozygotes | Allele Frequency | Al. Freq. 99% |
|---|---|---|---|---|---|
| European (non-Finnish) | 167 | 102224 | 0 | 0,001633667 | 0,001354033 |
| European (Finnish) | 4 | 21592 | 0 | 0,000185254 | 3,81275E-05 |
| Latino | 133 | 34076 | 1 | 0,00390304 | 0,003159106 |
| African | 95 | 14850 | 0 | 0,006397306 | 0,004970048 |
| South Asian | 15 | 30352 | 1 | 0,000494201 | 0,000246334 |
| East Asian | 0 | 17676 | 0 | 0 | 0 |
| Ashkenazy Jewish | 301 | 9544 | 8 | 0,031538139 | 0,027463899 |
| Other | 42 | 5582 | 3 | 0,007524185 | 0,005088945 |
| Male | 397 | 128712 | 9 | 0,003084405 | 0,002735746 |
| Female | 360 | 107184 | 4 | 0,00335871 | 0,002960669 |

**Figura 20.** Captura de pantalla de la pestaña "*Evidence*". Muestra la tabla obtenida de la población *non-cancer* de la base de datos de exomas del documento de la variante c.1810C>T de *ATM*.

Si no se cumple BA1, comprueba si se cumple BS1. Para su cálculo también utiliza la AF al 99% de algunas de las mismas poblaciones. Para el gen *CDH1* y *PTEN*, se le solicita un mínimo de 5 alelos con la variante y una frecuencia superior a 0.001 para alguna de las subpoblaciones. Además, para *PTEN*, en caso de que la frecuencia alélica sea superior a 0.000043, se le da un peso de "*supporting*" en vez de "*strong*". Para *ATM*, se requiere solamente una frecuencia superior a 0.0005. Para *TP53*, de nuevo se le exige el mínimo de 5 alelos, pero una frecuencia superior a 0.0003. Finalmente, para el resto de genes, solamente se necesita una frecuencia superior a 0.005.

Para el cálculo de PM2, a diferencia de las otras dos, utiliza directamente las frecuencias alélicas generales de la población y de algunas subpoblaciones (*European (non-Finnish), Latino, African, South Asian* y *East Asian*). Para *TP53*, se requiere que las frecuencias de todas las poblaciones sean 0. Para el resto de los genes, se exige que la frecuencia sea inferior a 0.00001 en la población general e inferior a 0.00002 en alguna de las subpoblaciones. Además, para *CHEK2* en caso de que la frecuencia de la población general sea inferior o igual a 0.00005, se le da un peso de "*supporting*" en vez de

"*moderate*", y para el resto de genes sucede lo mismo, pero cuando esta frecuencia sea inferior o igual a 0.00002. Además, en los casos en los que no se haya encontrado la variante en ninguna de las dos bases de datos, si la zona se encuentra bien cubierta se cumple PM2.

| BENIGN EVIDENCE | Program analysis | | User analysis | | Program analysis |
|---|---|---|---|---|---|
| Code | Analyzed (met/unmet) | Strength evidence | Analyzed (met/unmet) | Strength evidence | Comments |
| **BA1** | met | stand alone | NC | | Les frequencies al 99% de les subpoblacions son: nfe - 0.00132433925878548; amr - 0.00318601176476377; afr - 0.00533633334419203; sas - 0.00024633395704493; i eas - 0. Com a mínim en una d'elles es dóna una freqüència superior de 0.005. |
| **BS1** | NC | | NC | | |
| **BS2** | met | supporting | NC | | El nombre d'homozigots son 13. |

**Figura 21.** Fragmento de la tabla de la pestaña "*Evidence*", en el que se muestra el veredicto de las evidencias BA1, BS1 y BS2 de la variante c.1810C>T de *ATM.*

El veredicto de las diferentes evidencias se muestra en la pestaña "*Evidence*". Tomando como ejemplo la variante de la figura 20, se puede observar cómo cumple BA1 (*Figura 21*).

**BS2**

BS2 también es calculada para todos los tipos de variantes *(Figura 19)*. El código utilizado para realizar esto se encuentra en los apartados *Obtención de datos, Exoma+genoma,* y *Evidence (subapartado BS2)* del documento Programa1.Rmd. Para su cálculo, se utiliza el número de individuos homocigotos para la variante (nhomalt) en la población general. Este valor lo obtiene directamente de la información que le proporciona gnomAD. Para la mayoría de genes, utiliza el nhomalt del subconjunto de individuos *non-cancer*. Solamente para el gen *ATM* se utiliza el *non-neuro*. Para *ATM,* se le exige un mínimo de 2 individuos para que se cumpla la evidencia. Para *CDH1*, requiere que como mínimo haya 3 homocigotos en *non-cancer*. Para *CHEK2* y *TP53* no se calculará. Para el resto de genes, se necesitan como mínimo 2 homocigotos. En todos los casos el peso que se le da es de "*supporting*".

En la figura 21, se puede observar cómo determina el veredicto para esta evidencia.

**PS1, PM5 (ClinVar)**

Las evidencias PS1 y PM5 son evidencias que son calculadas para variantes *missense*. La información que necesita para calcularlas la obtiene de ClinVar *(Figura 22)*. El código utilizado para realizar esto se encuentra en los apartados *ClinVar Variants, ClinVar, ClinVar Extra, ClinVar Variants (II), ClinVar (II),* y *ClinVar Extra (II)* del documento Programa1.Rmd.

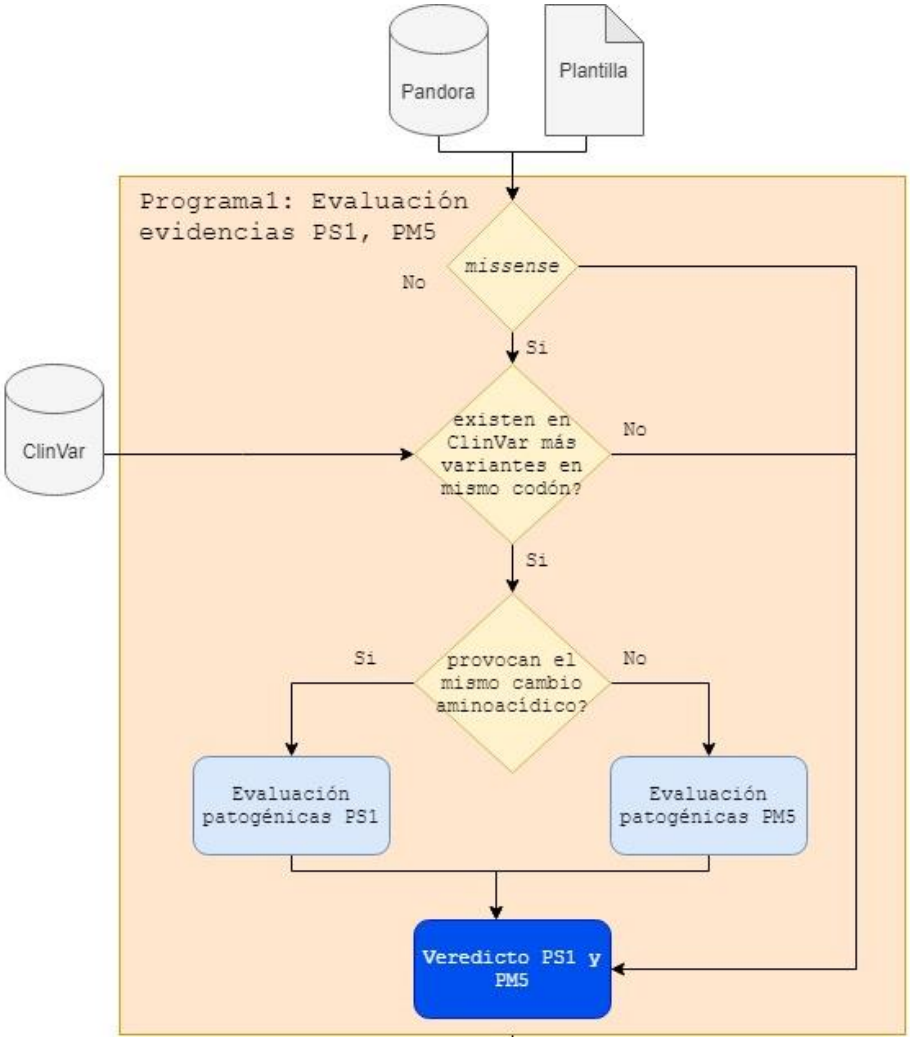



**Figura 22.** Ejemplo de organigrama que muestra la lógica detrás del Programa1 para el cálculo de las evidencias PS1 y PM5.

Primero, obtiene una tabla con todas las variantes que se producen en el mismo codón que la variante y que se encuentran clasificadas en ClinVar. Contiene la información de la localización en el cromosoma, el simbolo del gen, el cambio producido en la proteína, las condiciones en las que se ha encontrado, el significado clínico, el *review status,* y el *Accession number (Figura 23).* Para ello, con la información que se dispone de la variante, genera la URL que da acceso a la página web de ClinVar,

`https://www.ncbi.nlm.nih.gov/clinvar/?term=`**`prot`**`+%5Bvariant+name%5D+and+`**`gen`**

donde *gen* es sustituido por el símbolo del gen proporcionado por el usuario en la Plantilla, y *prot* es la anotación de la proteína indicada con el aminoácido de referencia y la posición del codón. Para obtener esta, se utiliza la anotación de la proteína que o bien ha proporcionado el usuario o ha obtenido de Pandora. En caso de que el aminoácido se dé con el código de una sola letra, el programa lo convierte al código de tres letras.

Segundo, el programa busca la información referente a la variante de estudio en ClinVar. Con ese fin, rastrea si la variante se encuentra en la tabla obtenida del apartado anterior. Si no la encuentra, indica que la variante no está

registrada en ClinVar. Si la encuentra, busca la información del *Accession Number* de la tabla, y genera una nueva URL.

`https://www.ncbi.nlm.nih.gov/clinvar/variation/`**`accession`**`/`

| Web ClinVar | https://www.ncbi.nlm.nih.gov/clinvar/?term=Lys276+%5Bvariant+name%5D+and+ATM | | | | | |
|---|---|---|---|---|---|---|
| **Variation Location** | **Gene** | **Protein change** | **Condition** | **Clinical Significance (Last Review)** | **Review status** | **Accession** |
| NM_000051.3(ATM):c.826A>C (p.Lys276Gln)GRCh37: Chr11:108115678GRCh38: Chr11:108244951 | ATM | K276Q | not provided, Ataxia-telangiectasia syndrome Hereditary cancer-predisposing syndrome, | Uncertain significance (Jun 13, 2019) | criteria provided, multiple submitters, no conflicts | VCV000419650 |
| NM_000051.3(ATM):c.826A>G (p.Lys276Glu)GRCh37: Chr11:108115678GRCh38: Chr11:108244951 | ATM | K276E | Ataxia-telangiectasia syndrome, not provided | Uncertain significance (Jun 13, 2018) | criteria provided, multiple submitters, no conflicts | VCV000143025 |

**Figura 23.** Captura de pantalla de la tabla de las diferentes variantes producidas en el codón 276 de *ATM* registradas en ClinVar. Resultado obtenido de la pestaña "*ClinVar*" de la variante c.826A>G de *ATM.*

Esta URL es utilizada para obtener la primera, la cuarta y la sexta tabla de la web. La primera contiene las diferentes nomenclaturas que recibe la variante. Esta información la muestra en la pestaña "*Classification summary*" a modo informativo. La cuarta corresponde a todos los registros que se han producido sobre la variante. Se especifica la clasificación, el *review status*, las enfermedades en las que se ha encontrado, el remitente e información extra. A partir de esta información, realiza el recuento de las diferentes combinaciones posibles entre los diferentes niveles de significado clínico de la variante y los diferentes *review status* (ver apartado 3.1). Con los valores que obtiene, calcula el número de estrellas y el significado clínico de la variante. Toda esta información la muestra en la pestaña "*ClinVar*" de la Plantilla *(Figura 24)*. Finalmente, la sexta tabla proporciona un listado de citas bibliográficas en las que aparece la variante. Contiene el nombre del artículo, el nombre del autor, la revista, el año de publicación y el *link* a PMID (identificador de PubMed) *(Figura 25)*. Esta información la muestra en la pestaña "*Citations Variant*".

