Universitat Oberta de Catalunya

uoc.edu

PGVWeb: Aplicación web para la priorización de variantes genéticas en un contexto clínico

Carmen Núñez García

Máster en Bioinformática y Bioestadística

Bioinformática clínica

Guerau Fernández Isern

Marc Maceira Duch

05/01/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*PGVWeb: Aplicación web para la priorización de variantes genéticas en un contexto clínico*
Nombre del autor:	*Carmen Núñez García*
Nombre del consultor/a:	*Guerau Fernández Isern*
Nombre del PRA:	*Marc Maceira Duch*
Fecha de entrega (mm/aaaa):	05/01/2021
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Bioinformática clínica*
Idioma del trabajo:	*Castellano*
Palabras clave	*variante genética, Next-Generation Sequencing (NGS), herramienta web*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

La secuenciación de segunda generación ha permitido un gran avance en el diagnóstico clínico. Mediante la detección de variaciones genéticas es posible determinar la causa de una patología en un gran número de ocasiones. No obstante, la gran cantidad de variaciones genéticas que se detectan dificulta la labor que llevan a cabo los analistas. Por ello, el diseño de una interfaz gráfica sencilla e intuitiva que facilite el manejo y visualización de las variantes se hace imprescindible. Otro de los aspectos indispensables en el diagnóstico clínico es la priorización de variantes candidatas, es decir, priorizar aquellas variantes con más probabilidad de ser la causa de la patología.

En este Trabajo de Fin de Máster se ha desarrollado una aplicación web, en lenguaje de programación R, que permite importar y visualizar, de manera interactiva, todas las variantes genéticas detectadas mediante la secuenciación. Además, se han incorporado filtros y algoritmos de priorización para mostrar en primer lugar las variantes que tienen más probabilidad de ser causantes de patología.

Para llevar a cabo este proyecto se han utilizado diversos paquetes de R, entre los que destaca el paquete *Shiny*, empleado para crear una interfaz de usuario interactiva. También se han empleado diferentes bases de datos como *ClinVar, HPO* y *OMIM* que hacen posible generar filtros específicos y priorizar en función del fenotipo del paciente.

Como resultado de este Trabajo, se ha desarrollado una aplicación basada en web que combina sencillez de uso y una gran flexibilidad, con el objetivo de facilitar y enfocar el análisis a variantes causantes de enfermedad.

Abstract (in English, 250 words or less):

Next-generation sequencing has led to a breakthrough in clinical diagnosis. Detection of genetic variations allows to determine the cause of a pathology of large number of cases. However, the huge amount of genetic variants detected difficult the task of analysts. Therefore, the design of a simple and intuitive graphical interface that facilitates the operation and visualization of variants becomes essential. Another indispensable aspect in clinical diagnosis is the prioritization of candidate variants, that is, prioritizing those variants most likely to be the cause of the pathology.

In this project we have developed a web application in R programming language, which allows you to import and visualize, interactively, all the genetic variants detected by sequencing. In addition, filters and prioritization algorithms have been added to show the variants that are more likely to be the cause of pathology.

Various R packages have been used to carry out this project, including the *Shiny* package, used to create an interactive user interface. Different databases have also been used such as *ClinVar*, *HPO* and *OMIM* that make it possible to generate specific filters and prioritize based on the patient's phenotype.

As a result of this work, a web application has been developed that combines simplicity of use and great flexibility in order to facilitate and focus the analysis on disease-causing variants.

Índice

1. Introducción 1
1.1. Contexto y justificación del Trabajo 1
1.1.1 Contexto 1
1.1.2 Justificación 3
1.2. Objetivos del Trabajo 4
1.3. Enfoque y método seguido 4
1.4. Planificación del Trabajo 5
1.4.1. Tareas 5
1.4.2. Calendario 6
1.4.3. Hitos 7
1.4.4. Análisis de riesgos 8
1.5. Resultados esperados 9
1.6. Breve descripción de los otros capítulos de la memoria 9
2. Aspectos generales y estado del arte 10
3. Datos de partida 13
3.1. Formato VCF 13
3.2. Anotación de variantes 15
3.3. Cálculo de cobertura 16
4. Diseño de la aplicación web 17
4.1. Paquetes de R más relevantes 17
4.2. Front-end 19
4.2.1. Panel de opciones 20
4.2.2. Panel principal 21
4.3. Back-end 23
5. Importación y visualización de los datos 26
6. Filtrado y priorización de variantes 29
6.1. Filtrado de variantes 29
6.2. Priorización de variantes 30
6.2.1. Implementación del algoritmo de priorización 35
7. Pruebas de funcionalidad 37
7.1. Caso de prueba 1 37

7.2. Caso de prueba 2 38
7.3. Caso de prueba 3 39
7.4. Caso de prueba 4 40
8. Pruebas de rendimiento 41
9. Conclusiones 42
10. Glosario 44
11. Bibliografía 45

Lista de figuras

Figura 1. Calendario del Plan de Trabajo 6
Figura 2. Diagrama de Gantt 7
Figura 3. Ejemplo de archivo en formato VCF. 14
Figura 4. Diseño de la herramienta web 20
Figura 5. Panel de opciones de la aplicación web 21
Figura 6. Tabla de variantes 22
Figura 7. Tabla de cobertura 22
Figura 8. Esquema de las bases de datos empleadas en la herramienta web. 25
Figura 9. Desplegable *VCFiles.* 26
Figura 10. Esquema de priorización de variantes 30
Figura 11. Algoritmo de priorización de variantes 36
Figura 12. Representación del tiempo de ejecución (en segundos) frente al número de variantes del archivo VCF importado en la aplicación. 41

Lista de tablas

Tabla 1. Campos del archivo VCF. 14
Tabla 2. Requisitos de filtrado de calidad 29
Tabla 3. Criterios de clasificación del ACMG 34
Tabla 4. Resultados caso 1 38
Tabla 5. Tabla de variantes *probablemente patogénicas* o *patogénicas* detectadas en el caso a estudio 1. 38
Tabla 6. Resultados caso 2 39
Tabla 7. Tabla de variantes *probablemente patogénicas* o *patogénicas* detectadas en el caso a estudio 2. 39
Tabla 8. Resultados caso 3 39
Tabla 9. Tabla de variante *probablemente patogénica* detectada en el caso a estudio 3. 40
Tabla 10. Resultados caso 4 40
Tabla 11. Tabla de variante *patogénica* detectada en el caso a estudio 4. 40

1. Introducción

1.1. Contexto y justificación del Trabajo

1.1.1 Contexto

Comprender la información contenida en el ADN es uno de los objetivos principales de la ciencia, por ello han sido desarrollados numerosos métodos de secuenciación del ADN a lo largo de la historia. A partir de 2005, surgieron las técnicas de secuenciación de segunda generación (*Next-Generation Sequencing, NGS),* que permitieron un gran avance en la genética médica, entre otras áreas de la ciencia. Gracias al desarrollo de estas técnicas de secuenciación masiva es posible el análisis de muchas secuencias de ADN en paralelo, produciendo datos genómicos a gran escala. El principal desafío que presenta esta gran cantidad de datos genómicos es su manipulación, pues son necesarios recursos computacionales eficientes que permitan almacenar, gestionar e interpretar los datos de manera correcta y eficaz.

La secuenciación simultánea de ciertos genes (*Targeted Sequencing Panels, TSP*), la secuenciación completa del exoma (*Whole Exome Sequencing, WES*) y del genoma (*Whole Genome Sequencing, WGS*) mediante NGS son tres de las técnicas más empleadas actualmente en el diagnóstico clínico y la investigación biomédica, pues permiten, en ocasiones, determinar la causa de una patología concreta mediante la detección de las variaciones genéticas en múltiples genes simultáneamente (1,2). El número de variaciones genéticas que se detectan en la secuenciación de un exoma es alrededor de 150.000, y de un genoma superior a 4 millones (3), dando lugar a nuevos retos y oportunidades en el diagnóstico clínico. Existen distintos tipos de variaciones genéticas entre las cuales se destacan, variaciones estructurales (SV's, CNV's), pequeñas inserciones o deleciones (INDEL) y sustituciones (SNV's). Muchas de estas variaciones es posible que no tengan relevancia sobre el fenotipo del paciente ya sea porque no tienen efectos a nivel proteico o a nivel

sistémico. Sin embargo, otras pueden ser causales de una enfermedad mendeliana como la fibrosis quística o de patologías más complejas y/o heterogéneas como el cáncer. Conocer la clasificación y priorizar aquellas variantes que puedan ser causales de la enfermedad a estudio es un paso esencial en el proceso de diagnóstico.

La *American College of Medical Genetics and Genomics* (*ACMG*) es una organización compuesta por genetistas, bioquímicos, clínicos, citogenetistas, médicos moleculares, asesores genéticos y otros profesionales de la salud comprometidos con la práctica de la genética médica. Describen las bases de la clasificación de variantes genéticas utilizando el criterio de expertos y de datos empíricos. Las variantes se clasifican en función de su efecto en: *patogénicas*, *probablemente patogénicas*, *de significado clínico incierto, probablemente benignas* y *benignas* (4). Entre los criterios descritos podrían considerarse de suma importancia el tipo de variante en cuestión, su frecuencia poblacional y el origen de dicha variante (*de novo* o heredada). En consecuencia, se requiere conocer el contexto de cada variante genética para seguir adecuadamente las guías de clasificación *ACMG*.

Respecto a las herramientas bioinformáticas desarrolladas para el análisis y procesamiento de los datos obtenidos en la secuenciación, han evolucionado significativamente en los últimos años. *GATK* es una de las herramientas de análisis de datos genómicos más usadas hoy en día, incluye flujos de trabajo optimizados que permiten obtener resultados más precisos con la mayor eficiencia computacional (5). Uno de los procesos más importantes, el *variant calling*, detecta las variantes genéticas respecto a una referencia y las almacena en un archivo con formato VCF (*Variant Call Format*) (6,7). Este archivo, además de contener toda la información relativa a la posición cromosómica y el cambio detectado, almacena otros parámetros como la calidad de la secuenciación, profundidad de lecturas y otros datos que pueden resultar útiles para detectar datos de mala calidad. Asimismo, este archivo puede ampliarse para añadir información mediante el empleo de bases de datos en un proceso denominado anotación, que permitirá posteriormente filtrar y priorizar las variantes.

Por otro lado, herramientas como *SnpEff* (8), *ANNOVAR* (9) o *VEP* (10) son herramientas de libre acceso (*open source,* del inglés) que permiten anotar las variantes genéticas empleando bases de datos genómicas, aportando así información biológica, clínica, poblacional y predictiva asociada a cada una de ellas; como la frecuencia poblacional (*gnomAD* (11)), clasificación y anotación por otros laboratorios (*ClinVar* (12)*, dbSNP* (13)), consecuencias a nivel de gen o proteína, así como el fenotipo asociado y sus modos de herencia (*HPO* (14), *OMIM* (15)).

Una vez procesados los datos, se debe realizar un filtrado y priorización de variantes en base a la información clínica disponible del caso a estudio. Este proceso se lleva a cabo en el ámbito de un laboratorio de diagnóstico genético, por analistas cualificados que pueden no tener nociones de bioinformática ni de tratamiento de datos. Por ello se hace indispensable una interfaz gráfica que permita visualizar e integrar toda la información sobre las variantes genéticas y facilite la priorización de variantes.

Durante este Trabajo de Fin de Máster se desarrollará una aplicación que permita visualizar e integrar toda la información sobre las variantes genéticas, así como el desarrollo de algoritmos que permitan: (1) filtrar aquellas variantes que no sean relevantes o que no cumplan los requisitos mínimos de calidad establecidos; priorizar (2) variantes en función de un fenotipo o enfermedad y (3) según los estándares internacionales (*ACMG*).