Tercero, el programa obtiene la información del resto de variantes que se dan en el mismo codón registradas en ClinVar. Del mismo modo que para la variante de estudio, genera la URL para cada una de ellas utilizando el *Accession Number*. Sin embargo, de estas solamente captura la cuarta tabla de la web, proporcionando la misma información que se obtiene de la variante que se clasifica. También procede a calcular el recuento de las diferentes combinaciones posibles entre los diferentes niveles de significado clínico de la variante y los diferentes *review status*; así como el número de estrellas y la clasificación de la variante. Como información extra, a partir de la matriz de conservación BLOSUM62, calcula las puntuaciones de sustitución del cambio producido en la proteína de la variante y del cambio producido del resto de variantes que se dan en el mismo codón. Finalmente, muestra la comparación entre la variante de estudio con cada una de las variantes producidas en el mismo codón. Realiza la comparación del cDNA, de la proteína y de la puntuación de sustitución de BLOSUM62 *(Figura 26)*. Para recoger toda la

información, el programa utiliza como referencia la pestaña "*ClinVar Extra*". Genera una pestaña para cada variante que se da en el mismo codón que la variante de estudio y recoge su información. Además, renombra la pestaña con el nombre de la variante y el nombre de la proteína.

| | |
|---|---|
| **Review status (stars)** | 2 |
| **Clinical significance** | uncertain significance |

| | practice guideline | expert panel | single submitter |
|---|---|---|---|
| **pathogenic** | 0 | 0 | 0 |
| **likely pathogenic** | 0 | 0 | 0 |
| **uncertain significance** | 0 | 0 | 3 |
| **likely benign** | 0 | 0 | 0 |
| **benign** | 0 | 0 | 0 |

| **Web ClinVar** | https://www.ncbi.nlm.nih.gov/clinvar/variation/143025/ | | | |
|---|---|---|---|---|
| **Interpretation (last evaluated)** | **Review status (Assertion criteria)** | **Condition (Inheritance)** | **Submitter** | **Supporting information** |
| Uncertain significance (Jun 13, 2018) | criteria provided, single submitter (Invitae Variant Classification Sherloc (09022015)) Method: clinical testing | Ataxia-telangiectasia syndrome Allele origin: germline | Invitae Accession: SCV000828249.1 Submitted: (Aug 29, 2018) | Evidence detailsPublicationsPubMed (1) Comment: This sequence change replaces lysine with glutamic acid at codon 276 of the ATM protein (p.Lys276Glu). The lysine residue is highly conserved and there is â€¦ (more)This sequence change replaces lysine with glutamic acid at codon 276 of the ATM protein (p.Lys276Glu). The lysine residue is highly conserved and there is a small physicochemical difference between lysine and glutamic acid. This variant is not present in population databases (ExAC no frequency). This variant has not been reported in the literature in individuals with ATM-related disease. ClinVar contains an entry for this variant (Variation ID: 143025). Algorithms developed to predict the effect of missense changes on protein structure and function are either unavailable or do not agree on the potential impact of this missense change (SIFT: "Deleterious"; PolyPhen-2: "Probably Damaging"; Align-GVGD: "Class C15"). Algorithms developed to predict the effect of sequence changes on RNA splicing suggest that this variant may create or strengthen a splice site, but this prediction has not been confirmed by published transcriptional studies. In summary, the available evidence is currently insufficient to determine the role of this variant in disease. Therefore, it has been classified as a Variant of Uncertain Significance. (less) |
| Uncertain significance (Feb 21, 2017) | criteria provided, single submitter (GeneDx Variant Classification (06012015)) Method: clinical testing | Not Provided Allele origin: germline | GeneDx Accession: SCV000618295.2 Submitted: (Jan 29, 2019) | Evidence details Comment: This variant is denoted ATM c.826A>G at the cDNA level, p.Lys276Glu (K276E) at the protein level, and results in the change of a Lysine to â€¦ (more)This variant is denoted ATM c.826A>G at the cDNA level, p.Lys276Glu (K276E) at the protein level, and results in the change of a Lysine to a Glutamic Acid (AAA>GAA). This variant has not, to our knowledge, been published in the literature as pathogenic or benign. ATM Lys276Glu was not observed in large population cohorts (Lek 2016, The 1000 Genomes Consortium 2015, NHLBI Exome Sequencing Project). Since Lysine and Glutamic Acid differ in polarity, charge, size or other properties, this is considered a non-conservative amino acid substitution. ATM Lys276Glu occurs at a position that is conserved across species and is not located in a known functional domain. Based on currently available evidence, it is unclear whether ATM Lys276Glu is pathogenic or benign. We consider it to be a variant of uncertain significance. (less) |

**Figura 24.** Fragmento de la pestaña "*ClinVar*" de la variante c.826A>G de *ATM*.

| **Literature Reference** | | | | |
|---|---|---|---|---|
| **Title** | **Author** | **Journal** | **Year** | **Link** |
| Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. | Nykamp K | Genetics in medicine : official journal of the American College of Medical Genetics | 2017 | PMID: 28492532 |

**Figura 25.** Captura de pantalla de la tabla con el listado de citas bibliográficas obtenido de ClinVar para la variante c. 826A>G de *ATM*. Se muestra en la pestaña "*Ctations Variant*".

| | Summary |
|---|---|
| cDNA | Diferent canvi nucleotídic |
| protein | Diferent canvi d'aminoàcid |
| BLOSUM | La variant problema implica un canvi menys conservatiu que el de la variant classificada (probabilitat igual o més petita) |

**Figura 26.** Captura de pantalla de la comparación entre las variantes c.826A>G y c.826A>C de *ATM*. Información contenida en la pestaña "*ClinVar c.826A>C (K276Q)*" de la variante c.826A>G de *ATM*.

Esta última información es la que el programa utiliza para calcular las evidencias PS1 y PM5. Para el cálculo de PS1, el programa comprueba que el cambio en la secuencia de la proteína de la variante de estudio sea el mismo que se produce en la variante de ClinVar. Si se da el mismo cambio, se comprueba que el número de estrellas de la variante de ClinVar sea igual o superior a 3, y que el significado clínico sea: para *ATM* patogénico o probablemente patogénico, para *TP53* no se calcula, y para el resto de genes si es patogénico le da una fuerza "*strong*" y si es probablemente patogénico le da una fuerza de "*moderate*".

Para el cálculo de PM5, primero el programa comprueba que el cambio en la proteína sea distinto entre las dos variantes. Segundo, utiliza la puntuación de sustitución de BLOSUM62 de cada una para determinar PM5. Cuando el valor de la variante es menor o igual al valor de la variante de ClinVar, se comprueban las estrellas y el significado clínico. Cuando el número de estrellas sea igual o superior a 3 y el significado clínico sea patogénico o probablemente patogénico: para *TP53* se le da una fuerza de "*supporting*"; y para el resto de genes se le da una fuerza de "*moderate*".

**PVS1**

En este programa esta evidencia es calculada para variantes *nonsense, frameshift* (producida por pequeñas INDELs), y con afectación en el codón de inicio. Sin embargo, la metodología de cálculo varía entre ellas, especialmente entre *nonsense/frameshift* y afectación del codón de inicio.

Para las variantes *nonsense* y *frameshift,* utiliza la información del archivo LRG_genes.txt. El código utilizado para realizar esto se encuentra en los apartados *Evidence (subapartado PVS1),* y *NMD* del documento Programa1.Rmd. Sin embargo, para las variantes *frameshift,* primero debe generar una URL que acceda a la herramienta Name Checker de Mutalyzer.

```
https://mutalyzer.nl/name-checker?description=NM%28gen%29%3Avar
```

Para generarla, necesita la nomenclatura del transcrito (*NM*) y el símbolo del gen (*gen*). El gen lo proporciona siempre el usuario, y el NM lo proporciona el usuario o se obtiene de Pandora. Además, se necesita la nomenclatura de la variante (*var*), que siempre introduce el usuario. Finalmente, para pequeñas deleciones e inserciones, comprueba que estas se encuentren *out of frame*.

Generada la URL, accede a la herramienta para obtener la información de la variante. Si encuentra en alguna línea "fs", significa que se genera *frameshift* y, por lo tanto, se guarda la línea en una variable. A partir de esta línea, obtiene el número de codones desde la posición de la variante hasta que se genera el *stop*. Si en cambio encuentra "*", indica que en el lugar de la inserción o la deleción se genera un *stop* y crea una variable que indica que el *stop* se genera en el mismo codón. Si no se encuentra ninguno de los dos patrones,

indica que la deleción o la inserción no provocan un codón *stop* prematuro. Si en cambio la variante es *nonsense* (e.g. p.T488*, donde la * indica que se genera un *stop* en esa posición), indica en una variable que el *stop* se genera en el mismo codón.

En caso de encontrar que se genera el codón *stop*, lee el archivo LRG_genes.txt. Si se está analizando una variante del gen *TP53,* guarda la información del nombre de los transcritos 1, 3 y 4 de LRG y las coordenadas de los exones de todos ellos. Para el resto de genes solamente se guarda del transcrito canónico que es el que hay en el archivo. Las coordenadas las procesa para obtener en cada fila dos columnas, una con la posición de inicio y otra con la final de cada uno de los exones del transcrito.

A continuación, calcula la longitud del transcrito y la posición nucleotídica en la que se genera el codón *stop*. Una vez calculados, calcula el exón en el que se ha generado el codón *stop*. En el caso de variantes *nonsense*, el exón en el que se genera el *stop* es el mismo que en el que se encuentra la variante. Finalmente, calcula la proporción de proteína que se ha eliminado respecto a la longitud total del transcrito.

Toda esta información es la que utiliza para determinar el cumplimiento o no de esta evidencia. Para determinarla, sigue el árbol de decisión de la figura 27. Para saber si se produce o no NMD, se basa en el exón en el que se produce el *stop*. Si este se localiza antes del penúltimo exón, se genera NMD. Si se localiza en el último exón, no se genera NMD. Y si se genera en el penúltimo exón se comprueba si este se ha producido antes o después de los últimos 50 nucleótidos. Si se produce después, no se produce NMD, pero si se produce antes, sí.

Cuando se da NMD, comprueba si se localiza o no en un exón biológicamente relevante. Si no lo es, no se cumple la evidencia. Para *BRCA1*, se considera que no son regiones relevantes cuando caiga en los exones 9 y 10. Para *TSC2* cuando se encuentre en los exones 25 y 31. Y para *TP53* cuando se localice en el exón extra (el número 3) que tienen los transcritos 3 y 4. En el resto de exones y genes que se dé NMD, se le da un peso de "*very strong*".

Cuando no se da NMD, se comprueba el porcentaje de proteína que se pierde. Si este es inferior al 10%, se le da un peso de "*moderate*". Si es superior, se le da un peso de "*strong*".