1.1.2 Justificación

En la secuenciación completa de un exoma o un genoma se detectan gran cantidad de variantes genéticas que podrían ser *patogénicas*. Determinar cuál o cuáles de ellas son causales de una patología concreta es una tarea compleja y muy costosa. Una aplicación adecuada que permita visualizar e integrar toda la información sobre las variantes genéticas, e incluya algoritmos de priorización puede facilitar en gran medida el trabajo en el ámbito clínico y mejorar las tasas diagnósticas de los estudios genéticos en enfermedades congénitas. Por ello, el desarrollo de una herramienta de este tipo es interesante y motivadora. Además, es interesante profundizar en el

conocimiento de lenguajes y tecnologías para el tratamiento y visualización de datos como *Shiny*, y el uso de herramientas colaborativas de código abierto como GitHub, en la que otros programadores puedan contribuir con posibles mejoras y/o cambios.

1.2. Objetivos del Trabajo

A continuación, se concretan los objetivos generales del proyecto y los objetivos específicos para lograr cada uno de ellos:

Objetivo general:

1. Diseño de una herramienta web capaz de priorizar variantes genéticas dado un caso a estudio concreto.

Objetivos específicos:

a. Comprender y dominar el proceso de *variant calling*, el formato de los archivos VCF y el proceso de anotación.
b. Fijar criterios mínimos de calidad para el filtrado de variantes genéticas
c. Incorporar filtros y algoritmos de priorización, permitiendo orientar el análisis al fenotipo y/o enfermedad.
d. Adaptar, en la medida de lo posible, las guías *ACMG* a la priorización de variantes (tipo de variante, frecuencia poblacional, predicción *in silico*).
e. Diseñar un *front-end* sencillo e intuitivo, que muestre toda la información disponible para cada variante de manera ordenada, escalonada y a demanda del usuario.

1.3. Enfoque y método seguido

Para el desarrollo de este proyecto se necesita un archivo en formato VCF como punto de partida. Por esta razón, se detalla de forma teórica el proceso de obtención de los archivos en formato VCF y su estructura, para su posterior manipulación a través de la aplicación. Además, es imprescindible disponer de

uno o varios archivos en formato VCF que se emplearán durante el desarrollo. Estos ficheros se descargan del repositorio *Genome in a Bottle (NIST),* que contiene un archivo VCF para el caso NA12878.

El archivo VCF es el punto de partida a partir del cual se desarrolla la herramienta. No obstante, debe estar anotado antes de su utilización, es decir, debe contener toda la información necesaria acerca de cada variante genética. Esto se realiza mediante el programa *SnpEff* utilizando las bases de datos *ClinVar, dbSNP, dbNSFP, gnomAD* y diversos predictores. Entre los múltiples programas existentes para la anotación de variantes, *SnpEff* se considera el más adecuado por permitir un proceso de anotación completamente personalizable, no consumir muchos recursos y emplear una sintaxis de línea de comando sencilla.

Finalmente, se emplea el paquete *Shiny* de R para implementar la aplicación web con una interfaz de usuario interactiva. Se trata de un paquete de R diseñado específicamente para la visualización interactiva de grandes cantidades de datos, y se puede personalizar de forma sencilla con sintaxis HTML. Para conseguir un almacenamiento eficiente de la información contenida en los archivos VCF, se diseña una base de datos relacional y estructurada, a la que poder acceder desde la aplicación de una manera rápida y eficaz. Además, se incluyen los algoritmos de filtrado y priorización previamente programados.

1.4. Planificación del Trabajo

1.4.1. Tareas

- Documentación y estado del arte
- Instalación del entorno de programación, paquetes y/o módulos necesarios. Búsqueda y descarga de VCFs con su posterior anotación.
- Diseñar la estructura del programa, la apariencia, interfaz y la estructura de las bases de datos relacionales.

- Programar la importación de variantes en formato VCF, su almacenamiento en bases de datos y posterior visualización.
- Diseñar e implementar filtros automáticos que permitan la clasificación automática de variantes en una primera fase.
- Diseñar e implementar filtros personalizables por el usuario.
- Implementar la posibilidad de clasificar las variantes individuales, o aquellas enmarcadas dentro de un filtrado, empleando la clasificación en 5 niveles del *ACMG*.
- Programar la exportación de los resultados del análisis a un archivo Excel/LibreOffice.
- Comprobar el correcto funcionamiento de la aplicación.
- Redacción y finalización de la memoria

1.4.2. Calendario

Task name	Start date	Finish date	Duration
PEC0. Propuesta de TFM	16/09/2020	28/09/2020	9
PEC1. Plan de trabajo	29/09/2020	13/10/2020	11
Documentación y estado del arte	16/09/2020	13/10/2020	20
Definición de las tareas a realizar y su duración	01/10/2020	09/10/2020	7
PEC2. Desarrollo del proyecto - Fase 1	14/10/2020	16/11/2020	24
Preparación del entorno. Descarga de VCFs	14/10/2020	19/10/2020	4
Definición de parámetros y criterios de filtrado y priorización	20/10/2020	22/10/2020	3
Diseño de la estructura de la aplicación	22/10/2020	30/10/2020	7
Programación del algoritmo de priorización	02/11/2020	16/11/2020	11
PEC3. Desarrollo del proyecto - Fase 2	17/11/2020	14/12/2020	20
Implementación del algoritmo	17/11/2020	30/11/2020	10
Implementación de filtros automáticos y personalizables	17/11/2020	30/11/2020	10
Programación e implementación de la exportación	02/12/2020	02/12/2020	1
Comprobación del correcto funcionamiento	03/12/2020	14/12/2020	8
PEC4. Cierre memoria	15/12/2020	05/01/2021	16
PEC5. Presentación del proyecto	06/01/2021	20/01/2021	11
Elaboración de la presentación	06/01/2021	11/01/2021	4
Defensa pública	13/01/2021	20/01/2021	6

Figura 1. Calendario del Plan de Trabajo

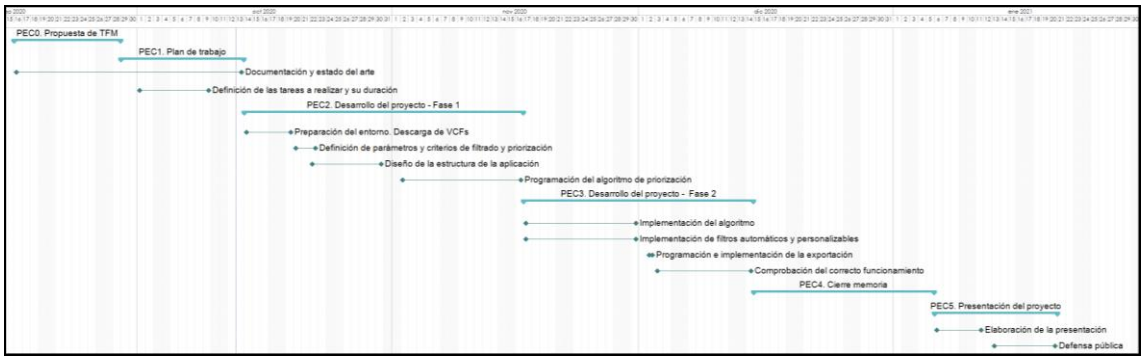


Figura 2. Diagrama de Gantt. Representación esquemática del Plan de Trabajo establecido

1.4.3. Hitos

- Inicio PEC0. Propuesta de TFM. 16/09/2020
  - Documentación y estado del arte
  - Redacción de la propuesta.
- Fin PEC0. Propuesta de TFM. 28/09/2020
- Inicio PEC1. Plan de trabajo. 29/09/2020
  - Definición de las tareas a realizar y su duración.
  - Fin de etapa de documentación y estado del arte.
- Fin PEC1. Plan de trabajo. 13/10/2020
- Inicio PEC2. Desarrollo del proyecto – Fase 1. 14/10/2020
  - Preparación del entorno de trabajo. Descarga y anotación de VCFs. 19/10/2020
  - Definición de parámetros y criterios de filtrado y priorización. 22/10/2020
  - Diseño de la estructura de la aplicación: apariencia, interfaz y bases de datos relacionales. 01/11/2020
  - Programación del algoritmo de priorización de variantes
- Fin PEC2. Desarrollo del proyecto – Fase 1. 16/11/2020
- Inicio PEC3. Desarrollo el proyecto – Fase 2. 17/11/2020
  - Implementación del algoritmo.
  - Implementación de filtros automáticos y personalizables. 01/12/2020
  - Programación e implementación de la exportación. 02/12/2020
  - Comprobación del correcto funcionamiento de la herramienta web.

- Fin PEC3. Desarrollo el proyecto – Fase 2. 14/12/2020
- PEC4. Cierre de la memoria. 15/12/2020 – 05/01/2021
- PEC5a. Elaboración de la presentación. 06/01/2021 – 10/01/2021
- PEC5b. Defensa pública. 13/01/2021 – 20/01/2021

1.4.4. Análisis de riesgos

Los principales riesgos identificados que se pueden presentar a lo largo del proyecto son:

- Planificación inicial
    - Descripción: Planificación de calendario no realista
    - Impacto: No cumplimiento de plazos
    - Probabilidad: Media
    - Acción mitigadora: Asignar suficiente margen temporal para compensar imprevistos o una mala planificación de ciertas tareas.
- Alcance del proyecto
    - Descripción: Mala valoración inicial del alcance y complejidad del proyecto.
    - Impacto: No cumplimiento de plazos y/o objetivos iniciales.
    - Probabilidad: Baja
    - Acción mitigadora: Consensuar de forma detallada y realista con el director del TFM los plazos y objetivos del proyecto.
- Valoración de herramientas
    - Descripción: Selección de herramientas que posteriormente sean inadecuadas
    - Impacto: No cumplimiento de plazos y/o objetivos iniciales.
    - Probabilidad: Baja
    - Acción mitigadora: Asignar suficiente margen temporal para buscar alternativas si fuera necesario.
- Pérdida de datos accidental y problemas del equipo informático

- Descripción: Pérdida de datos del desarrollo, problemas con el equipo informático donde se programa y se elabora este TFM.
- Impacto: No cumplimiento de plazos y/o objetivos iniciales.
- Probabilidad: Baja
- Acción mitigadora: Empleo de repositorios de código en la nube, equipos de sustitución y copias de seguridad.

1.5. Resultados esperados

- Plan de trabajo
- Memoria del proyecto
- Producto: Herramienta web en lenguaje R (app.R)
- Repositorio en GitHub con la herramienta web y una guía de usuario
- Presentación virtual del proyecto
- Autoevaluación del proyecto

1.6. Breve descripción de los otros capítulos de la memoria

1. **Introducción**.
2. **Datos de partida:** Explicación del proceso de obtención de los datos necesarios, su estructura y su anotación.
3. **Diseño de la aplicación web:** Almacenamiento y estructura de datos. Diseño del *front-end* para la visualización.
4. **Importación y visualización de los datos:** Programación de la importación de las variantes a la base de datos y su posterior visualización.
5. **Filtrado y priorización de variantes:** Implementación de algoritmos de filtrado y priorización.
6. **Pruebas de funcionalidad:** Testeo de la funcionalidad de la herramienta web, depuración de posibles errores no detectados y casos de prueba.
7. **Resultados:** Exposición de los resultados obtenidos mediante la aplicación web.
8. **Pruebas de rendimiento:** Testeo del rendimiento de la aplicación.
9. **Conclusiones:** Conclusión del proyecto y líneas futuras.