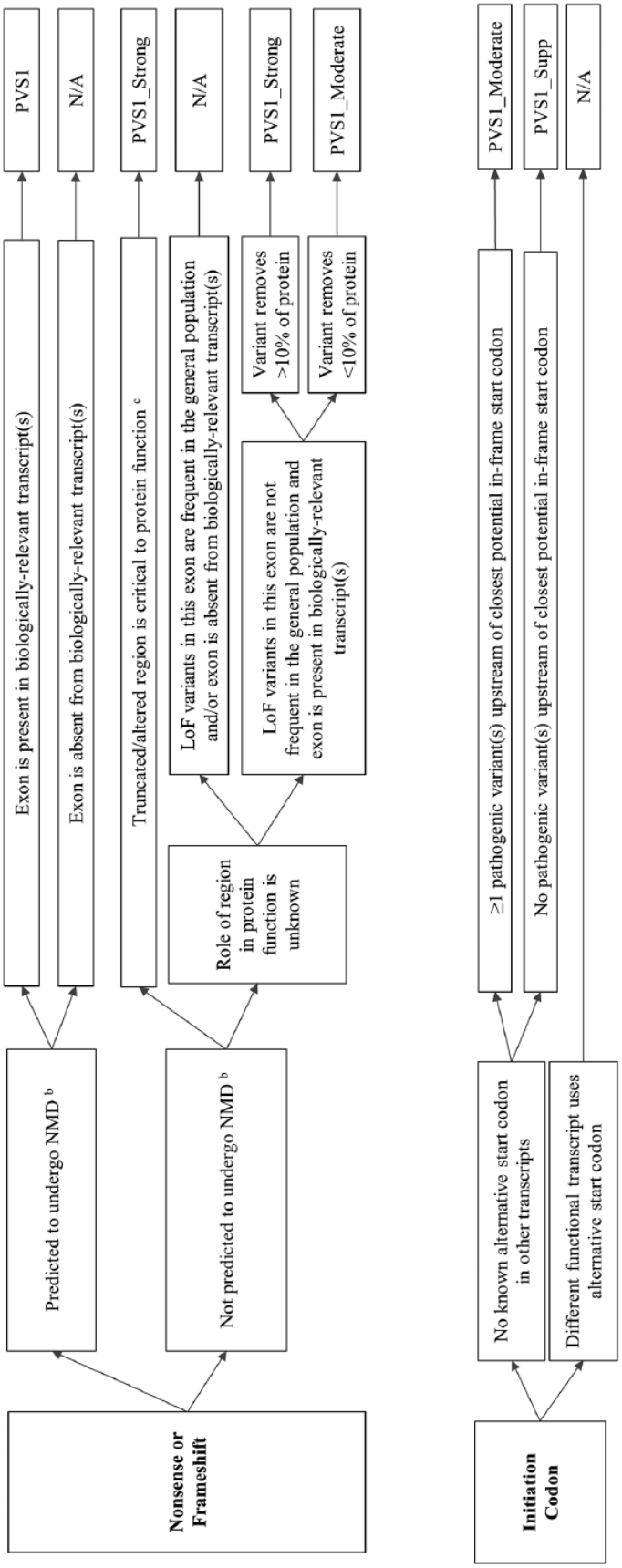



**Figura 27.** Fragmento del árbol de decisión de Tayoun, 2018 referente a las mutaciones *nonsense* y *frameshift*, y a las mutaciones con afectación del codón de inicio.

La información referente a la evidencia PVS1 para variantes *nonsense* y *frameshift* la recoge en la pestaña "*NMD*" de la Plantilla. Se proporciona el nombre del transcrito en LRG, la posición nucleotídica en el que se da la variante, el exón en el que se encuentra, el número de codones que hay hasta que se genera el *stop*, el exón en el que se produce el *stop*, el número de exones que tiene el transcrito, si se produce NMD o no, y las coordenadas de los exones del transcrito *(Figura 28)*. En cuanto al veredicto de la clasificación, se proporciona en la pestaña "*Evidence*".

| Transcript | Position | Exon | Stop codon* | Exon (stop) | Total exons | NMD |
|---|---|---|---|---|---|---|
| LRG_135t1 | 108121655 | 10 | 0 | 10 | 63 | SI |

| Exon positions | |
|---|---|
| Start | End |
| 108093559 | 108093913 |
| 108098322 | 108098423 |
| 108098503 | 108098615 |
| 108099905 | 108100050 |
| 108106397 | 108106561 |
| 108114680 | 108114845 |
| 108115515 | 108115753 |
| 108117691 | 108117854 |
| 108119660 | 108119829 |
| 108121428 | 108121799 |
| 108122564 | 108122758 |
| 108123544 | 108123639 |
| 108124541 | 108124766 |

**Figura 28.** Fragmento de la pestaña "*NMD*" obtenido de la variante *nonsense* c.1463G>A de *ATM.*

Para las variantes en la que se ve afectado el codón de inicio, utiliza un procedimiento distinto y se basa en la información que obtiene del archivo Start_Codon.xlsx *(Figura 27)*. El código utilizado para realizar esto se encuentra en los apartados *Evidence (subapartado PVS1),* y *Evidence (II) (subapartado Pathogenic)* del documento Programa1.Rmd. En este tipo de variantes, se ve afectado el primer, segundo o tercer nucleótido de la secuencia codificante.

Primero, el programa comprueba la posición en la que se encuentra la segunda metionina en pauta en la secuencia. Esta se podría usar como nuevo codón de inicio, constituyendo la posibilidad más conservadora de efecto en la proteína. Para calcularla, obtiene la secuencia aminoacídica del UCSC Genome Browser. Por ello, genera el siguiente URL, donde *ensembl* es el ID del transcrito canónico en Ensembl, que se obtiene del archivo dbNSFP:

```
http://genome.cse.ucsc.edu/cgi-
bin/hgGene?hgsid=955202475_hTXiT4KTxZ25hWn7UF2LSHCY9EFR&hgg_do_g
etProteinSeq=1&hgg_gene=ensembl
```

A partir de la secuencia aminoacídica, procede a encontrar la segunda metionina, guardando la posición en la que se encuentra.

Segundo, lee el archivo Codon_Start.xlsx, el cual contiene un listado con variantes patogénicas o probablemente patogénicas con 2 o más estrellas, obtenido de Simple ClinVar.

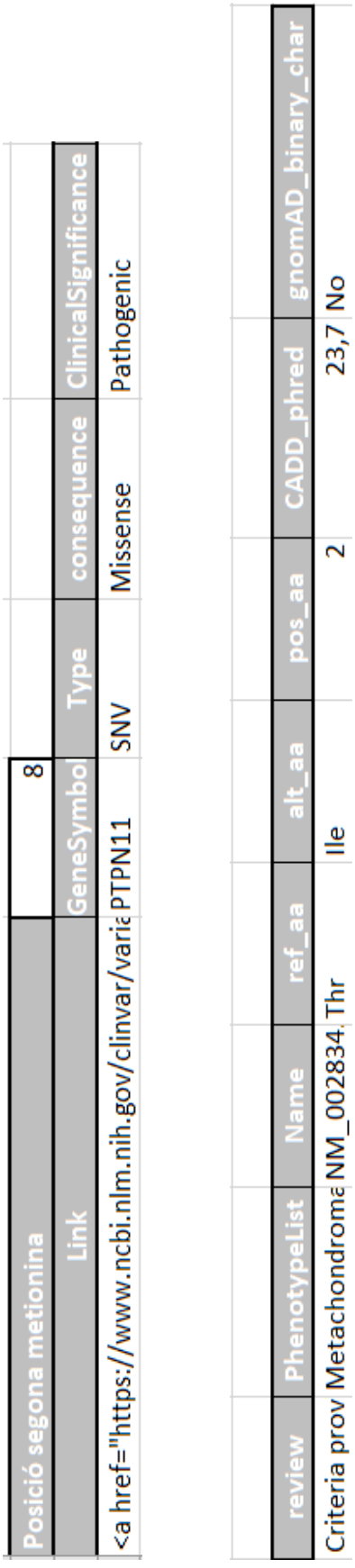
Finalmente, realiza la búsqueda para el gen que se está analizando y filtra el resultado, quedándose solo con aquellas variantes que se encuentren antes de la posición de la segunda metionina. De estas variantes proporciona el *link* a la web de ClinVar, el símbolo del gen, el tipo de variante, la consecuencia, el significado clínico, el *review status*, el listado fenotípico, el nombre de la variante, el aminoácido de referencia y el alternativo, la posición aminoacídica, y el valor del predictor CADD. Esta información la proporciona en la pestaña "*Start Codon*" *(Figura 29)*.

Para dar el veredicto de la evidencia, sigue el árbol de decisión que se encuentra en la figura 27 para este tipo de variante. En este tipo de variantes, la proteína putativa que se generaría tendría delecionada toda la región antes de la segunda metionina. Por ello, si se demuestra la existencia de variantes *missense* patogénicas o probablemente patogénicas en esta región, sería un indicador de la putativa patogenicidad de la pérdida del codón de inicio original. En caso de que se hayan encontrado variantes patogénicas o probablemente patogénicas antes del segundo codón de inicio indica que se cumple la evidencia, con un peso de "*moderate*". En caso de que no se hayan encontrado, se le asigna que cumple la evidencia, pero con un peso menor ("*supporting*").

**PM1**

La evidencia PM1 es calculada para todas las variantes, siempre y cuando no cumpla PVS1. Esta evidencia se aplica a aquellas variantes que afectan aminoácidos que caen en regiones críticas, *hot-spot* o con dominios funcionales críticos. El código utilizado para realizar esto se encuentra en el apartado *Evidence (II) (subapartado Pathogenic)* del documento Programa1.Rmd. Primero, se encarga de guardar en una variable el codón en que se produce la variante.

Segundo, en función del gen, comprueba si esta se encuentra dentro de unas regiones en concreto que están definidas como regiones importantes de la proteína. Para *CDH1* esta evidencia no es calculada. Para *PTEN*, se consideran regiones críticas los codones entre el 90 y el 94, y entre el 123 y el 130, ya que corresponden a residuos de motivos catalíticos de la proteína. Por lo que, si la posición en que se produce la variante cae en alguna de estas regiones, se cumple la evidencia PM1. Para *ATM*, cuando el residuo afectado sea el 3008 o el 1981, se cumple PM1, con un peso "*moderate*". Pero cuando se vea afectado algún residuo entre el 2712 y el 2962, o entre el 3024 y el 3056, los cuales pertenecen a dominios funcionales críticos (quinasa y FATC, respectivamente), se cumple la evidencia, pero con un peso de "*supporting*". Para *CHEK2*, las regiones críticas serían entre los codones 115 y 175, y los codones 226 y 486, los cuales se tratan de dominios altamente conservados (FHA y quinasa, respectivamente). De modo que, de verse afectado algún codón de estas regiones, hace que se cumpla PM1. Finalmente, para *TP53*, las posiciones críticas son codones en concreto (*hot-spots* mutacionales): 175, 248, 273, 245, 282 y 249. De verse afectada alguna de estas posiciones, se cumplirá PM1. Para el resto de genes, al no estar definidas estas regiones, quedarán a valoración del usuario.



| Posició segona metionina | | 8 | | | | |
|---|---|---|---|---|---|---|
| Link | GeneSymbol | Type | consequence | ClinicalSignificance | | |
| <a href="https://www.ncbi.nlm.nih.gov/clinvar/varia | PTPN11 | SNV | Missense | Pathogenic | | |

| review | PhenotypeList | Name | ref_aa | alt_aa | pos_aa | CADD_phred | gnomAD_binary_char |
|---|---|---|---|---|---|---|---|
| Criteria prov | Metachondroma | NM_002834. | Thr | Ile | 2 | 23,7 | No |