2. Aspectos generales y estado del arte

La secuenciación de segunda generación (*Next-Generation Sequencing, NGS)* (16)*,* un método para secuenciar simultáneamente millones de fragmentos de ADN, se ha adoptado rápidamente en el laboratorio clínico debido a que permite analizar múltiples genes y muestras a la vez en una única prueba. Existen gran variedad de técnicas de secuenciación, así como formatos de archivos que se obtienen tras la secuenciación. En concreto, este Trabajo de Fin de Máster se centra en la secuenciación completa de exoma (*WES*), donde se obtiene entre un 25% y un 40% de rendimiento diagnóstico (17). Esta técnica se basa en la secuenciación de las regiones codificantes de los genes, obteniendo así las secuencias de los exones de cada gen. Se estima que alrededor del 85% de las mutaciones patogénicas o causantes de patología se localizan en regiones codificantes del genoma, por ello esta técnica tiene un gran potencial en el diagnóstico clínico de trastornos genéticos raros (18).

Tras la secuenciación, el archivo de salida más utilizado por los secuenciadores en el análisis de secuencias es un archivo en formato FASTQ, un archivo de texto que contiene el identificador de la secuencia, la secuencia y la calidad de esta. En caso de que se hayan secuenciado simultáneamente varias muestras, se lleva a cabo un proceso llamado *demultiplexing,* que consiste en separar las diferentes muestras según una secuencia de identificación añadida a las lecturas. Una vez obtenidas las secuencias, el siguiente paso es alinearlas a un genoma de referencia, generando un archivo en formato BAM donde se posicionan las lecturas (o secuencias) en base a las coordenadas genómicas de la referencia. Finalmente, el objetivo de la secuenciación es generar un archivo VCF que incluya todas las variantes genéticas detectadas (SNV’s, INDEL’s, CNV’s, SV’s).

Mientras que los pipelines bioinformáticos para la obtención de variantes son relativamente uniformes y producen los mismos archivos, los métodos para la priorización e interpretación de variantes en estudios clínicos son muy diversos.

En general, los estudios de NGS generan un gran número de variantes, muchas de las cuales, en el contexto de un estudio clínico, no son relevantes ya que es muy poco probable que tengan un efecto funcional a nivel proteico y/o sistémico, y muy pocas de ellas tendrán relevancia en el estudio de un fenotipo/enfermedad en concreto. Es por ello que, para detectar las variantes causantes de enfermedad entre una enorme cantidad de posibilidades, se requieren pasos posteriores de priorización. Por último, dada una lista priorizada de variantes, los analistas inspeccionan manualmente cada una y seleccionan un subconjunto para ser informadas. Los procesos de priorización están basados en una amplia gama de fuentes de anotación, principalmente en (2):

a) Anotación funcional de la variante. Realizado por programas y bases de datos de secuencias genómica; predice el efecto funcional provocado por la variante sobre el gen.
b) Tipo de variante. Las variantes *missense*, *nonsense*, *stop-loss*, *frameshift* y *splicing* tienen potencial para afectar la función de las proteínas.
c) Bases de datos poblacionales. Variantes muy frecuentes tienen poca probabilidad de ser causantes de enfermedad.
d) Predictores computacionales. Basados en conservación filogenética, propiedades fisicoquímicas y alteración de elementos de secuencia entre otros, especialmente útiles para las variantes *missense*, inserciones y delecciones *in-frame* y variantes en regiones no codificantes.
e) Bases de datos de asociaciones de genes, fenotipos y modos de herencia. La información clínica e historia familiar del caso a estudio es esencial.

Diversos programas de análisis y priorización de variantes han sido desarrollados en los últimos años, empleando diferentes técnicas de priorización con menor o mayor grado de facilidad de uso. Todos ellos, o la gran mayoría, emplean términos HPO para incorporar el fenotipo del paciente. De esta manera se aprovecha el genotipo del paciente para establecer relación con el fenotipo, priorizando aquellos genes que probablemente tengan más

relevancia. De todas las herramientas disponibles para la priorización de variantes se destacan:

- **Exomiser**. Mediante el algoritmo PHIVE (*PHenotypic Interpretation of Variants in Exomes*) el programa prioriza las variantes en función de una puntuación de variante adquirida según su rareza y su predicción *in silico,* en combinación con una puntuación de fenotipo que se obtiene mediante el cálculo de la similitud entre el fenotipo HPO (*Human Phenotype Ontology),* definido previamente en el paciente, y el fenotipo MPO (*Mammalian Phenotype Ontology*), de organismos modelos modificados genéticamente, como son los ratones (19).
- **PhenIX**. Es similar a *Exomiser*, pues filtra las variantes en función de su rareza y predicción de patogenicidad. Sin embargo, no realiza una puntuación de fenotipo como *Exomiser*, sino que evalúa el espectro fenotípico asociado a los genes resultantes buscando similitudes con las anomalías fenotípicas descritas en el paciente mediante los términos HPO (20).
- **Phenolyzer.** Prioriza las variantes en función de la enfermedad o el fenotipo proporcionado por el usuario como texto libre, sin necesidad de estar codificado mediante un término HPO. *Phenolyzer* utiliza información sobre la relación entre diversos genes, la interacción entre las proteínas, rutas de biológicas en común así como información sobre regulación transcripcional (21).
- **OMIM Explorer**. Emplea la similitud semántica para puntuar los genes más relevantes, del mismo modo que el resto de los programas. Su característica más relevante es que utiliza una interfaz interactiva práctica que produce resultados íntimamente relacionados con los obtenidos en una revisión de expertos (22).

3. Datos de partida

En este Trabajo de Fin de Máster se ha escogido el formato VCF como punto de partida. El repositorio *Genome in a Bottle (NIST)* es un consorcio público-privado-académico organizado por NIST que aporta una infraestructura técnica autorizada (estándares de referencia, métodos de referencia y datos de referencia) para su uso en la validación analítica y el desarrollo, optimización y demostración de nuevas tecnologías (23). *GIAB* ha caracterizado actualmente un genoma piloto (NA12878/HG001) del proyecto HapMap, y dos tríos hijo/padre/madre de ascendencia judía asquenazí (GM24385/HG002-GM24149/HG003-GM24143/HG004) y china (GM24631/HG005-GM24694-GM24695) del *Personal Genome Project*. Se ha seleccionado el genoma NA12878 para el desarrollo del proyecto debido a que está ampliamente estudiado.

3.1. Formato VCF

El formato VCF (6) es el formato estándar para almacenar las variaciones genéticas detectadas en la secuenciación de numerosas muestras junto con sus respectivas anotaciones. Existen otros formatos para almacenar datos genómicos, como el formato GFF, del inglés *Genetic Feature Format* (24), pero no está estandarizado para almacenar información sobre variantes ni múltiples muestras. Por esta razón, el formato VCF se convierte en uno de los formatos de referencia para secuenciación de segunda generación. La información contenida en un archivo VCF se presenta estructurada por una sección para la metainformación (*header*) y otra para los datos obtenidos de la secuenciación. El encabezado (*header*), delimitado por el carácter doble "##", contiene información sobre el conjunto de datos y las fuentes de referencia empleadas, así como la descripción de todas las etiquetas y anotaciones utilizadas para caracterizar y cuantificar las propiedades de cada variante. Tras las líneas del encabezado, existe una línea adicional, delimitada por el símbolo "#", que define los campos de la sección datos (ver Tabla 1). Las primeras ocho columnas de los registros representan propiedades referentes a la variante

detectada. La información específica de la muestra, como el genotipo, se encuentra en la columna FORMAT.

Campo	Definición
CHROM	Cromosoma
POS	Posición genómica de inicio
ID	Código de identificación de la variante en dbSNP
REF	Alelo de referencia
ALT	Listado de alelos alternativos separados por coma
QUAL	Calidad de la secuenciación
FILTER	Información de filtrado (PASS/FILTER)
INFO	Información adicional (anotación)
FORMAT	Información sobre el genotipo

Tabla 1. Campos del archivo VCF.

De todos los campos del archivo VCF, el campo INFO es el designado para almacenar información adicional como es la anotación de las variantes. La anotación de variantes implica incluir metadatos adicionales que aporten nueva información sobre las variantes y permitan mejorar la evaluación de estas. Si bien la información sobre las frecuencias poblacionales, la asociación con determinadas enfermedades o fenotipos y los efectos deletéreos predichos se agregan comúnmente, las anotaciones más fundamentales como los genes, transcritos o regiones (intrón, exón, sitios de *splicing,* etc) a los que afectan las variantes también ayudarán a evaluar y clasificar dichas variantes.

Figura 3. Ejemplo de archivo en formato VCF. Imagen reproducida de Danecek P. *et al.* (2011)**.** Se presenta el encabezado (*Header*) con la información delimitada por "##" y "#", así como la sección de datos (*Body*) donde se detalla cada registro en una línea diferente. En las columnas posteriores a FORMAT se describe información específica de cada muestra.

3.2. Anotación de variantes

Entre los múltiples programas existentes para la anotación de variantes, se ha escogido *SnpEff* porque permite una anotación flexible y personalizable, no consume muchos recursos y emplea una sintaxis de línea de comando sencilla. En cuanto a las bases de datos empleadas, se han escogido *ClinVar, dbSNP, dbNSFP/dbscSNV* y gnomAD. Por un lado, *ClinVar* es una base de datos que proporciona una asociación entre las variantes genéticas y determinados fenotipos o enfermedades. Esta base de datos aporta información clínica sobre la variante, si se ha descrito anteriormente, clasificación por otros laboratorios, etc. Por otro lado, *dbSNP* (*Single Nucleotide Polymorphism Database*) contiene un gran número de variantes genéticas identificadas a las que asigna un código de identificación ("rs" seguido de un número). En cuanto *dbNSFP/dbscSNV,* es una base de datos desarrollada para la predicción funcional de las variantes, contiene puntuaciones de predicción de 37 algoritmos de predicción. Por último, *gnomAD* es una base de datos que aporta información sobre frecuencias poblacionales.

Los comandos empleados para anotar el archivo VCF han sido:

- Anotación con S*npEff*:

```
java.exe -jar snpEff.jar ann -v -canon -no PROTEIN_PROTEIN_INTERACTION_LOCUS -no PROTEIN_STRUCTURAL_INTERACTION_LOCUS -no-downstream -no-intergenic -no-upstream -noStats -ss 10 -noMotif -noNextProt hg38_RefSeqCurated archivo.vcf > archivo_snpeff.vcf
```

- Anotación con *ClinVar*:

```
java.exe -jar SnpSift.jar annotate -v -noLog -tabix -noId clinvar.vcf.gz archivo_snpeff.vcf > archivo_snpeff_clinvar.vcf
```

- Anotación con *gnomAD*:

```
java.exe -jar SnpSift.jar annotate -v -noLog -tabix -noId gnomAD.vcf.gz archivo_snpeff_clinvar.vcf > archivo_snpeff_clinvar_gnomad.vcf
```

- Anotación con *dbNSFP*:

```
java.exe -jar SnpSift.jar dbnsfp -v -db dbNSFP4.3a_predictors.txt.gz -f SIFT_converted_rankscore,Polyphen2_HVAR_rankscore,MutationTaster_converted_rankscore, CADD_phred,fathmm-MKL_coding_score,GERP++_RS,phyloP30way_mammalian,
```

```
phastCons30way_mammalian archivo_snpeff_clinvar_gnomad.vcf > archivo_snpeff_clinvar_gnomad_dbnsfp.vcf
```

- Anotación con *dbSNP*:

```
java.exe -jar SnpSift.jar annotate -v -noLog -tabix -id dbsnp151.vcf.gz archivo_snpeff_clinvar_gnomad_dbnsfp.vcf.vcf> archivo_dbsnp_annotate.vcf
```

3.3. Cálculo de cobertura

Opcionalmente, la web permite la subida de un archivo que refleje las coberturas obtenidas de cada gen, con el fin de poder identificar directamente en la aplicación las regiones secuenciadas con profundidad inferior a 10X, por si fuera necesaria su secuenciación por otras técnicas. Para realizar el cálculo de la cobertura es necesario el archivo de alineamiento y un archivo de intervalos con las regiones codificantes del genoma, obtenido de *RefSeq*. Se ha empleado el programa *CollectHsMetrics* de la suite *Picard Tools* (25). Posteriormente, el resultado es procesado para mostrar las regiones con coberturas inferiores a 10X mediante código R.