**Figura 29.** Captura de pantalla de la tabla con el listado de variantes patogénicas o probablemente patogénicas registradas en Simple ClinVar que se encuentran antes de la segunda metionina del gen *PTPN11*. Se puede encontrar en la pestaña "*Start Codon*" de la variante c.1A>C de *PTPN11.*

Para esta evidencia, no existe una pestaña particular en donde recoger la información. Se proporciona una línea informativa en la columna de comentarios de la pestaña "*Evidence*", en relación a su cumplimiento o no *(Figura 30)*.

| Program analysis | | User analysis | | Program analysis |
|---|---|---|---|---|
| Analyzed (met/unmet) | Strenght evidence | Analyzed (met/unmet) | Strenght evidence | Comments |
| met | supporting | NC | | La posició aminoacídica es 2912, i es troba entre un dels dominis funcionals (quinasa o FATC). |

**Figura 30.** Veredicto de la evidencia PM1 de la variante c.8734A>G de *ATM*, presente en la pestaña "*Evidence*".

**PM4**

PM4 solamente es calculada para deleciones o inserciones pequeñas que se encuentran *in-frame*. En caso de que sean variantes *missense*, *silent* o *nonsense*, automáticamente da como veredicto que no se cumple la evidencia. El código utilizado para realizar esto se encuentra en los apartados *Evidence (subapartado PM4),* y *Evidence (II) (subapartado Pathogenic)* del documento Programa1.Rmd. Para su cálculo, primero comprueba si la INDEL es *in-frame* o no. Para ello constata si el número de nucleótidos que se delecionan o se insertan es 3 o múltiplos de 3. Segundo, examina si se genera o no un codón *stop* consultando en Mutalyzer (Name Checker), como hace en el cálculo de PVS1 para mutaciones *frameshift.*

Al igual que con PM1, no existe una pestaña específica para recoger la información de esta evidencia. De modo que también proporciona una línea informativa en la columna de comentarios de la pestaña "*Evidence*", en relación a su cumplimiento o no *(Figura 31)*. Si se trata del gen *TP53,* esta evidencia no es calculada. Para el resto de genes, en caso de comprobarse que se trata de una INDEl *in-frame* que no genera un codón *stop* prematuro, la declara como *met* con una fuerza de "*moderate*".

| Program analisys | | User analysis | | Program analysis |
|---|---|---|---|---|
| Analyzed (met/unmet) | Strenght evidence | Analyzed (met/unmet) | Strenght evidence | Comments |
| met | moderate | NC | | El nombre de nucleòtids inserits o delecions es 3 o múltiple de 3 (nombre nucleòtids: 6) i no es genera un codó stop en el lloc d'inserció o deleció, o acaba generant un frameshift (Mutalyzer: https://mutalyzer.nl/name-checker?description= NM_007294.3%28BRCA1%29%3Ac.2018_2023del) |

**Figura 31.** Fragmento de la tabla que aparece en la pestaña "*Evidence*" con el veredicto de la evidencia PM4 de la variante c.2018_2023del de *BRCA1*.

**PP3, BP4 (predictores de proteína)**

El cálculo de estas evidencias depende del predictor de proteína REVEL. El código utilizado para realizar esto se encuentra en los apartados *Predictores de proteína,* y *Predictores de proteína y algoritmos de conservación (subapartado REVEL)* del documento Programa1.Rmd. Es por ello que, primero, en función

del cromosoma en el que se encuentre la variante, el programa lee el archivo que corresponde con la información de la base de datos dbNSFP. Estas evidencias son calculadas para variantes *missense,* y *silent.* Para las variantes *silent*, automáticamente PP3 no se cumple y BP4 se cumple. Por lo que, solamente dependen de REVEL las variantes *missense*.

Segundo, genera una tabla con los genes en los que están establecidos unos umbrales gen-específico para REVEL, para determinar si la variante es benigna (inferior al valor $T_{BE}$) o patogénica (superior al valor $T_{DE}$). Estos valores son los que se proporcionan en la Tabla 2.

Tercero, el programa realiza la búsqueda de la variante en el archivo de dbNSFP. Para ello, hace coincidir la posición de inicio y final de la variante con la posición de la columna que corresponde al genoma de referencia hg19, y el nucleótido de referencia y el alternativo con sus respectivas columnas. Cuando se encuentre, guarda el valor de REVEL en una variable.

**Tabla 2.** Tabla en la que se resumen los valores *cutoffs* de REVEL gen-específicos.

| Revel Values (cutoffs) | | |
|---|---|---|
| Gene | Value | |
| | $T_{BE}$ | $T_{DE}$ |
| ATM | 0,359 | 0,689 |
| BRCA1 | 0,628 | 0,824 |
| BRCA2 | 0,581 | 0,974 |
| MLH1 | 0,109 | 0,815 |
| MSH2 | 0,562 | 0,862 |
| MSH6 | 0,556 | 0,881 |
| MUTYH | 0,214 | 0,661 |
| NF1 | 0,261 | 0,605 |
| RET | 0,481 | 0,732 |
| TP53 | 0,536 | 0,667 |
| TSC2 | 0,703 | 0,97 |
| General | 0,46 | 0,741 |

Con esta información, procede a calcular si se cumplen o no PP3 y BP4. Para los genes que tienen *cutoffs* específicos, se utilizan los valores de la Tabla 2 para determinar el umbral benigno ($T_{BE}$) y el patogénico ($T_{DE}$). Mientras que para el resto de genes, se utiliza como umbral benigno 0.460 y como patogénico 0.741, que son los umbrales generales. En caso de que el valor de REVEL sea inferior al umbral benigno, se le da predicción de benigna y se cumple BP4. En caso de que el valor sea superior al umbral patogénico, se le da predicción de patogénica y se cumple PP3. En los casos en que el valor de REVEL quede entre ambos umbrales no se le concede ni PP3 ni BP4.

**BP7 (algoritmos de conservación de nucleótido)**

BP7 depende del valor de los algoritmos de conservación de nucleótidos. De modo que también recurre a la información del archivo dbNSFP. Esta evidencia es calculada para variantes *silent*. El código utilizado para realizar esto se encuentra en los apartados *Predictores de conservación de nucleótidos,* y

*Predictores de proteína y algoritmos de conservación (subapartado Algoritmos de conservación)* del documento Programa1.Rmd.

Primero, el programa procede a obtener los valores de los algoritmos de conservación de proteína phyloP, GERP++ y phastCons. Para realizar la búsqueda en el archivo, solamente busca por la posición nucleotídica de la variante en la columna que corresponde a la posición del genoma de referencia hg19. No importa el *match* en los nucleótidos de referencia y alternativos, ya que este tipo de predictor calcula la conservación de la posición nucleotídica, es decir, del nucleótido de referencia. De modo que independientemente del cambio nucleotídico que se produzca, el valor de estos predictores es el mismo. Todos ellos los guarda en una variable.

Segundo, calcula el cumplimiento o no de la evidencia. Para cualquier variante que no sea *silent*, da el no cumplimiento de la evidencia. En caso contrario, para el gen *ATM*, comprueba que el valor phyloP sea inferior a 6.66. De ser menor implica que no está altamente conservada esa posición nucleotídica, implicando el cumplimiento de BP7. Para *PTEN*, verifica los valores tanto de phyloP como de phastCons. Si se cumple que el valor de phyloP sea inferior a 0.1 y el valor de phastCons sea distinto a 1, indica que no es una posición altamente conservada y cumple la evidencia. Finalmente, para el resto de genes, se constata el valor de GERP++. Si este es inferior a 5.5, indica que no está altamente conservado y se cumple la evidencia.

Los valores de los predictores de proteína y de los algoritmos de conservación de nucleótido se muestran en la pestaña "*Protein predictors*" de la Plantilla *(Figura 32)*. En concreto, para los genes *ATM* y *PTEN,* se muestran los algoritmos de conservación de phastCons y phyloP (aunque para *ATM*, solo tiene en cuenta phyloP); para el resto de genes, el valor de GERP++.

| Predictor | Gene | Value | Pathogenicity |
|---|---|---|---|
| revel | ATM | No s'ha trobat | |
| phylop | ATM | 2,027 | Not highly conserved |
| phastcons | ATM | 1 | |

**Figura 32.** Tabla en el que se resumen los valores de los predictores de proteína y de conservación de nucleótidos. Se puede encontrar en la pestaña "*Protein predictors*" de la variante c.609C>T de *ATM*.

### Evidence

La pestaña "*Evidence*", como se ha observado en los apartados anteriores, es la que recoge el veredicto de las distintas evidencias. Como también se ha observado en los ejemplos (e.g. Figuras 30 y 31), rellena la columna de comentarios con un resumen de las razones por las que se ha indicado que se cumple la evidencia. Además, el programa elimina las columnas informativas con los criterios específicos de clasificación de los genes que no interesan, dejando solamente la del gen en el que se encuentra la variante.

Finalmente, en determinados casos, muestra un conjunto de preguntas que el usuario debe responder para verificar el cumplimiento de alguna evidencia (como sucede con BP7) o en otros casos para poder calcular la evidencia (como sucede en PM1 para aquellos genes en que no están establecidas las regiones críticas). La información de la respuesta a estas cuestiones es procesada por el Programa2.

### Classification

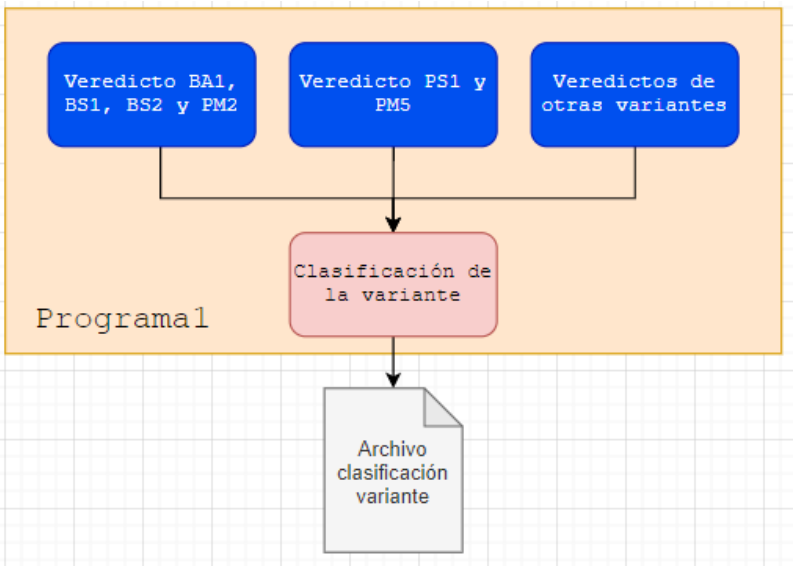



**Figura 33.** Organigrama en el que se muestra el paso final del Programa1, después de obtener los veredictos de todas las evidencias.

Para rellenar esta pestaña, el programa realiza el recuento de las diferentes evidencias que se cumplen con una fuerza determinada *(Figura 33)*: para patogénicas, las que tengan una fuerza de *very strong*, de *strong*, de *moderate*, y de *supporting*; y para las benignas, las que tengan una fuerza de *stand alone*, de *strong*, y de *supporting*. Para ello utilizará la tabla resumen de la pestaña "*Classification Summary*" de 0s y 1s.

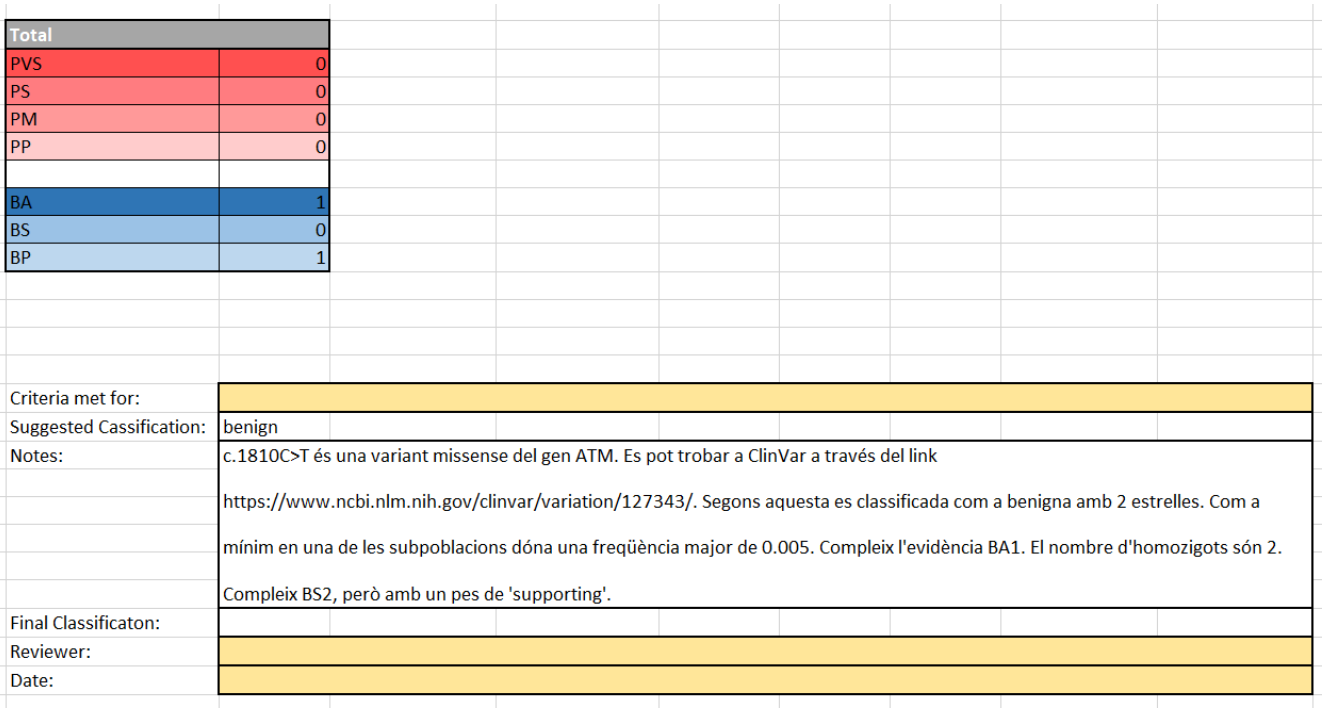
Finalmente, en función del número de evidencias que se cumplan con una determinada fuerza, calcula la clasificación de la variante *(Tabla 1)*.

### Classification Summary

En la pestaña "*Classification Summary*", el programa recoge el cumplimiento o no de las evidencias, siguiendo un código de 0s cuando no se cumple, y de 1s cuando se cumple *(Figura 15)*. También vuelca el recuento de las diferentes evidencias que se cumplen según su fuerza *(Figura 34)*.

Además, en aquellos casos en que el usuario solamente proporciona el símbolo del gen y la variante, rellena el resto de campos de la tabla con los valores que obtiene de Pandora. También el campo de transcriptId lo rellena para aquellas variantes en que calcula PVS1 y realiza la búsqueda en el archivo de LRG.

Finalmente, muestra el veredicto de clasificación y un comentario. Este comentario se va construyendo a medida que se va ejecutando el código con información referente a las evidencias que se van cumpliendo y la razón por la que se cumplen *(Figura 34)*.



| Total | |
|---|---|
| PVS | 0 |
| PS | 0 |
| PM | 0 |
| PP | 0 |
| | |
| BA | 1 |
| BS | 0 |
| BP | 1 |

| | |
|---|---|
| Criteria met for: | |
| Suggested Cassification: | benign |
| Notes: | c.1810C>T és una variant missense del gen ATM. Es pot trobar a ClinVar a través del link https://www.ncbi.nlm.nih.gov/clinvar/variation/127343/. Segons aquesta es classificada com a benigna amb 2 estrelles. Com a mínim en una de les subpoblacions dóna una freqüència major de 0.005. Compleix l'evidència BA1. El nombre d'homozigots són 2. Compleix BS2, però amb un pes de 'supporting'. |
| Final Classificaton: | |
| Reviewer: | |
| Date: | |

**Figura 34.** Captura de pantalla de la pestaña "*Classification Summary*" de la variante c.1810C>T de *ATM*. Se puede observar la tabla con el recuento de las evidencias en función de su fuerza, el veredicto de clasificación y el comentario generado que da soporte a este.

## 4.4 Funcionamiento del ProgramaBateria

El funcionamiento del ProgramaBateria es muy parecido al del Programa1 *(Figura 17)*. Este programa emite el primer veredicto de clasificación, calculando las mismas evidencias que calcula el Programa1. La diferencia es que el ProgramaBateria permite el cálculo en *batch* de este primer veredicto para un conjunto de variantes. Este programa se puede encontrar como material suplementario en formato RMarkdown, bajo el nombre de ProgramaBateria.Rmd.

Teniendo en cuenta la funcionalidad de cada programa, una de las principales diferencias a nivel de código respecto al Programa1 es cómo procesa el *input* en el que se le proporciona la información de las variantes. En el Programa1 esta información se le facilita a través de la Plantilla.xlsx, rellenando los campos de la tabla de la Figura 18. A este otro programa, en cambio, se le proporciona un archivo de texto (*.txt*). Este archivo de texto contiene una fila para cada variante que se quiere analizar, con la información de: el símbolo del gen, el ID del transcrito, la nomenclatura de la variante, la nomenclatura de la proteína, el cromosoma en el que se encuentra, las posiciones de inicio y final de la variante, y los nucleótidos de referencia y alternativos de la variante. Todo ello debe estar separado por tabulador y sin encabezados de columna *(Figura 35)*.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ATM | NM_000051.3 | c.1810C>T | p.P604S | 11 | 108123551 | 108123551 | C | T |
| ATM | NM_000051.3 | c.1899T>G | p.C633W | 11 | 108124541 | 108124541 | T | G |
| ATM | NM_000051.3 | c.2012T>A | p.I671K | 11 | 108124654 | 108124654 | T | A |
| ATM | NM_000051.3 | c.2250G>A | p.K750= | 11 | 108127067 | 108127067 | G | A |
| ATM | NM_000051.3 | c.2362A>C | p.S788R | 11 | 108128319 | 108128319 | A | C |
| ATM | NM_000051.3 | c.2386A>C | p.N796H | 11 | 108129722 | 108129722 | A | C |

**Figura 35.** Ejemplo del formato que debe tener el archivo de texto con las diferentes variantes que se le pasen al ProgramaBateria.

Primero, el programa lee los archivos iniciales. Lee el archivo de texto con la información de las diferentes variantes y los archivos que contienen la información de gnomAD (genoma.xlsx y exoma.xlsx).

Segundo, a partir del archivo con el listado de variantes, el programa va accediendo a la información de cada una de ellas y la guarda en variables. Además, se generan un conjunto de variables con la información de "-", ya que esta información el Programa1 normalmente la obtiene de Pandora. Pero los campos que ocupan en la tabla que se crea con la información de la variante son importantes para calcular el resto de procesos.

Tercero, el programa le proporciona a la función `bucle` las variables con: la información de la variante, el símbolo del gen, la nomenclatura del transcrito, dos de las variables vacías, la nomenclatura de la proteína, otra variable vacía, el cromosoma, la posición de inicio y la de final de la variante, el nucleótido de referencia y el nucleótido alternativo, en este orden específico. La función `bucle` es una función creada en el propio programa que contiene todo el proceso que sigue el Programa1 a partir de la búsqueda de la variante en los archivos genoma.xlsx y exoma.xlsx hasta que realiza el guardado de la información en el *workbook*. Por ello, se le debe proporcionar toda la información de la variante que utiliza el Programa1. Como *output*, la función devuelve el comentario que se va generando de la variante en función de las evidencias que se cumplen.

Una vez realiza la función `bucle` para la primera variante, procede a realizarla con la segunda, y así sucesivamente para todas las variantes del archivo.

Finalmente, se obtiene un conjunto de archivos, uno para cada variante con el veredicto de clasificación y la información de las diferentes bases de datos para declarar cada evidencia. Durante la ejecución del bucle, para cada variante crea una copia del documento Plantilla, la renombra (<`SYMBOL_variant_yyyy-mm-dd.xlsx`>) y la rellena con la información que va obteniendo, del mismo modo que hace el Programa1.

**Validación con variantes clasificadas**

El ProgramaBateria ha sido utilizado para realizar la validación del correcto funcionamiento de la semi-automatización de la clasificación en variantes de cáncer hereditario. Gracias a las 40 variantes clasificadas validadas, brevemente, se ha comprobado que calcula correctamente todas las evidencias según están especificadas en el artículo Feliubadaló, 2020. Obviamente, no se han tenido en cuenta todas aquellas evidencias que no se han automatizado en el programa.

Del total de 40 variantes, en 26 se les predicen todas las evidencias automatizables que se cumplen en el artículo (c.61A>G, c.162T>C, c.609C>T,

c.826A>G, c.1463G>A, c.1810C>T, c.2250G>A, c.2386A>C, c.3802_3802del, c.4060C>A, c.4802G>A, c.4852C>T, c.5623C>T, c.6315G>C, c.6679C>T, c.6848C>T, c.6860G>C, c.7135C>G, c.7191A>G, c.7375C>G, c.7381C>T, c.8734A>G, c.8876_8879del, c.9007_9034del, c.9023G>A, c.9079_9080ins). Destacar que la variante c.7135C>G, era una sustitución en la que el Programa1 calcula el cumplimiento de BP4. Sin embargo, se trata de una variante no canónica en la que se ve afectado el *splicing*. Al ejecutar el Programa2, cumple las evidencias que se indican en el artículo (entre ellas, PP3).

Sin embargo, existen algunas discrepancias en las 14 variantes restantes. En una de ellas (c.1380G>C) no se calculó la evidencia BP7. Esto se debe a que la posición nucleotídica de esta variante no se encuentra en la base de datos dbNSFP. Por consiguiente, no se pudo obtener los valores de los algoritmos de conservación de nucleótidos y calcular BP7. Sin embargo, el resto de evidencias que debía cumplir sí que las calculó correctamente.

En dos variantes se ha declarado que la evidencia PM2 se cumplía, cuando en el artículo se especificaba que no. En ambos casos, se ha comprobado manualmente los valores que proporciona el programa sobre gnomAD y se esperaría que fueran correctos. Una fue la variante c.1564_1565delGA. En este caso, el programa no encontró esta variante en gnomAD y, tomando los valores de las variantes que hay justo antes y después, calculó la evidencia PM2. Sin embargo, esta variante sí que se encuentra en gnomAD. La razón por la que no se ha encontrado es por la dificultad que supone en el caso de deleciones e inserciones pasar la nomenclatura HGVS a nomenclatura de gnomAD. Y aunque se puede llegar a programar, requiere de un código adicional que queda pendiente para implementar en un futuro. La otra variante que calculó PM2 sin cumplirse aparentemente en el artículo es c.2012T>A, la cual la calcula con una fuerza de "*supporting*". Para que se cumpla la evidencia con esta fuerza solamente se tiene en cuenta la FA general. En estos casos se exige una FA igual o inferior a 0.00002 y, teniendo en cuenta los datos de la tabla, se cumple. El resto de evidencias que calcula el artículo para estas variantes son calculadas correctamente por el programa.