4. Diseño de la aplicación web

La herramienta web desarrollada en este Trabajo de Fin de Máster se denomina PGVWeb, siglas de *Priorization of Genetic Variants Web.* Esta aplicación ha sido implementada en lenguaje R mediante el programa RStudio. Los paquetes empleados para su desarrollo han sido numerosos, entre ellos destaca el paquete *Shiny* (26) para la implementación de la aplicación web, el paquete *vcfR* (27) para la lectura y manipulación de los datos del archivo VCF, el paquete *RSQLite* (28) para la integración de la información en el sistema de gestión de bases de datos SQLite, y el paquete *DT* (29) que permite mostrar los datos en una tabla customizada.

El lenguaje R es un lenguaje de programación con un enfoque estadístico que proporciona una gran cantidad de herramientas estadísticas y gráficas muy útiles en el campo de la Bioinformática. Además, es un lenguaje que puede integrarse con distintas bases de datos, como es SQLite. Por otro lado, RStudio es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, que facilitará tanto la tarea de uso interactivo de R como la programación de scripts en dicho lenguaje. De los diferentes proyectos de RStudio, el que presenta mayor potencial es *Shiny*.

4.1. Paquetes de R más relevantes

Paquete *Shiny*:

Shiny es un *framework* que hace más sencillo el desarrollo de aplicaciones web interactivas basadas en R. De todas las propiedades de *Shiny* destacan dos: la interactividad y la reactividad, ambas estrechamente relacionadas. El primero de los términos permite manipular los datos de la aplicación sin necesidad de modificar el código. Esto se hace posible gracias a la reactividad, en la que cada modificación por parte del usuario renueva todo el proceso.

Gracias al uso de las funcionalidades de *Shiny*, se puede crear directamente en R no sólo el *back-end* de la aplicación web, sino también la interfaz de usuario, sin necesidad de conocimientos de HTML o CSS.

Una aplicación *Shiny* se almacena en un script denominado `app.R` que consta de tres partes:

- Objeto para la interfaz de usuario (`ui`). En él se define la estructura y el aspecto de la aplicación.
- Función del servidor (`server`). Contiene las instrucciones para generar la aplicación.
- Llamada a la aplicación mediante la función `ShinyApp()`

Paquete RSQLite:

SQLite es un sistema de gestión de bases de datos relacionales de dominio público. Lee y almacena los datos directamente en archivos de disco ordinario con extensión ".sqlite". Esta característica lo diferencia de otros sistemas que tienen una estructura cliente-servidor (MySQL, PostgreSQL, etc.), es decir, en los que se necesita un servidor para acceder a la base de datos. En este aspecto, SQLite se enlaza directamente con el programa pasando a ser parte integral del mismo.

El paquete *RSQLite* incorpora el sistema de gestión SQLite en R, lo que proporciona una interfaz compatible con DBI. Este paquete permite realizar las consultas de diferentes formas. La forma escogida ha sido mediante DBI, porque es una implementación nativa de R y emplea una sintaxis sencilla.

Paquete vcfR:

El desarrollo de los archivos en formato VCF genera la necesidad de crear herramientas para trabajar con ellos. Si bien existe un número creciente de software para leer datos de VCF, muchos solo extraen los genotipos sin incluir los datos asociados como la calidad de estos. El paquete *vcfR* ofrece un conjunto de herramientas en lenguaje R diseñadas para leer, escribir,

manipular y analizar datos de un archivo en formato VCF. También incluye funciones para extraer parte de los datos y trazar estadísticas de estos.

La función `read.vcfR()` se emplea para leer el archivo VCF y almacenar la información como un objeto vcfR. Otra de las funciones características es `extract.gt()`, que permite extraer información sobre el genotipo de las variantes (frecuencia alélica, `AF`, o cigosidad, `GT`). Por último, la función `vcfR2tidy()` se emplea para convertir todos los datos del objeto vcfR en un *tibble*, lo que se conoce en R como *data.frame* pero en su versión más actualizada. Además, permite especificar qué partes de los datos se quieren convertir.

Paquete DT:

La forma predeterminada de R para mostrar matrices y/o *data.frames* no es una forma eficaz de presentar el contenido. Por ello, RStudio ha desarrollado este paquete de R que proporciona una interfaz para la biblioteca *JavaScript DataTables.* DT permite a los usuarios tener tablas interactivas en formato HTML que incluyen búsqueda, clasificación, filtrado y exportación de los datos. La función `datatable()` contiene numerosas opciones para definir las propiedades de la tabla (estilo, tamaño, ancho de columna y fila, etc.)

4.2. Front-end

La idea principal consiste en ofrecer una interfaz de usuario interactiva que muestre en una primera instancia todas aquellas variantes que puedan ser candidatas para diagnosticar un determinado fenotipo y/o enfermedad. Además, también se pretende que el usuario pueda seleccionar diferentes filtros para mostrar uno u otro tipo de variantes, dependiendo de su herencia, frecuencia, clasificación, etc. En cuanto a la interfaz gráfica principal de la aplicación, está formada por un encabezado, un panel de opciones y el panel principal.

En primer lugar, el encabezado contiene el nombre de la aplicación (PGVWeb) así como un botón de *logout* para cerrar sesión en la aplicación. En segundo lugar, el panel de opciones incluye un apartado para subir el archivo VCF del

paciente, otro para el archivo de cobertura, así como un desplegable con los diferentes filtros y opciones de priorización que el usuario podrá seleccionar en función de sus requerimientos. Por último, en el panel principal se muestra la tabla de variantes detectadas en el paciente, así como información adicional sobre cobertura de genes (ver Figura 4).

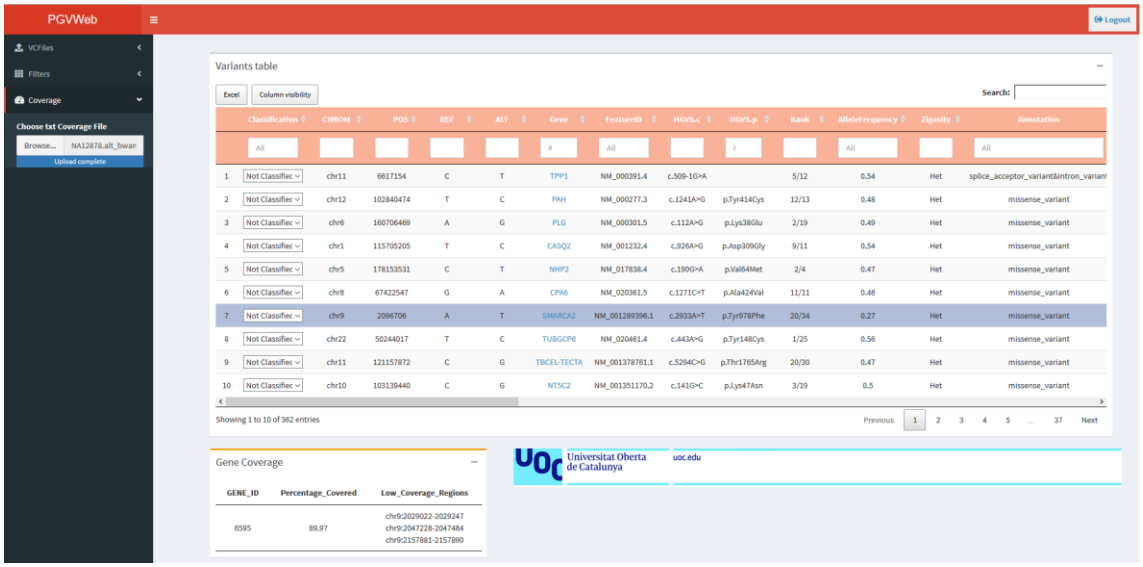


Figura 4. Diseño de la herramienta web. Se muestran las diferentes partes de la aplicación. En la parte izquierda se encuentra el panel de opciones, en la parte central la tabla de variantes y abajo a la izquierda la tabla de cobertura del gen *SMARCA2*.

4.2.1. Panel de opciones

Las tres partes que componen el panel de opciones tienen, a su vez, una serie de subapartados con los que el usuario puede interaccionar (ver Figura 5A). Para analizar los datos de un archivo VCF existen dos formas: mediante la subida del archivo directamente, con su correspondiente nombre del caso, o bien buscando el archivo concreto en la base de datos del usuario, siempre y cuando éste se haya subido previamente. Una vez que la información del VCF está disponible en la aplicación, se pueden aplicar los filtros para el filtrado y priorización de variantes. Los parámetros seleccionados por defecto para cada uno de ellos se muestran en la Figura 5B. Además, se puede incluir un archivo de cobertura para completar el análisis de variantes mediante la importación de un archivo de cobertura en formato de texto.

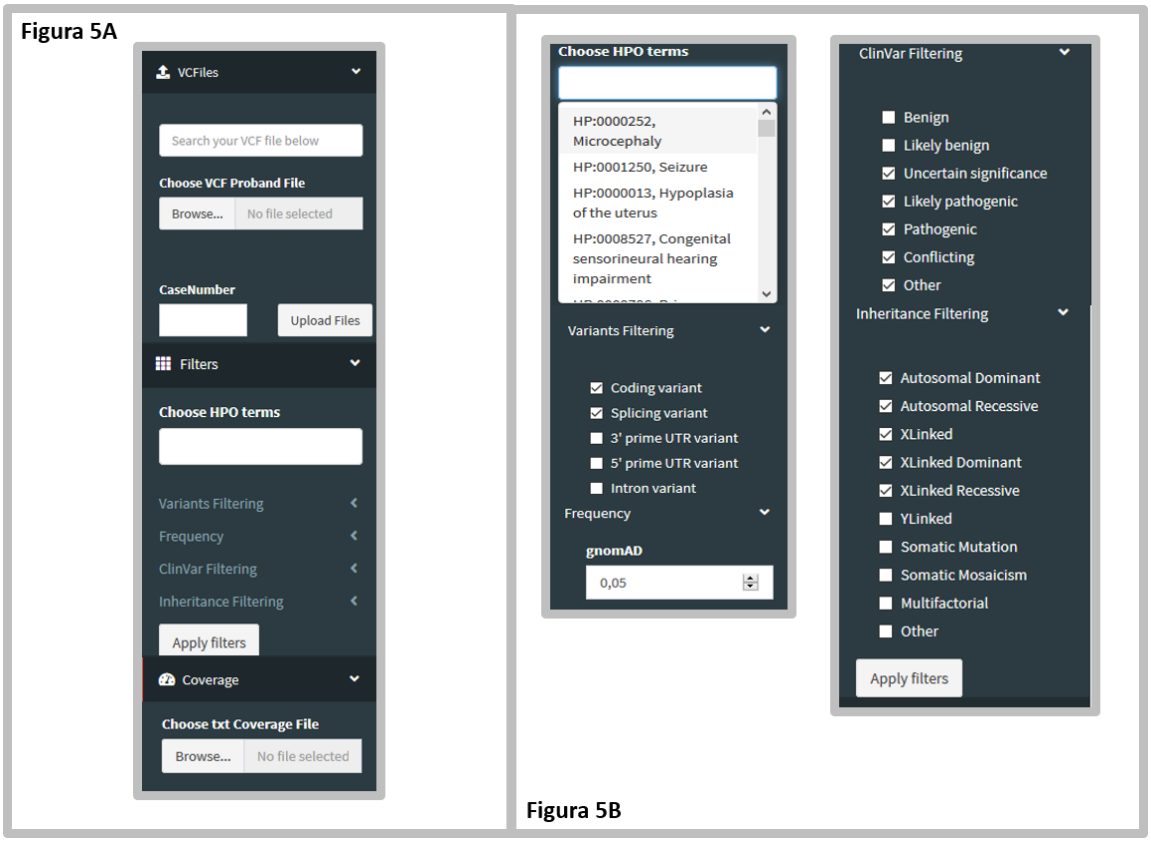


Figura 5. Panel de opciones de la aplicación web. A) Subapartados presentes en el panel de opciones. B) Desglose de los diferentes filtros y opciones de priorización del apartado “Filtros”.