En el caso particular de la variante c.998C>T según el artículo se cumpliría la evidencia BS2_Supporting. Sin embargo, el programa indica que no se cumple. Comprobando manualmente, el cálculo del programa a partir de gnomAD está realizado correctamente, ya que para que se cumpla BS2_supporting se exigen como mínimo 2 individuos homocigotos en el subconjunto *non-neuro* y solamente encuentra 1 en gnomAD. La razón por la que en el artículo se calculó esta evidencia es porque para ello utilizaron la información de individuos fenotipados en su propia base de datos, la cual no está al alcance del programa.

Finalmente, quedan por comentar las discrepancias en el cálculo de las evidencias que dependen del predictor de proteína REVEL, que son PP3 y BP4. En relación a estas, se ha observado que en 4 variantes (c.998C>T, c.1899T>G, c.2362A>C, c.5071A>C) el programa indica el cumplimiento de la evidencia BP4 cuando no se encuentra en el artículo; y en 7 (c.4396C>G, c.5558A>T, c.6067G>A, c.6115G>A, c.6203T>C, c.7390T>C, c.8122G>A) el no cumplimiento de PP3 cuando si se indica en el artículo. Las discrepancias en el cumplimiento o no de estas con el laboratorio, se deben principalmente a que para *ATM* utilizaron la combinación de tres predictores para determinar las evidencias (REVEL, VEST4, PROVEAN), ya que se ha comprobado que, en el

caso particular de *ATM*, dan mejores resultados y quedan pendientes de implementar en el código en un futuro. Sin embargo, cabe destacar que en el caso particular de la variante c.5558A>T, se cumple BA1 y, según la combinación de los 3 predictores, PP3. Teniendo en cuenta la fuerza de BA1, no es un resultado lógico obtener también una evidencia patogénica. Por lo que la predicción de REVEL y del programa no dando el cumplimiento de PP3 parece más acertada.

## 4.5 Funcionamiento del Programa2

El Programa2 utiliza la información que le proporciona el usuario después de ver los resultados generados por el Programa1 o el ProgramaBateria *(Figura 17)*. Para ello, primero, el usuario debe comprobar el resultado de la ejecución de alguno de los anteriores programas y si no está de acuerdo con alguna de las evidencias o fuerzas que se le asigna a alguna evidencia, puede modificar la fuerza rellenando las columnas de evidencia y fuerza en la pestaña "*Evidence*" *(Figura 16)*. Además, también puede rellenar estas columnas para aquellas evidencias no automatizables que estime que se cumplen según la información recogida. Finalmente, debe responder todas las preguntas que el programa inicial le habrá indicado y aparezcan al final de la tabla "*Evidence*".

Una vez rellenado, debe introducir en el Programa2 el nombre del archivo de la variante como *file1*. Independientemente que la variante proceda del Programa1 o el ProgramaBateria, siempre se deberá introducir el nombre del archivo de la variante, ya que el rellenado manual siempre se hará individualmente.

A continuación, se detallan los archivos que consulta el Programa2, donde el *file1* debe ser el archivo de la variante que se analiza:

```
file1<-"BRCA1_c.211+1G-A_2020-12-13.xlsx"
file2<-"genoma.xlsx"
file3<-"exoma.xlsx"
file4<-"BLOSUM62_probabilities.csv"
file5<-"tsv_exomes.txt"
file6<-"tsv_genomes.txt"
file7<-"revel.txt"
file8<-"phylop.txt"
file9<-"LRG_genes.txt"
```

Después de la intervención del usuario, es el turno del programa. Primero, lee el archivo por la pestaña "*Evidence*", ya que es en esta donde se recoge la información proporcionada por el usuario. También lee y carga el mismo archivo para poder ir escribiendo la nueva información.

A partir de aquí el programa está dividido en diferentes apartados, uno para cada una de las evidencias. El código empleado se puede consultar en el archivo Programa2.Rmd, proporcionado como material suplementario. En general, para todas las evidencias el programa prioriza la información de *met/unmet* y fuerza de la evidencia que proporcione el usuario. En caso de que no haya rellenado las columnas, utiliza, si existe, la información de *met/unmet* y fuerza de la evidencia que hubiera calculado uno de los dos programas anteriores. Finalmente, tendrá en cuenta la respuesta a las preguntas, cuando la haya, para recalcular algunas evidencias ya calculadas por el Programa1 o ProgramaBateria, o para poder calcular nuevas evidencias.

Para la evidencia PVS1, si se indica que se ve afectado el *splicing*, el usuario proporciona el nombre de la variante en RNA que se predice o se ha hallado experimentalmente. En este caso, procede a comprobar si se genera o no algún codón *stop* prematuro. Para ello utiliza el nombre de la variante en RNA (*RNAannot*) como si fuera una anotación del cDNA y utiliza la nomenclatura del transcrito en LRG (*transcritLRG*). Con ello genera el *link*

`https://mutalyzer.nl/name-checker?description=`**`transcritLRG`**`%3A`**`RNAannot`**

que utiliza en Mutalyzer (Name Checker) para buscar si se genera o no un codón *stop*. A partir de aquí, calcula si se da NMD o no del mismo modo que lo calculaban los otros dos programas para mutaciones *nonsense* y *frameshift*. También recoge la información en la pestaña "*NMD*" como hacían los otros dos programas.

Para PS1, tiene en cuenta las respuestas que haya respondido el usuario. Para *TP53*, si el usuario indica que existen variantes patogénicas en ClinGen TP53 en esa posición, que se ve afectado el *splicing*, y especifica que los datos utilizados son datos experimentales de RNA indica que se cumple con fuerza "*strong"*. En cambio, si los datos provienen de modelos *in silico*, le da una fuerza "*moderate".* Para el resto de genes, si el programa inicial indicaba que cumplía PS1, comprueba también las respuestas del usuario. Si se ve afectado el *splicing* y comprueba que los modelos in silico predicen que el impacto en la variante es igual o mayor que variantes de *splicing* patogénicas en la misma posición, para *ATM* cumple la evidencia con fuerza "*supporting"*, y para el resto la cumple con una fuerza "*strong"*. En caso de que no se vea afectado el *splicing*, da la fuerza que le daba el Programa1 o ProgramaBateria.

Para PM1, cuando no se ha podido calcular, se pregunta al usuario si se ve afectado algún *hot-spot* o dominio de la proteína. En caso de que indique que sí, cumple la evidencia.

Para la evidencia PM5, el programa solamente realiza las preguntas para *TP53.* Se le pregunta si se encuentra en una zona *hot-spot*, si se afecta el *splicing* y si se encuentran 2 o más variantes patogénicas en ClinGen TP53 con una puntuación de sustitución de BLOSUM62 igual o mejor que la variante que se clasifica. En caso de que no se encuentre en una zona *hot-spot,* que afecte el *splicing*, y que exista alguna variante de *splicing* patogénica, le dará un peso de "*moderate"*.

Cuando no se cumpla PP3, se le pregunta al usuario si se ve afectado el *splicing* según los predictores de *splicing.* En caso de que diga que sí, se cumple PP3.

Para BP4, también se le pregunta si se ve afectado el *splicing* según los predictores de *splicing,* tanto si el programa inicial predecía que se cumplía como si no. En caso de que no se vea afectado el *splicing* para las variantes *silent*, se cumple BP4. Para las *missense* toma el resultado del primer programa. Si se ve afectado el *splicing*, no se cumple.

En BS2, para *TP53* se le pregunta al usuario el número de homocigotos que encuentra en una muestra de población femenina sana de más de 60 años. Cuando da un valor superior o igual a 8, se determina que se cumple la evidencia con un peso "*strong".* Cuando el valor que se da es entre 2 y 7, se cumple también, pero con un peso de "*supporting"*.

Finalmente, en el cálculo de BP7 en algunos genes aparecen preguntas para confirmar su cumplimiento. Para *CDH1*, se le pregunta si la variante es el nucleótido de referencia en un primate y/o en más de 3 mamíferos. En caso de que no lo sea no se cumple. Para el resto de genes, se le pregunta si se ve afectado el *splicing.* En caso de que se vea afectado, no se cumple la evidencia.

Para acabar el programa procede a hacer el recuento de evidencias que se cumplen y a dar el veredicto final de clasificación del mismo modo que lo hace el Programa1. El resultado del *workbook* lo guarda en el archivo que se le ha proporcionado inicialmente de la clasificación de la variante.

## 4.6 Elaboración del Manual del usuario

El Manual del usuario se puede encontrar como material suplementario y consta de dos partes: Contenido y Funcionamiento. En el apartado de Contenido se explica brevemente cada uno de los archivos que formarán parte del archivo *.exe*. En esta explicación, se proporciona la fuente utilizada para elaborarlo, qué información contendrá cada archivo y la función que desempeñarán, sin entrar en detalles en el proceso de obtención de ellos.

En el apartado de Funcionamiento se explica cómo utilizar cada uno de los programas para la clasificación de variantes. Es por ello que este queda dividido en tres subapartados: una única variante (cuando se utilice el Programa1); batería de variantes (cuando se utilice el ProgramaBateria); y modificación de los resultados (cuando se ejecute el Programa2).

En cada uno de estos subapartados de Funcionamiento, se le explica al usuario qué *input* debe proporcionarle al programa, y la forma y el formato en el que se debe proporcionar. Además, se explica brevemente el *output* que obtendrá. Sin embargo, no se entra en detalles de cómo los diferentes programas obtendrán la información y calcularán las evidencias y el veredicto de clasificación.

# 5. Conclusiones

En este proyecto, se ha desarrollado una herramienta de uso práctico en laboratorios de diagnóstico que permite semi-automatizar el Proceso de Clasificación de Variantes en Cáncer Hereditario.

Con ese fin se han abordado tres aspectos generales:

1. El desarrollo de un programa para clasificar variantes a partir de evidencias automatizables.

2. El desarrollo de un programa para reclasificar las variantes tras la revisión de las automatizadas y adición de información por el usuario.

3. La documentación del código y creación de un manual dirigido al usuario.

En relación al primer punto:

- (a): Se ha desarrollado un programa que calcula todas las evidencias automatizables para una variante de tipo *silent, missense, nonsense, frameshift,* pequeñas deleciones e inserciones *in-frame* y variantes que afectan en el codón de inicio, y emite un veredicto de clasificación para ella en función de las evidencias que cumple.

- (b): Se ha desarrollado un segundo programa que evalúa todas las evidencias automatizables para una lista de variantes en *batch*, y emite un veredicto de clasificación para cada una de ellas.

En relación al segundo punto:

- (a): Se ha desarrollado un tercer programa que 1) permite al usuario revisar y completar el resultado generado de alguno de los programas anteriores, 2) (re)evalúa evidencias y 3) emite un veredicto de clasificación.

En relación al tercer punto:

- (a): Se ha documentado el código de toda la herramienta en formato RMarkdown para facilitar la lectura, revisiones y actualizaciones posteriores.

- (b): Se ha elaborado un manual dirigido al usuario para utilizar correctamente la herramienta.

**Reflexiones finales: logro de objetivos**

Se han logrado alcanzar todos los objetivos, a excepción de la programación de las evidencias PS3 y BS3. Estas dos corresponden a evidencias que dependen de la búsqueda bibliográfica de información para determinar su cumplimiento, no siendo, efectivamente, evidencias automatizables. Sin embargo, en este proyecto lo que se pretendía era automatizar la búsqueda de artículos bibliográficos en los que apareciera la variante, para poder proporcionarlos al usuario y facilitarle su búsqueda.

La inversión de más de una semana en la elaboración del archivo de la base de datos dbNSFP implicó ampliar unos días la fecha límite del cronograma y descartar este objetivo. La razón principal fue que se creyó el menos relevante

por cumplir una función informativa y no ser una evidencia automatizable como tal. Además, en los casos en que se encuentra la variante en ClinVar, se le proporciona al usuario el listado de citas bibliográficas de esta web a modo informativo. Por lo que, indirectamente estaría cumpliendo la función que se le quería dar al objetivo de BS3 y PS3.

**Reflexiones finales: planificación y metodología**

Dos imprevistos durante el desarrollo, obligaron a reorganizar el cronograma para garantizar el cumplimiento de la mayor parte del proyecto. Primero, se tuvieron que posponer las evidencias PP3 y BP4, porque para la fecha prevista todavía no se tenía decidido qué predictor de proteína se iba a utilizar. Al final se acabaron realizando después de la elaboración del archivo dbNSFP. Este cambio supuso poder avanzar antes con la evidencia PVS1, la cual era un gran bloque de trabajo. Segundo, se tuvo que regenerar el/los archivo/s de la base de datos de dbNSFP, porque el filtrado no recogía las zonas de interés. Durante las prácticas, este archivo había sido generado con R. Pero al partir de diferentes archivos de gran tamaño (uno por cromosoma), se tuvieron que ir leyendo por fragmentos, aunque supusiera una gran inversión de tiempo y una tarea bastante tediosa.

Sin embargo, al tener que reelaborarlo en este proyecto y el tiempo que suponía trabajando desde R, se buscaron alternativas para reducir el tiempo que se debía de invertir y afectar en lo menor posible el cronograma inicial. Una de las opciones que se planteó fue comprobar si era posible desde la línea de comandos de Linux y así se hizo. Aunque todo el proceso se podría haber realizado con un menor número de pasos, se decidió hacer de la forma que se ha detallado en el Anexo 8.2 para optimizar los procesos de lectura y agilizar el tiempo de ejecución general. Además, se decidió crear un archivo a cada paso para que, en caso de ser necesario modificar algún archivo final, se pueda disponer de todos los intermediarios y poder partir del archivo que más convenga. También se generó un archivo por cromosoma por dos razones: para agilizar la carga en R durante la ejecución del programa; y para que, en caso de que se añada o se elimine algún gen en el diagnóstico, solamente se deba realizar el proceso en el cromosoma en el que se encuentre el gen.

A pesar de reducir el tiempo para elaborar el archivo, el tiempo global invertido fue considerable y supuso reorganizar el plan de trabajo inicial. Paralelamente a su elaboración, se estuvo implementando la programación de la evidencia PM4, por no depender de este archivo, intentando mitigar el retraso. Una vez generado el archivo, se pudieron finalizar las evidencias PVS1, PP3, BP4, y BP7, que dependían de información contenida en esta base de datos.

Otra modificación respecto a la planificación inicial fue la realización de la tarea en la que se debía realizar el recuento de las evidencias que se cumplían y dar el veredicto final de clasificación, tanto en el Programa1 como en el Programa2. Esta tarea se dejó para el final para poder disponer del suficiente número de evidencias, a partir de las cuales pueda dar un veredicto real de clasificación.

La última modificación respecto a la metodología fue crear tres programas distintos. Desde un inicio, se pensó en la realización de dos programas: el principal que calcularía la mayoría de evidencias automatizables, y el secundario que, a partir de la información del primero y la información que proporcionase el usuario, calcularía nuevas evidencias y daría el nuevo veredicto. Sin embargo, al querer, implementar un método que permitiera

proporcionarle diferentes variantes en *batch* surgió la idea de generar un tercer programa, aunque fuera posible hacerlo en uno solo. Aunque es más difícil mantener en paralelo dos programas con el mismo código (solamente difieren en la forma de obtención de la información de la variante), de haberse creado solamente un programa habría sido el ProgramaBateria. Sin embargo, la razón principal por la que se decidió implementar tres programas es que al crear el ProgramaBateria se pierde la estructura por títulos y *chunks* en RMarkdown que tiene el Programa1 para cada apartado, la cual es muy útil, especialmente para poder encontrar fácilmente la zona del código que se quiere modificar cuando hay un error o se quiere implementar algo nuevo. También, el mantener los dos programas facilita después encontrar la misma zona en el ProgramaBateria, ya que puedes saber que hay antes y después del código para localizarlo.

En cuanto a los recursos utilizados, uno de los grandes cambios que se han hecho respecto a las prácticas fue el cambiar la fuente de obtención de la información de variantes que se producen en el mismo codón. Inicialmente se utilizaba la web de ClinVar Miner. Desde esta, el programa realizaba una búsqueda de entre todas las variantes registradas del gen, aquellas que afectaran al mismo codón que la que se estaba analizando. Sin embargo, se comprobó que había una forma mucho más directa de obtener el listado de estas variantes desde ClinVar. Además, a través de ClinVar, se obtiene directamente el listado de nomenclaturas de la variante, el cual se obtenía antes desde la herramienta Position Converter de Mutalyzer [38]. Finalmente, y no menos importante, también se obtiene una tabla informativa de un listado de citas bibliográficas en las que aparece la variante que se analiza. Por lo que en conjunto con este cambio se redujo en aproximadamente un 50% el tiempo de ejecución de una variante, pasando de 8.53 minutos a 4.30 minutos; además de obtener información extra de utilidad para el usuario.

Por último, se decidió utilizar REVEL como predictor de proteína, porque es un metapredictor que posee una precisión que supera a la de los otros predictores o metapredictores con los que se ha comparado. Además, REVEL se ha calibrado clínicamente, utilizando variantes clasificadas como patogénicas o probablemente patogénicas, y benignas o probablemente benignas de un total de 66 genes. Así se logró que los umbrales para BP4 y PP3 se correspondieran con probabilidades de patogenicidad de 0.2 y 0.8, respectivamente. De este modo, para 20 genes se generaron unos umbrales específicos de patogenicidad, por disponer de 10 o más variantes benignas o probablemente benignas ya clasificadas, y 10 o más variantes patogénicas o probablemente patogénicas para realizar el análisis. Para el resto de genes, se calcularon unos umbrales generales [7].

Específicamente, para *ATM* se ha demostrado que, en función del dominio de la proteína que se analiza, la combinación específica de este predictor junto con dos más (VEST4 y PROVEAN) obtiene mejores resultados [1]. Sin embargo, por falta de tiempo, se decidió renunciar a esta implementación específica que solo afecta a un gen, dejándola para más adelante.

**Líneas de trabajo futuras**

Se ha desarrollado una herramienta útil, funcional, relativamente rápida y fácil de utilizar por cualquier usuario, que se podría empezar a aplicar en el diagnóstico clínico. A continuación, se recogen algunas de las mejoras y extensiones que se podrían introducir.

Sería interesante implementar la búsqueda de citas bibliográficas para las evidencias de BS3 y PS3. En esta búsqueda se podría incorporar una función que realizase el recuento de las veces que apareciese la variante en cada fuente (artículo o DB), y si esta se localiza en el título, *abstract* o cuerpo del artículo.

Se podría añadir la generación de comentarios en la pestaña "*Evidence*" para las evidencias que calcule o recalcule el Programa2, como ya hace el Programa1.

Se debería finalizar el cálculo de evidencias para las variantes intrónicas, dado que para ellas solamente se calcula PVS1.

En *TP53,* para algunas evidencias que ya están automatizadas para el resto de genes, se utilizan bases de datos específicas que no han sido automatizadas. En estos casos, como ocurre con BS2_Supporting, se requiere de la información del usuario para comprobar si se cumple o no. Se propone automatizar la obtención de esta información de sus respectivas fuentes.

Una adición que ahorraría un esfuerzo considerable al usuario sería la automatización de la predicción de alteraciones del *splicing*, que acaba impactando en varias evidencias. Actualmente, se requiere que el usuario indique si está afectado o no y suministre la predicción del transcrito alterado. Una posible opción sería utilizar SpliceAI [24], que ha mostrado los mejores resultados publicados hasta la fecha.

Dado que frecuentemente se establecen criterios personalizados para nuevos genes, sería recomendable ir implementando los criterios específicos para cada evidencia. Ejemplos actuales son: *CDH1* [27]*, PTEN* [28]*, ATM* [1]*, CHEK2* [29] y *TP53* [30]*.*

Finalmente, el objetivo a medio plazo del laboratorio es integrar esta automatización, con las adaptaciones necesarias, a Pandora, para completar las funciones actuales de LIMS (Laboratory Information Management System) del laboratorio, la GUI (Graphical User Interface) del *pipeline* de la NGS (Next-Generation Sequencing) y la base de datos (DB) de las variantes del laboratorio.

# 6. Glosario

| | |
|---|---|
| **AC** | Allele Count |
| **ACMG** | American College of Medical Genetics |
| **AF** | Allele Frequency |
| **AMP** | Association for Molecular Pathology |
| **AN** | Allele Number |
| **BLOSUM62** | BLOcks SUbstitution Matrix 62 |
| **BS** | Benign Strong |
| **BP** | Benign Supporting |
| ***BRCA1*** | BReast CAncer type 1 |
| ***CDH1*** | CaDHerin-1 |
| **cDNA** | complementary DNA |
| **CDS** | CoDing Sequence |
| ***CHEK2*** | CHEckpoint Kinase 2 |
| **CIBERONC** | Centro de Investigación Biomédica en Red Cáncer |
| **ClinGen** | Clinical Genome |
| **DB** | DataBase |
| **dbNSFP** | database of human Nonsynonymous SNVPs and their Functional Predictions |
| **DNA** | DeoxyriboNucleic Acid |
| **FATC** | Focal Adhesion Targeting Carboxyterminal domain |
| **FHA** | ForkHead-Associated domain |
| **gnomAD** | genome Aggregation Database |
| **GRCh37** | Genome Reference Consortium human build 37 |
| **GRCh38** | Genome Reference Consortium human build 38 |
| **GUI** | Graphical User Interface |
| **hg18** | Human Genome reference build 18 |
| **hg19** | Human Genome reference build 19 |
| **hg38** | Human Genome reference build 38 |
| **IC** | Intervalo de Confianza |
| **IDIBELL** | Institut D'Investigació biomédica de BELLvitge |
| **LIMS** | Laboratory Information Management System |
| **LOF** | Loss Of Function |
| **LRG** | Locus Reference Genomic |
| **MAF** | Minor Allele Frequency |
| **NGS** | Next-Generation Sequencing |
| **nhomalt** | número de homocigotos para el alelo alternativo |
| **NIH** | National Institutes of Health |
| **NMD** | Nonsense-Mediated mRNA Decay |
| **pb** | Pares de Bases |
| **PM** | Pathogenic Moderate |
| **PMID** | PubMed IDentifier |
| **PP** | Pathogenic Supporting |
| **PS** | Pathogenic Strong |
| ***PTEN*** | Phosphatase and TENsin homolog |
| **PVS** | Pathogenic Very Strong |
| **RNA** | RiboNucleic Acid |
| **SNVP** | Single-Nucleotide Variant Proportion |
| **TFM** | Trabajo Final de Máster |
| ***TP53*** | Tumor protein P53 |
| **UCSC** | Universitiy of California, Santa Cruz |