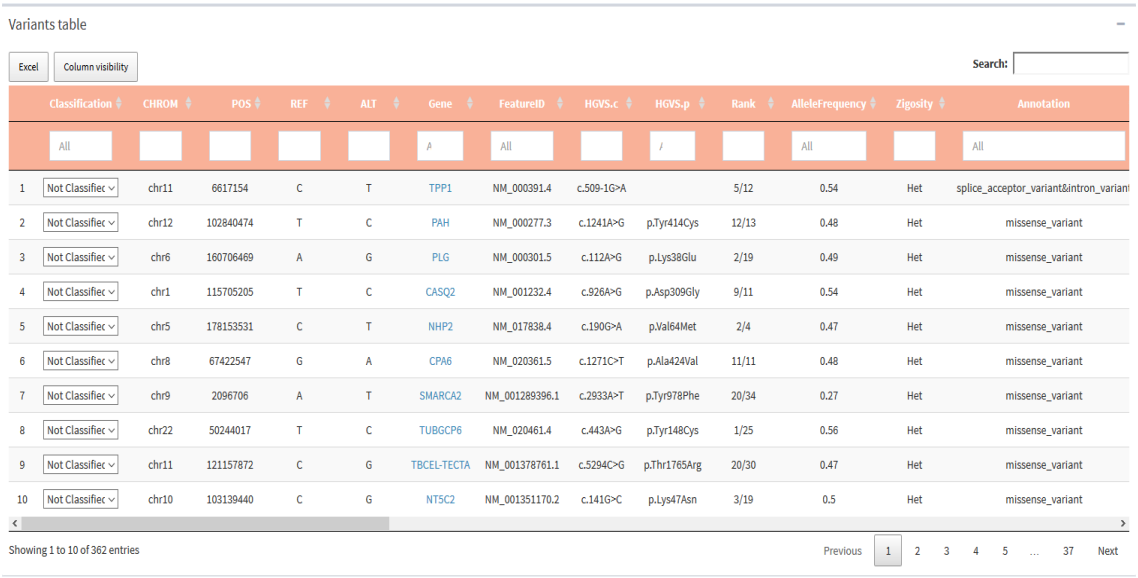
4.2.2. Panel principal

La parte más importante de esta aplicación es el panel principal, formado por la tabla de variantes detectadas y la tabla de cobertura de los genes. Como se ha descrito en apartados anteriores, una de las ventajas del paquete DT de R es que ofrece la posibilidad de generar tablas interactivas en formato HTML y permite customizar las mismas. En este proyecto la tabla de variantes presenta las siguientes propiedades (ver Figura 6):

- Botones y cuadro de búsqueda superior: Permite exportar los datos a un archivo Excel, seleccionar las columnas visibles y buscar coincidencias de texto en la tabla, como genes, enfermedades, consecuencias de las variantes, clasificaciones, etc.
- Encabezado: Permite ordenar la tabla y realizar búsquedas y filtrados basados en una columna específica.
- Tabla de variantes: Se muestra toda la información disponible para cada variante y se irá actualizando de manera dinámica en función de los filtros

y opciones de priorización seleccionados. Además, algunos campos presentan enlaces embebidos que dirigirán a la entrada de la base de datos de origen mediante navegador externo.

- Pie de tabla: Muestra información de las variantes visibles sobre el total.


Variants table

Excel | Column visibility | Search:

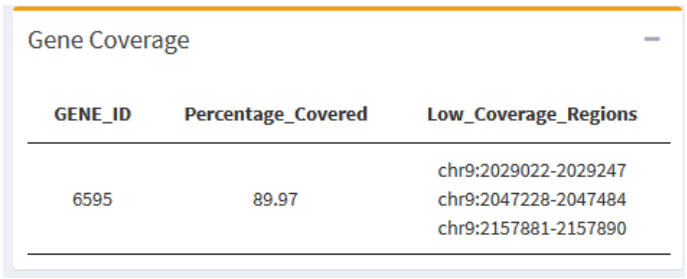
	Classification	CHROM	POS	REF	ALT	Gene	FeatureID	HGVS.c	HGVS.p	Rank	AlleleFrequency	Zigosity	Annotation
	All					A	All		/		All		All
1	Not Classifiec	chr11	6617154	C	T	TPP1	NM_000391.4	c.509-1G>A		5/12	0.54	Het	splice_acceptor_variant&intron_variant
2	Not Classifiec	chr12	102840474	T	C	PAH	NM_000277.3	c.1241A>G	p.Tyr414Cys	12/13	0.48	Het	missense_variant
3	Not Classifiec	chr6	160706469	A	G	PLG	NM_000301.5	c.112A>G	p.Lys38Glu	2/19	0.49	Het	missense_variant
4	Not Classifiec	chr1	115705205	T	C	CASQ2	NM_001232.4	c.926A>G	p.Asp309Gly	9/11	0.54	Het	missense_variant
5	Not Classifiec	chr5	178153531	C	T	NHP2	NM_017838.4	c.190G>A	p.Val64Met	2/4	0.47	Het	missense_variant
6	Not Classifiec	chr8	67422547	G	A	CPA6	NM_020361.5	c.1271C>T	p.Ala424Val	11/11	0.48	Het	missense_variant
7	Not Classifiec	chr9	2096706	A	T	SMARCA2	NM_001289396.1	c.2933A>T	p.Tyr978Phe	20/34	0.27	Het	missense_variant
8	Not Classifiec	chr22	50244017	T	C	TUBGCP6	NM_020461.4	c.443A>G	p.Tyr148Cys	1/25	0.56	Het	missense_variant
9	Not Classifiec	chr11	121157872	C	G	TBCEL-TECTA	NM_001378761.1	c.5294C>G	p.Thr1765Arg	20/30	0.47	Het	missense_variant
10	Not Classifiec	chr10	103139440	C	G	NT5C2	NM_001351170.2	c.141G>C	p.Lys47Asn	3/19	0.5	Het	missense_variant

Showing 1 to 10 of 362 entries

Previous 1 2 3 4 5 ... 37 Next

Figura 6. Tabla de variantes. Se muestran las diez primeras variantes y las trece primeras columnas de la tabla de variantes. Abreviaturas: CHROM, Cromosoma; POS, Posición Inicial de la variante; REF, Alelo de referencia; ALT, Alelo alternativo; Gene, Gen afectado; FeatureID, Transcrito afectado según RefSeq; HGVS.c, anotación a nivel codificante; HGVS.p, anotación a nivel proteico; Rank, exón/intrón afectado.

La tabla de cobertura presenta únicamente tres columnas en las que se especifica el número de acceso del gen seleccionado (*GENE ID*), el porcentaje cubierto a 10X y, de no ser un 100%, se muestran las regiones que han quedado insuficientemente cubiertas.


Gene Coverage

GENE_ID	Percentage_Covered	Low_Coverage_Regions
6595	89.97	chr9:2029022-2029247 chr9:2047228-2047484 chr9:2157881-2157890

Figura 7. Tabla de cobertura. El cuadro representa la cobertura obtenida para el gen *SMARCA2*, el cual está cubierto al 89% con una profundidad mínima de 10X y presenta 3 regiones que no quedan completamente cubiertas.

4.3. Back-end

Para el almacenamiento de usuarios junto con sus credenciales, los archivos subidos por estos, y las bases de datos *HPO* y *OMIM* necesarias para el proceso de priorización, se ha optado por SQLite, un software de administración de bases de datos relacionales que permitirá el almacenamiento de todos estos datos de una manera compacta y escalable. Entre las ventajas de este software libre se destaca que almacena toda la información en un único fichero en formato ".sqlite", lo que facilita el almacenamiento, las copias de seguridad y asegura una correcta migración de una instancia a otra de la aplicación sin generar problemas de incompatibilidad.

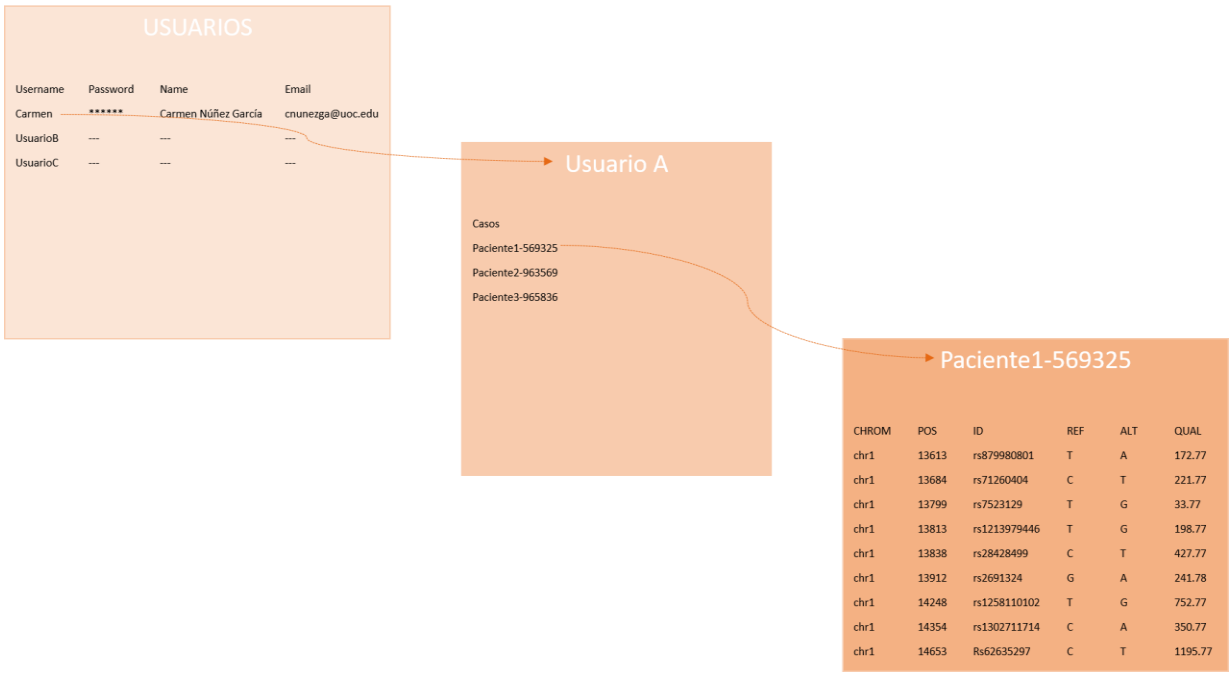
En este proyecto se han desarrollado dos bases de datos relacionales que permiten la gestión de la mayor parte del mismo.

Por un lado, se desarrolla una base de datos que almacena toda la información sobre los usuarios y los archivos importados. La estructura interna de la base de datos consta de una tabla llamada "USUARIOS", donde se almacenan los usuarios registrados y que tienen acceso a la aplicación. Además, por cada usuario existirá una tabla relacionada que abarcará todos aquellos casos que ese usuario haya subido a la aplicación. Por último, para cada archivo VCF se generará una tabla que contendrá toda la información referente a dicho archivo (Ver Figura 8). De esta manera se consigue ahorrar tiempo de procesamiento, dado que una vez que el usuario suba un archivo a la aplicación, éste quedará almacenado en la base de datos para su uso posterior.

Por otro lado, SQLite también se emplea para generar una base de datos que guarda la información contenida en las bases de datos *HPO* y *OMIM*, las cuales se explicarán detalladamente en apartados posteriores. En esta base de datos existen tres tablas, una que contiene la información sobre los términos *HPO*, otra designada para los datos de *OMIM* y una última que relaciona ambas tablas, denominada "HPO-OMIM". Las tablas se han pre-procesado para facilitar su posterior manipulación obteniendo como resultado los datos que se muestran en la Figura 8.