# 7. Bibliografía


1. Feliubadaló, L., et al. (2020). A Collaborative Effort to Define Classification Criteria for *ATM* Variants in Hereditary Cancer Patients. *Clinical Chemistry*, hvaa250: 1-16.
2. Richards, S., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17*(5)*: 405-424.
3. Furness, L.M.. (2017). Bridging the gap: the need for genomic and clinical -omics data integration and standardization in overcoming the bottleneck of variant interpretation. *Expert Review of Precision Medicine and Drug Development,* 2*(2)*: 79-89.
4. Collins, R.L. et al. (2020). A structural variation reference for medical and population genetics. *Nature*, 581: 444-451.
5. Rehm, H.L., et al. (2015). ClinGen – The Clinical Genome Resource. *The New England Journal of Medicine,* 372: 2235-2242.
6. Liu, X. et al. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. *Hum Mutat,* 37*(3)*: 235-241.
7. Tian, et al. (2019). REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Nature, 9: 12752, DOI: https://doi.org/10.1038/s41598-019-49224-8.
8. Henrie, A. et al. (2018). ClinVar Miner: Demonstrating Utility of a Web-Based Tool for Viewing and Filtering ClinVar Data. *Hum Mutat,* 39*(8)*: 1051-1060.
9. Landrum, M.J., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res,* 42: D980-D985.
10. Lefter, M. et al. (2020), Mutalyzer 2: Next Generation HGVS Nomenclature Checker. Preprint from bioRxiv, DOI: https://doi.org/10.1101/2020.06.24.168583.
11. Aken, B.L., et al. (2016). The Ensemble gene annotation system. *Database (Oxford)*, baw093, doi: 10.1093/database/baw093.
12. Karolchik, D., et al. (2012). The UCSC Genome Browser. *Curr Prolog Hum Gene*t, doi: 10.1002/0471142905.hg1806s71.
13. Perez-Palma, E., et al. (2019). Simple ClinVar: an interactive web server to explore and retrive gene and disease variants aggregated in ClinVar database. *Nucleic Acids Res.*, 47: W99-W105.
14. Whiffin, N., et al. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Nature, 19*, 1151-1158.
15. Henikoff, S., Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89*(22)*: 10915-9.
16. MacArthur, J.A.L., et al. (2014). Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.,* 42: D873-D878.
17. Campbell, M. 2020, Missense, Nonsense and Frameshift Mutations: A Genetic Guide, accessed 27 december 2020, <https://www.technologynetworks.com/genomics/articles/missense-nonsense-and-frameshift-mutations-a-genetic-guide-329274>.
18. Dufner-Almeida, L.G., et al. (2019). Understanding human DNA variants affecting pre-mRNA splicing in the NGS era. *Adv Genet*, 103: 39-90.
19. Parsons, M., et al. (2013). Consequences of germline variation disrupting the constitutional translational initiation codon start sites of *MLH1* and *BRCA2*: use

of potential alternative start sites and implications for predicting variant pathogenicity. *Mol Carcinog*, 54*(7)*: 513-522.
20. Armour, J.A.L., et al. (2002). The detection of large deletions or duplications in genomic DNA. *Hum Mutat*, 20*(5)*: 325-337.
21. Tayoun, A.N.A., et al. (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMPvariant criterion. *Human Mutation, 39*: 1517-1524.
22. Feliubadaló, L., et al. (2017). Benchmarking of Whole Exome Sequencing and Ad Hoc Designed Panels for Genetic Testing of Hereditary Cancer. *Sci Rep*, 7: 37984.
23. Feliubadaló, L., et al. (2019). Opportunistic testing of *BRCA1, BRCA2* and mismatch repair genes improves the yield of phenotype driven hereditary càncer gene panels. *Int J Cancer*, 145*(10)*: 2682-2691.
24. Jaganathan, K., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176: 535-548.
25. Heidenbrand, J.R., et al. (2019). Recommendations for performance optimizations when using GATK3.8 and GATK4. *BMC Bioinformatics*, 20: 557.
26. Gilks, W.P., et al. (2016). Whole genome resequencing of a laboratory-adadpted *Drosophila melanogaster* population sample. *F1000Research,* 5: 2644.
27. Lee, K., et al. (2018). Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline *CDH1* sequence variants. *Hum Mutat*, 39*(11)*: 1553-1568.
28. Mester, J.L., et al. (2018). Gene-specific criteria for *PTEN* variant curation: Recommendations from the ClinGen *PTEN* Expert Panel. *Hum Mutat*, 39*(11)*: 1581-1592.
29. Vargas-Para, G., et al. (2020). Comprehensive analysis and ACMG-based classification of *CHEK2* variants in hereditary càncer patients. *Hum Mutat*, 41*(12)*: 2128-2142.
30. Fortuno, C., et al. (2020). Specifications of the ACMG/AMP variant interpretation guidelines for gemrline TP53 variants. *Hum Mutat*, DOI: https://doi.org/10.1002/humu.24152.
31. Gentleman, R. R Programming for Bioinformatics. 1st Edition. *Chapman and Hall/CRC;* 2008, 362p.
32. Rubba, C. (2020). htmltab: Assemble Data Frames from HTML Tables. R package version 0.7.1.1. https://CRAN.R-project.org/package=htmltab.
33. Conway, J., et al. (2017). RPostgreSQL: R Interface to the 'PostgreSQL' Database System. R package version 0.6-2. https://CRAN.R-project.org/package=RPostgreSQL.
34. Allaire, J.J., et al. (2020). rmarkdown: Dynamic Documents for R. R package version 2.1. URL https://rmarkdown.rstudio.com.
35. Mirai Solutions GmbH. (2020). XLConnect: Excel Connector for R. R package version 1.0.1. https://CRAN.R-project.org/package=XLConnect.
36. Shotts, W.E, Shotts, W.E. Jr. The Linux Command Line: A Complete Introduction. 2nd Edition. *Random House LCC US;* 2009, 482p.
37. Signorell, A., et al. (2020). DescTools: Tools for descriptive statistics. R package version 0.99.38.
38. Freeman, et al. (2018). VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. Hum Mutat, 39(1): 61-68.

# 8. Anexos

## 8.1. Generación del archivo LRG

A continuación se muestra el código desarrollado para la generación del archivo LRG_genes.txt:

```
LRG<-read.table("list_LRGs_transcripts_GRCh37.txt",
header= TRUE, sep="\t")
LRG_filt<-data.frame()
i<-1
while(i<=nrow(genes)){
 LRG_vector<- LRG[which(LRG$HGNC_SYMBOL==genes[i,1]),]
 LRG_filt<-rbind(LRG_filt, LRG_vector)
 i<-i+1
}
LRG_filt<-LRG_filt[,c(1,2,7,8)]
write.table(LRG_filt,  "./LRG_genes.txt",  col.names  =
TRUE, row.names = FALSE, sep="\t")
```

## 8.2. Procedimiento para la elaboración de los archivos de la base de datos dbNSFP

A continuación se detallan los pasos seguidos para la generación de los archivos de la base de datos dbNSFP:

- Se genera un archivo para cada cromosoma a partir del archivo BED (tomando como ejemplo el cromosoma 1):

```
gawk '{if($1 == "1") print $0;}' archivobed.bed > bed1.txt
```

- Se filtra el archivo de dbNSFP de cada cromosoma con las columnas que interesan y se guardan en un nuevo documento de texto:

```
cut    -f1-24,    29-31,    37-84,    86-155,    364-372
./dbNSFP4.0a_variant.chr1 > cromosoma1.txt
```

- Se guarda en un nuevo archivo la columna 9 con los valores de posición del ensamblaje genómico de hg19:

```
cut -f9 ./cromosoma1.txt > cromosoma1_columna.txt
```

- Se elimina el nombre de las columnas y se ordena el archivo utilizando los valores de la columna 9. Se guarda en un nuevo documento:

```
tail -n+2 cromosoma1.txt | sort -k9,9 > cromosoma1_ord.txt
```

- El archivo ordenado según hg19 se procesa con R para filtrarlo a partir del archivo *.BED* del cromosoma:

```
bed_1<-read.table("J:/Documentos/Estudios/Curso      2019-
2020/Máster bioinformática y bioestadística/Pràctiques en
```

```
empresa/projecte/plantilla/dbnsfp/bed1.txt", header= FALSE,
sep="\t")
chr1<-read.table("J:/Documentos/Estudios/Curso 2019-
2020/Máster bioinformática y bioestadística/Pràctiques en
empresa/projecte/plantilla/dbnsfp/cromosoma1_columna.txt",
header= FALSE, sep="\t")
chr1<-as.numeric(as.character(chr1[-1,1]))
chr1<-as.data.frame(chr1)
colnames(chr1)<-"position"
i<-1
vector<-data.frame()
while(i<=nrow(bed_1)){
  value<-t(t(chr1[which(chr1$position >= bed_1[i,2] &
chr1$position <= bed_1[i,3]),]))
  vector<-rbind(vector, value)
  i<-i+1
}
write.table(vector, "./vector1.txt", col.names = TRUE,
row.names = FALSE, sep="\t")
```

- Desde la línea de comandos de Linux, se ordena el archivo generado en R con las posiciones de hg19 filtradas con el archivo *.BED*, y se seleccionan solamente los valores únicos:

```
sort vector1.txt | uniq > vector_ord.txt
```

- Este archivo se vuelve a procesar con R para añadirle una columna numerando las diferentes posiciones:

```
vector_1<-read.table("J:/Documentos/Estudios/Curso 2019-
2020/Máster bioinformática y bioestadística/Pràctiques en
empresa/projecte/plantilla/dbnsfp/vector_ord1.txt", header=
FALSE, sep="\t")
i<-1
vector<-data.frame()
while(i<=nrow(vector_1)){
  value<-cbind(as.numeric(as.character(vector_1[i,1])), i)
  vector<-rbind(vector, value)
  i<-i+1
}
write.table(vector, "./vector1_ord_filtrat.txt", col.names
= TRUE, row.names = FALSE, sep="\t")
```

- Se seleccionan las filas del archivo que contiene las variantes del cromosoma en las que la columna 9 coincide con los valores del archivo generado en R:

```
join -1 9 -2 1 -i -t $'\t' cromosoma1_ord.txt
vector_ord_filtrat.txt > cromosoma1_ord_filtrat.txt
```

- Desde *Excel*, se pasa el formato *.txt* a *.xlsx*, ya que tiene un menor peso. En conjunto, con todos los cromosomas, se pasa de 3.09Gb a 1.2Gb.

## 8.3. Software utilizado en el proyecto

En la siguiente tabla se muestra todo el software utilizado para el desarrollo de este proyecto:

| Software | Version | Link |
|---|---|---|
| gnomAD | 2.1.1 | https://gnomad.broadinstitute.org/downloads/ |
| dbNSFP | Version 4.0a (5/12/2019) | https://sites.google.com/site/jpopgen/dbNSFP |
| LRG | 11/05/2020 | https://www.lrg-sequence.org/data/#lrg-data |
| Simple ClinVar | 05/12/2020 | http://simple-clinvar.broadinstitute.org/ |
| Archivo BED | v2.2+v2.3 | Propio del departamento |
| R | R-3.6.1 | https://cran.r-project.org/bin/windows/base/old/3.6.1/ |
| RMarkdown | v2.1 | https://rmarkdown.rstudio.com/ |
| RStudio | v1.1.463 | https://rstudio.com/products/rstudio/download/ |
| Linux | Ubuntu 18.0 | https://releases.ubuntu.com/18.04/ |
| Excel | 16.0 | https://www.microsoft.com/en-us/download/office.aspx |