En una primera instancia, a la hora de completar la tabla de variantes durante la ejecución de la aplicación, se consideró la opción de realizar consultas directamente a la base de datos. Sin embargo, este modo de almacenamiento, al estar basado en disco y que opera por medio de consultas, presenta un rendimiento inferior que se pone de manifiesto en aplicaciones que requieren una rápida respuesta como la tabla de variantes. Por ello se hace necesario trasladar, durante la ejecución, las variantes a variables como *dataframes* consiguiendo una mayor velocidad de visualización y manipulación de los datos, sin un coste computacional o de memoria apreciable.

Base de datos de usuarios



Base de datos OMIM-HPO

OMIM TABLE

Chrom	Start	End	MIMumber	EntrezGeneID	Phenotypes	Inheritance
chr1	0	27600000	155600	1243	{Melanoma, cutaneous malignant, 1}, 155600 (2), Autosomal dominant	AD
chr1	0	20100000	115665	7792	Cataract 8, multiple types, 115665 (2), Autosomal dominant	AD
chr1	0	46300000	608995	406874	{Dyslexia, susceptibility to, 8}, 608995 (2), Multifactorial, Autosomal dominant	AD, Mu
chr1	1013496	1014539	147571	9636	Immunodeficiency 38, 616126 (3), Autosomal recessive	AR
chr1	1020101	1056118	103320	375790	Myasthenic syndrome, congenital, 8, with pre- and postsynaptic defects, 615120 (3), Autosomal recessive	AR

HPO TABLE

EntrezGeneID	EntrezGeneSymbol	HPOTermName	HPOTermID
190	NR0B1	HP:0000147, HP:0000062, HP:0000026, HP:0008734...	Polycystic ovaries, Ambiguous genitalia, Male hypogonadism, Decreased testicular size...
182	JAG1	HP:0001636, HP:0000006, HP:0004209, HP:0000337...	Tetralogy of Fallot, Autosomal dominant inheritance, Clinodactyly of the 5th finger, Broad forehead...
274	BIN1	HP:0003803, HP:0008994, HP:0001561, HP:0002355...	Type 1 muscle fiber predominance, Proximal muscle weakness in lower limbs, Polyhydramnios, Difficulty walking...
162	AP1B1	HP:0008064, HP:0001406, HP:0011967, HP:0000962...	Ichthyosis, Intrahepatic cholestasis, Decreased circulating copper concentration, Hyperkeratosis...
288	ANK3	HP:0001290, HP:0000007, HP:0001257, HP:0000718...	Generalized hypotonia, Autosomal recessive inheritance, Spasticity, Aggressive behavior, Intellectual disability...

HPO-OMIM TABLE

EntrezGeneID	MIMNumber	GeneSymbols	Phenotypes	Inheritance	HPOTermName	HPOTermID
375790	103320	AGRN, CMS8	Myasthenic syndrome, congenital, 8, with pre- and postsynaptic defects, 615120 (3), Autosomal recessive	AR	HP:0000218, HP:0001265, HP:0010307...	High palate, Hyporeflexia, Stridor...
9636	147571	ISG15, G1P2, IFI15, IMD38	Immunodeficiency 38, 616126 (3), Autosomal recessive	AR	HP:0000007, HP:0002721, HP:0002716...	Autosomal recessive inheritance, Immunodeficiency, Lymphadenopathy...
8863	603427	PER3, FASPS3	?Advanced sleep phase syndrome, familial, 3, 616882 (3), Autosomal dominant	AD	HP:0000716, HP:0000006, HP:0031873	Depressivity, Autosomal dominant inheritance, Early chronotype
83715	606351	ESPN, USH1M	?Usher syndrome, type 1M, 618632 (3), Autosomal recessive; Deafness, autosomal recessive 36, 609006 (3), Autosomal recessive; Deafness, neurosensory, without vestibular involvement, autosomal dominant, 609006 (3), Autosomal recessive	AD, AR	HP:0007730, HP:0000375, HP:0000662	Iris hypopigmentation, Abnormal cochlea morphology, Nyctalopia

Figura 8. Esquema de las bases de datos empleadas en la herramienta web.

5. Importación y visualización de los datos

A continuación, se muestra un *pipeline* típico para importar y analizar los datos en la aplicación web.

En primer lugar, se debe acceder a la aplicación identificándose mediante un usuario previamente registrado. Una vez dentro de la aplicación, en el panel de opciones existen diferentes desplegables entre los cuales se encuentra "VCFiles" (ver Figura 9). Este apartado sirve para introducir los datos a analizar, ya sea un archivo VCF *de novo* o bien un archivo que ya esté disponible en la base de datos del usuario.

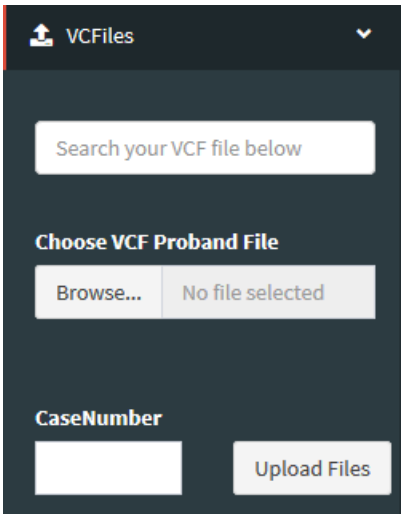


Figura 9. Desplegable *VCFiles.*

En el menú de opciones existen tres apartados relevantes. Un primer apartado destinado a la búsqueda de casos anteriores, subidos previamente; un segundo apartado que permite subir un archivo VCF para analizarlo posteriormente, ligado a un tercer apartado ("*CaseNumber*") que se utiliza para darle un nombre o código a este nuevo caso. Una vez seleccionada una de las dos opciones, se debe pulsar el botón "*Upload Files".* El proceso y carga de los casos puede tardar desde unos segundos, si se trata de un caso subido anteriormente, hasta unos minutos, si se trata de un caso nuevo, en función de la máquina y del número de variantes. El rendimiento de la aplicación se expondrá en apartados posteriores.

Una vez procesada e importada la lista de variantes, se mostrará una notificación de proceso completado, y se deberán seleccionar los filtros a aplicar. Para ello se ha diseñado un desplegable que permite seleccionar diferentes opciones en función de las preferencias del usuario (ver Figura 5B).

Tras aplicar todos los filtros se mostrará una tabla con los siguientes campos:

- **Classification:** Desplegable con 6 opciones de clasificación: *Not classified, pathogenic, likely pathogenic, uncertain significance, likely benign, benign.*
- **CHROM:** Cromosoma donde se ha detectado la variante
- **POS:** Posición inicial de la variante en el genoma
- **REF:** Alelo de referencia
- **ALT:** Alelo(s) alternativo(s)
- **Gene:** Gen afectado
- **FeatureID:** Tránscrito que se toma como referencia para las anotaciones HGVS.
- **HGVS.c:** Anotación a nivel codificante de la variante
- **HGVS.p:** Anotación a nivel proteico de la variante
- **Rank:** Exón/Intrón afectado
- **AlleleFrequency:** Frecuencia alélica
- **Zigosity:** Cigosidad (Heterocigosis/homocigosis)
- **Annotation:** Tipo de variante (p.ej. *missense variant, intron variant*)
- **dbSNP:** Número de acceso en la base de datos dbSNP (rs).
- **ClinVar:** Clasificación en la base de datos *ClinVar*
- **Phenotypes:** Fenotipos o enfermedades asociadas al gen afectado según la base de datos OMIM.
- **Inheritance:** Herencia de la enfermedad en OMIM.
- **MIMNumber:** Número de acceso a la enfermedad en la base de datos OMIM.
- **GENE_ID:** Número de acceso del gen afectado
- **AF_gnomAD:** Frecuencia poblacional en la base de datos gnomAD
- **Homozygotes:** Número de homocigotos sanos en la base de datos gnomAD

- **AnnotationImpact:** Impacto de la variante en cuatro niveles: HIGH, MODERATE, MODIFIER, LOW
- **CADD:** Puntuación obtenida por el predictor CADD
- **GERP:** Puntuación obtenida por el predictor GERP
- **MutationTaster:** Puntuación obtenida por el predictor Mutation Taster
- **Polyphen2:** Puntuación obtenida por el predictor Polyphen2
- **SIFT:** Puntuación obtenida por el predictor SIFT
- **FATHMM:** Puntuación obtenida por el predictor FATHMM
- **PhastCons30:** Puntuación obtenida por el predictor PhastCons
- **PhyloP30:** Puntuación obtenida por el predictor Phylop
- **QUAL:** Calidad de la lectura
- **DP:** Profundidad de lecturas de la variante

6. Filtrado y priorización de variantes

6.1. Filtrado de variantes

El filtrado de variantes se lleva a cabo para eliminar posibles artefactos o falsos positivos, consiguiendo así disminuir el número de variantes a analizar y evitando diagnósticos clínicos incorrectos. Todas aquellas variantes que no cumplan con los requisitos de calidad establecidos, se filtrarán. Los campos INFO y QUAL del archivo VCF contienen toda la información necesaria para el filtrado. Los criterios empleados en este proyecto corresponden con los recomendados por la plataforma *GATK* (5) (ver Tabla 2). Estos criterios se han escogido porque están ampliamente aceptados en el área de la secuenciación clínica.

Campo	Definición	Requisito
QUAL	Calidad de la secuenciación	Superior a 30.0
QD	Confianza de la variante (QUAL/DP)	Superior a 2.0
MQ	Calidad del alineamiento (*Mapping quality)*	Superior a 40.0
MQRankSum	Indicador de *strand-bias* (Test de Wilcoxon)	Superior a -12.5
FS	Indicador de *strand-bias* (*Fisher Strand*)	Inferior a 60.0
SOR	Indicador de *strand-bias* (Test *Symmetric Odds Ratio*)	Inferior a 3.0
ReadPosRank	Distancia desde el final de la lectura	Superior a -8

Tabla 2. Requisitos de filtrado de calidad

6.2. Priorización de variantes

En este Trabajo de Fin de Máster se lleva a cabo una aproximación que combina información a nivel de gen y a nivel de variante (ver Figura 10).

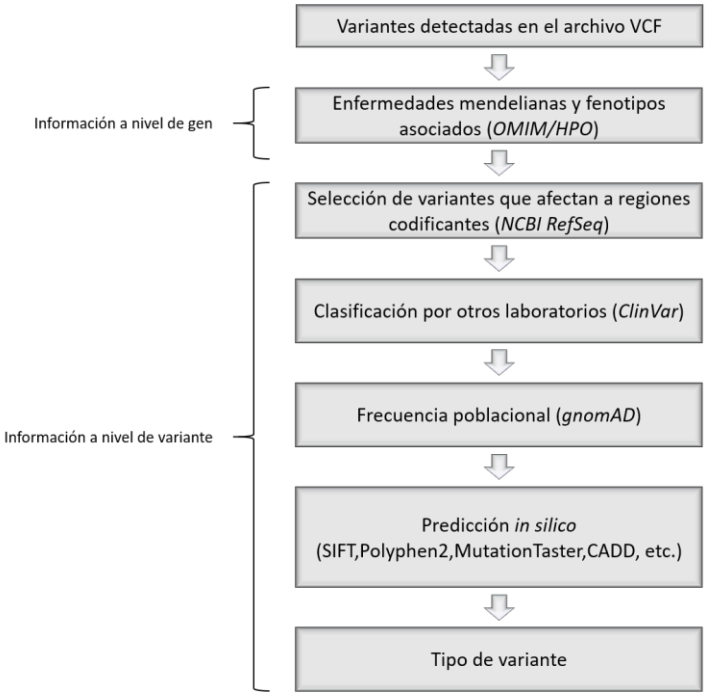


Figura 10. Esquema de priorización de variantes. Se detallan los pasos definidos para la priorización de variantes. A nivel de gen se tiene en cuenta la asociación con enfermedades mendelianas de la base de datos OMIM y HPO. A nivel de variante se tienen en cuenta la región a la que afecta, la clasificación en la base de datos *ClinVar,* la frecuencia poblacional determinada en *gnomAD*, la predicción *in silico* y el tipo de variante.

<u>Información a nivel de gen:</u>

Mediante la información a nivel de gen se puede conocer si existe asociación entre alguna enfermedad genética y el gen afectado, así como su modo (o modos) de herencia. Gracias a esta información se podrán priorizar las variantes que se encuentren en genes asociados a enfermedades con herencias autosómico-dominantes o ligadas al X (en estado de heterocigosis o hemicigosis) o las variantes que se encuentren en genes asociados con enfermedades de herencia autosómico-recesivas (en homocigosis o heterocigosis compuesta), que guarden relación con el fenotipo o enfermedad del paciente.

The Human Phenotype Ontology (*HPO*) es un proyecto que proporciona una ontología de fenotipos relevantes detectados en enfermedades humanas. Se trata de una base de datos que describe anomalías fenotípicas con un vocabulario estandarizado y su base molecular. Se presenta con una estructura similar a una jerarquía en la que todos los términos HPO de ésta están relacionados entre sí (30). La subontología *Phenotypic abnormality*, que contiene la descripción de las anomalías fenotípicas, es la que conferirá a la aplicación la posibilidad de priorizar en función del fenotipo clínico del paciente.

Por otro lado, el catálogo *Online Mendelian Inheritance in Man* (*OMIM*) (15) se centra en la asociación de una enfermedad concreta y el genotipo asociado a esta. Incluye todas enfermedades de base genética, y además ofrece información sobre su manifestación fenotípica cuando existe evidencia científica. Gracias a este catálogo se podrá priorizar en función de la herencia de la enfermedad asociada al gen afectado.

Información a nivel de variante:

La información a nivel de variante permite descartar un gran número de variantes así como priorizar otras en función de diversas propiedades que se describen a continuación.

En primer lugar, categorizar cada variante dependiendo de su relación con las secuencias codificantes del genoma y cómo puede afectar al producto génico (efecto molecular) permite priorizar aquellas variantes con un alto impacto sobre dicho producto génico (p.ej. deleciones; inserciones; *non-sense*; o variantes de *splicing*) y descartar aquellas variantes que tengan un impacto nulo *a priori* (p.ej. variantes sinónimas; aquellas que no afecten a regiones codificantes; o regiones intrónicas adyacentes, que no estén comprendidas en +-20 pares de bases del exón, fuera de las regiones donde se concentran la mayoría de variantes *patogénicas* descritas). Por otro lado, mediante bases de datos como *ClinVar* se puede conocer si una variante ha sido descrita previamente por algún laboratorio y, de ser así, obtener su clasificación por estos junto con los motivos para dicha clasificación y el fenotipo asociado. De

esta manera se podrán priorizar aquellas variantes descritas como causantes de enfermedad y descartar aquellas con indicios de benignidad.

Conocer la frecuencia poblacional de una variante, mediante la base de datos gnomAD, permitirá descartar aquellas con frecuencias muy altas pues es muy poco probable que alteraciones causales de una patología sean muy comunes en la población.

Por último, la predicción *in silico* para variantes *missense* o aquellas que puedan afectar al patrón normal de *splicing* es fundamental a la hora de priorizar variantes. Aunque ningún método de predicción *in silico* ha demostrado ser suficientemente confiable para clasificar por si solo estos tipos de variantes, sus puntuaciones de patogenicidad (*scores*) o probabilidades pueden ser empleados para priorizar variantes de significado incierto, o apoyar indicios obtenidos de otras fuentes.

Criterios *ACMG*:

Con la información disponible a nivel de gen y variante, existen una serie de guías y recomendaciones de la *ACMG* que establecen un conjunto sistemático de reglas cuya aplicación apoya o descarta la patogenicidad de una variante, y finalmente se combinan para clasificar la variante en una de las 5 categorías: *pathogenic* (P), *likely pathogenic* (LP), *variant of uncertain significance* (VUS), *likely benign* (LB), o *benign* (B) (ver Tabla 3). Aunque durante los últimos años se han desarrollado diversos métodos automáticos para asignar y combinar algunos tipos de evidencia, se trata de un desafío ya que muchos de los criterios son difícilmente automatizables al carecer de información disponible o necesitar de la interpretación de un profesional. Es por ello que se considera fundamental conocer estas recomendaciones, aplicadas extensamente en el ámbito clínico, y emplear correctamente la información mostrada y los filtros programados en la aplicación para una clasificación correcta y precisa de las variantes.

En este proyecto se han incorporado algunos de los criterios para priorizar las variantes. En primer lugar, el *ACMG* considera que variantes con frecuencias poblacionales superiores al 5% presentan una evidencia fuerte de impacto

benigno y variantes ausentes en las bases de datos poblacionales presentan una evidencia moderada de impacto *patogénico*, por lo que dependiendo de su frecuencia se pueden aplicar los criterios BA1, BS1, BS2 y PM2.

En segundo lugar, un criterio importante es la clasificación por otros laboratorios clínicos. Cada vez hay más laboratorios o fuentes acreditadas que comparten sus clasificaciones en las bases de datos públicas como *ClinVar,* de manera que se pueden utilizar sus clasificaciones para filtrar variantes *benignas* o *probablemente benignas,* o bien para filtrar variantes *patogénicas* o *probablemente patogénicas,* aplicando así los criterios BP6 y PP5 respectivamente.

En tercer lugar, uno de los criterios que puede ser útil en la priorización de variantes es la predicción *in silico.* No obstante, se debe tener en cuenta que las predicciones no se han validado como variantes *patogénicas*, que los diferentes algoritmos pueden dar resultados contradictores y que además pueden ser diferentes para distintos genes. En este aspecto, sólo se ha tenido en cuenta la predicción del algoritmo CADD, pues es uno de los algoritmos ampliamente utilizados, priorizando así los valores de puntuación más altos frente a los más bajos (PP3 frente BP4).

Por último, existen variantes llamadas *null* que a menudo tienen un impacto deletéreo sobre la proteína, lo que puede suponer la pérdida de la estructura y/o función de la proteína (PVS1). Por ello, se priorizan a su vez las variantes con un mayor impacto sobre la proteína (p.ej. *frameshift*, codón de stop), frente a las variantes que apenas presentan impacto a nivel proteico (p.ej. sinónimas, no codificantes).

	Benigna		Patogénica			
	Evidencia Fuerte	Evidencia de apoyo	Evidencia de apoyo	Evidencia moderada	Evidencia Fuerte	Evidencia Muy fuerte
Información poblacional	MAF es demasiado alta para la enfermedad (BA1/BS1) o la observación en controles sanos con la penetrancia de a enfermedad son inconsistentes (BS2)			Ausente en las bases de datos poblacionales (PM2)	La prevalencia en individuos afectos es estadísticamente superior a la observada en controles (PS4)	
Información predictiva y computacional		Numerosas fuentes predictivas sugieren que no hay un impacto en el producto génico (BP4)	Numerosas fuentes predictivas sugieren un efecto deletéreo en el producto génico (PP3)	*Missense* novel en un residuo aminoacídico donde se ha detectado anteriormente una *missense* diferente clasificada como *patogénica* (PM5) Cambio en la longitud de la proteína (PM4)	Mismo cambio de aminoácido clasificado como *patogénico* (PS1)	Predicción de variante *null* en un gen donde la pérdida de función es un mecanismo conocido de la enfermedad (PVS1)
Información funcional	Estudios funcionales demuestran que no existe un efecto deletéreo (BS3)		*Missense* en un gen con una baja tasa de variantes benignas de este tipo y donde la *missense* son mecanismos comunes de patogenicidad (PP2)	Región *hot-spot* o estudios funcionales del dominio demuestran que no existen variantes benignas (PM1)	Estudios funcionales demuestran un efecto deletéreo (PS3)	
Información de segregación	No segregación con la enfermedad (BS4)		Co-segregación con la enfermedad en múltiples miembros afectos de una familia (PP1)			
Información *de novo*				*De novo* sin confirmación paterna y materna (PM6)	*De novo* con confirmación paterna y materna (PS2)	
Infomación alélica		Observada en *trans* en combinación con una variante dominante (BP2)		Observada en *trans* en combinación con una variante *patogénica* en trastorno recesivo (PM3)		
Información de bases de datos		Fuentes fidedignas sin compartir los datos la clasifican como benigna (BP6)	Fuentes fidedignas la clasifican como *patogénica* (PP5)			
Otra información		Detectada en un caso con una causa alternatica (BP5)	Fenotipo del paciente o historia familia muy específica para el gen (PP4)			

Tabla 3. Criterios de clasificación del ACMG. El cuadro organiza cada uno de los criterios por el tipo de evidencia, así como por la fuerza que tienen los criterios para afirmar que una variante es *benigna* (lado izquierdo) o *patogénica* (lado derecho). Abreviaturas: BA, Evidencia independiente de impacto; BS, Fuerte evidencia de impacto benigno; BP, Evidencia de apoyo de impacto benigno; PM, Evidencia moderada de patogenicidad; PP, Evidencia de apoyo de patogenicidad; PS, Evidencia fuerte de patogenicidad; PVS, Evidencia muy fuerte de patogenicidad. Tabla reproducida de las guías ACMG (4).

6.2.1. Implementación del algoritmo de priorización

En apartados anteriores se han detallado las diferentes aproximaciones que se van a llevar a cabo para la priorización de variantes en este Proyecto. A continuación, se especifican las diferentes partes del algoritmo, de donde proviene la información empleada y el resultado obtenido.

Gracias a la anotación del archivo VCF, y el añadido posterior de los términos HPO y OMIM en la aplicación, se dispone de una gran cantidad de información que facilitará la tarea de filtrado y priorización. En la Figura 11 se resume el procedimiento estándar diseñado para la mayoría de los casos, basado en una serie de filtros opcionales cuyos parámetros serán seleccionados por el usuario.

En primer lugar, la aplicación web permite al usuario seleccionar un listado de términos HPO referentes al fenotipo del paciente, de forma que solo se mostrarán aquellas variantes que afecten a genes relacionados con los términos escogidos. En caso de no seleccionar ningún término HPO, se mostrarán todas las variantes del archivo VCF una vez realizado el filtrado de calidad.

Posteriormente, se puede aplicar un filtro relacionado con el tipo de variante (según la región del gen a la que afecta) que permitirá descartar aquellas variantes con pocas evidencias de causar efectos relevantes a la estructura o función de la proteína, como aquellas que afectan a regiones no codificantes. A continuación, si se sospecha de un patrón de herencia determinado por el fenotipo descrito o sus antecedentes familiares, se puede realizar un nuevo filtrado en función de modos de herencia. En la siguiente etapa, se pueden seleccionar variantes en función de su clasificación en la base de datos *Clinvar*, y posteriormente descartar aquellas variantes con una frecuencia poblacional incompatible con la incidencia de la enfermedad en la población. Por último, las variantes que han superado todo el proceso de filtrado anterior serán mostradas dando prioridad a variantes con un alto impacto sobre la proteína según la propia naturaleza de la variante, y como ordenamiento secundario a su efecto deletéreo predicho por la predicción *in silico*. Además, la aplicación

web incorpora una casilla para clasificar manualmente las variantes en los 5 niveles que describen las guías del *ACMG: pathogenic, likely pathogenic, uncertain significance, likely benign, benign.*

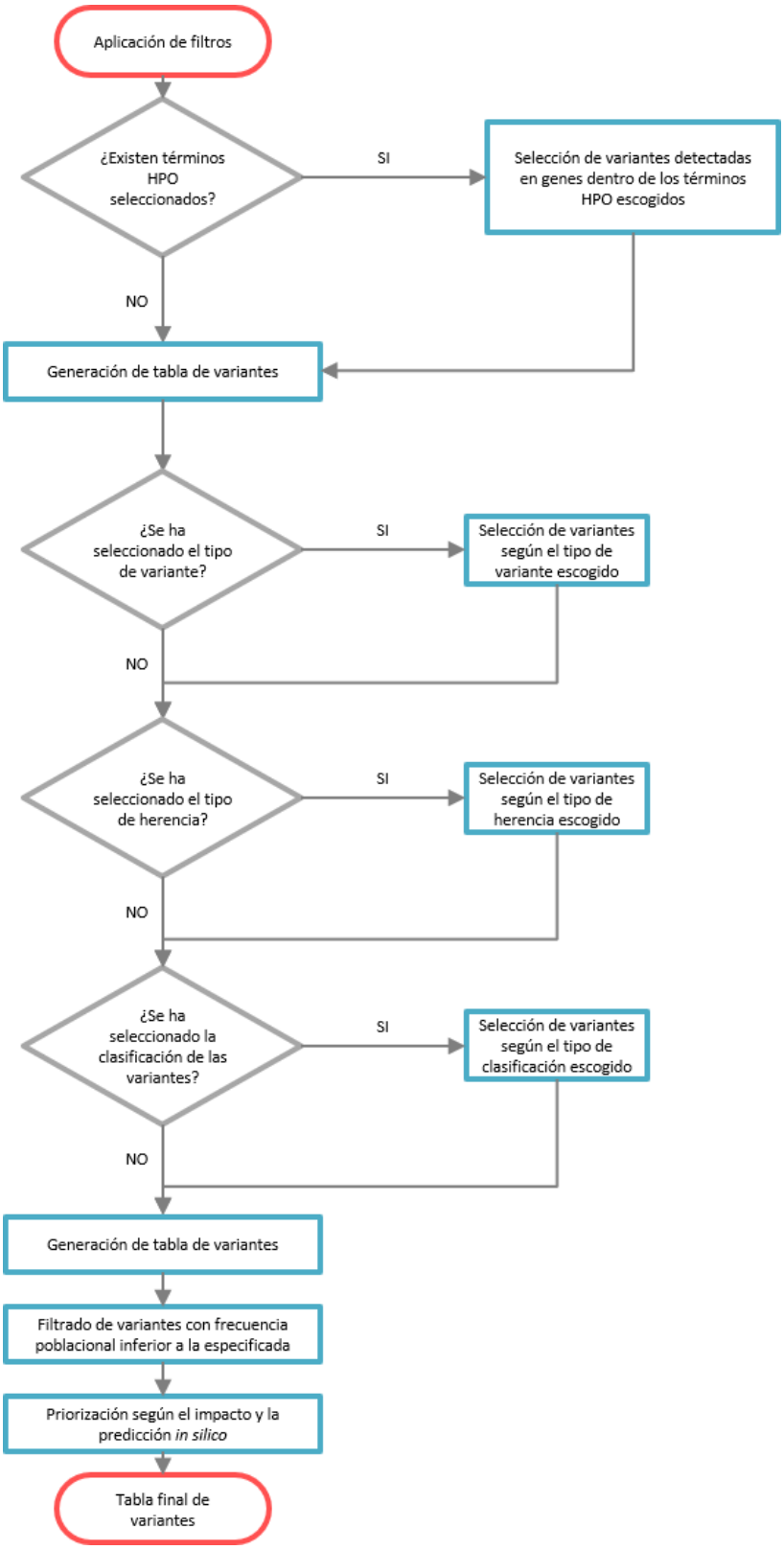


Figura 11. Algoritmo de priorización de variantes.

7. Pruebas de funcionalidad

Una de las ideas principales de este proyecto es complementar el análisis de variantes que realizan los analistas en el laboratorio. Por ello, se decide simular cuatro casos de pacientes con distintas patologías para evaluar la funcionalidad de la aplicación web. Los archivos se han generado partiendo del genoma de referencia NA12878, del repositorio *Genome in a bottle.* Se han escogido al azar 4 pacientes de la literatura médica con mutaciones causales conocidas y se han incorporado en el archivo VCF, simulando así un supuesto individuo afecto. Además, en uno de ellos, en concreto el caso a estudio número 3, se ha simulado un varón, eliminando aquellas variantes heterocigotas en el cromosoma X, quedando así solo variantes “en hemicigosis”.

El objetivo de estas pruebas de funcionalidad, por medio de los casos simulados, es evaluar si finalmente se obtienen las variantes causantes de la enfermedad descrita empleando los algoritmos de filtrado y priorización programados. A continuación, se muestran los resultados obtenidos en cada caso, junto a los pasos seguidos en el análisis.

7.1. Caso de prueba 1

Mujer sana sin enfermedad conocida con intereses reproductivos. En ocasiones, uno de los objetivos de la genética médica es detectar variantes *patogénicas* en *estatus* de portador asociadas a enfermedades autosómico recesivas o ligadas al X recesivas, con fines reproductivos. Con este objetivo, se buscan variantes *probablemente patogénicas* o *patogénicas* ya sea descritas en la base de datos *ClinVar* o que cumplan los criterios de clasificación del *ACMG*.

Filtro aplicado	Nº de variantes
Sin aplicar filtros	44,797
Aplicando filtros por defecto	362
Aplicando herencia autosómico recesiva y ligada al X recesiva	218
Aplicando *ClinVar pathogenic; likely pathogenic; Not ClinVar ID*	171
Aplicando *ClinVar pathogenic; likely pathogenic*	3

Tabla 4. Resultados caso 1. De las 171 variantes restantes, las 3 primeras son las variantes *patogénicas/probablemente patogénicas* descritas y en la séptima posición se encuentra la variante *null*.

En el estudio propuesto se han detectado tres alteraciones *probablemente patogénicas/patogénicas,* en estado de portador, descritas en la base de datos *ClinVar* así como una alteración que cumple criterios para ser variante *null*. Las alteraciones se muestran en la Tabla 5.

Variante detectada	Cigosidad	Enfermedad asociada	Herencia
NM_000391.4 (*TPP1*):c.509-1G>A	Het	Ceroid lipofuscinosis, neuronal, 2; Spinocerebellar ataxia, autosomal recessive 7	AR
NM_000277.3 (*PAH*):c.1241A>G p.(Tyr414Cys)	Het	Phenylketonuria	AR
NM_000301.5 (*PLG*):c.112A>G p.(Lys38Glu)	Het	Dysplasminogenemia; Plasminogen deficiency, type I	AR
NM_001309444.2 (*SPARC*):c.1024_1025delTA p.(Ter342fs)	Het	Osteogenesis imperfecta, type XVII	AR

Tabla 5. Tabla de variantes *probablemente patogénicas* o *patogénicas* detectadas en el caso a estudio 1.

7.2. Caso de prueba 2

Se trata de una niña de pocos meses con problemas de alimentación, retraso del crecimiento, y signos de daño hepático (ictericia, tendencia a hemorragias, hipoglucemia), y sin antecedentes familiares. Cuando las clínicas de los pacientes son muy ambiguas, como en este caso, una aproximación que facilita el análisis de variantes es el filtrado mediante un listado de genes. Mediante el empleo de un listado de términos HPO correspondientes a los diferentes signos y síntomas descritos en la paciente se consigue centrar el análisis en genes más específicos. Los términos HPO que se aplican en esta paciente son: HP:0011968, Feeding difficulties; HP:0000952, Jaundice; HP:0001892, Abnormal bleeding; HP:0001510, Growth delay; HP:0001943, Hypoglycemia.

Filtro aplicado	Nº de variantes
Sin aplicar filtros	44,799
Aplicando filtros por defecto	364
Aplicando términos HPO	85

Tabla 6. Resultados caso 2. De las 85 variantes restantes para el análisis, las variantes causales de patología se encuentran entre las 9 primeras.

En el estudio propuesto se han detectado dos alteraciones *probablemente patogénicas/patogénicas* descritas en la base de datos *ClinVar* compatibles con la clínica descrita. Ambas alteraciones podrían encontrarte en heterocigosis compuesta en la paciente. Las alteraciones se muestran en la Tabla 7.

Variante detectada	Cigosidad	Enfermedad asociada	Herencia
NM_000155.4 (*GALT*):c.563A>G(;)1030C>A p.(Gln188Arg)(;)(Gln344Lys)	Het, Het	Galactosemia	AR

Tabla 7. Tabla de variantes *probablemente patogénicas* o *patogénicas* detectadas en el caso a estudio 2.

7.3. Caso de prueba 3

Varón de 40 años a quien se le diagnosticó hemofilia a los 4 años y que padece hemartrosis desde los 13 años. Primo hermano también afecto.

En este caso, existen antecedentes familiares en los que otro varón de la familia también es afecto, por lo que cabría esperar que la variante estuviera ligada al sexo. Además, la mayoría de hemofilias descubiertas están ligadas al sexo. Es por ello que una posible aproximación sería la de filtrar en una primera instancia aquellas variantes en genes asociados con trastornos hemorrágicos mediante el término HPO HP:0001892, *Abnormal bleeding*. Y posteriormente, filtrar aquellas variantes que afecten a genes asociados a enfermedades ligadas al sexo, tanto dominantes como recesivas.

Filtro aplicado	Nº de variantes
Sin aplicar filtros	44,162
Aplicando filtros por defecto	356
Aplicando términos HPO	3
Aplicando filtro de herencias ligadas al X	1

Tabla 8. Resultados caso 3. La variante resultante tras el filtrado es la variante causante de la patología

En el estudio propuesto se han detectado una alteración *null probablemente patogénicas* compatible con la clínica descrita. La alteración se muestra en la Tabla 9.

Variante detectada	Cigosidad	Enfermedad asociada	Herencia
NM_000133.4 (*F9*):c.1262_1263delGG p.(Gly421fs)	Hem	Hemophilia B	XLR

Tabla 9. Tabla de variante *probablemente patogénica* detectada en el caso a estudio 3.

7.4. Caso de prueba 4

Paciente epiléptico con síndrome de Dravet. La madre del paciente tiene antecedentes de convulsiones febriles compatibles con GEFS+ (Epilepsia generalizada con crisis febriles plus). Dado que la madre del paciente presenta síntomas, es posible que la variante causante de la encefalopatía haya sido heredada mediante un patrón dominante, ya sea ligado al X o autosómico.

Filtro aplicado	Nº de variantes
Sin aplicar filtros	44,798
Aplicando filtros por defecto	363
Aplicando filtro de herencias de patrón dominante	201
Aplicando *ClinVar pathogenic; likely pathogenic*	1

Tabla 10. Resultados caso 4. La variante resultante tras el filtrado es la variante causante de la patología

En el estudio propuesto se ha detectado una alteración *patogénica* descrita en la base de datos *ClinVar.* La alteración se detalla en la Tabla 11.

Variante detectada	Cigosidad	Enfermedad asociada	Herencia
NM_001165963.4 (*SCN1A*):c.5126C>T p.(Thr1709Ile)	Het	Dravet syndrome; Epilepsy, generalized, with febrile seizures plus, type 2; Febrile seizures, familial, 3A; Migraine, familial hemiplegic, 3	AD

Tabla 11. Tabla de variante *patogénica* detectada en el caso a estudio 4.

8. Pruebas de rendimiento

A continuación, se muestra el tiempo consumido por la aplicación en función del número de variantes del archivo de entrada. Además, se muestra un desglose del tiempo empleado en cada etapa de la aplicación. Las pruebas de rendimiento de la aplicación se realizaron con un equipo que dispone de un procesador AMD A4-5000 1.5GHz (4 núcleos) con 8 GB de memoria RAM.

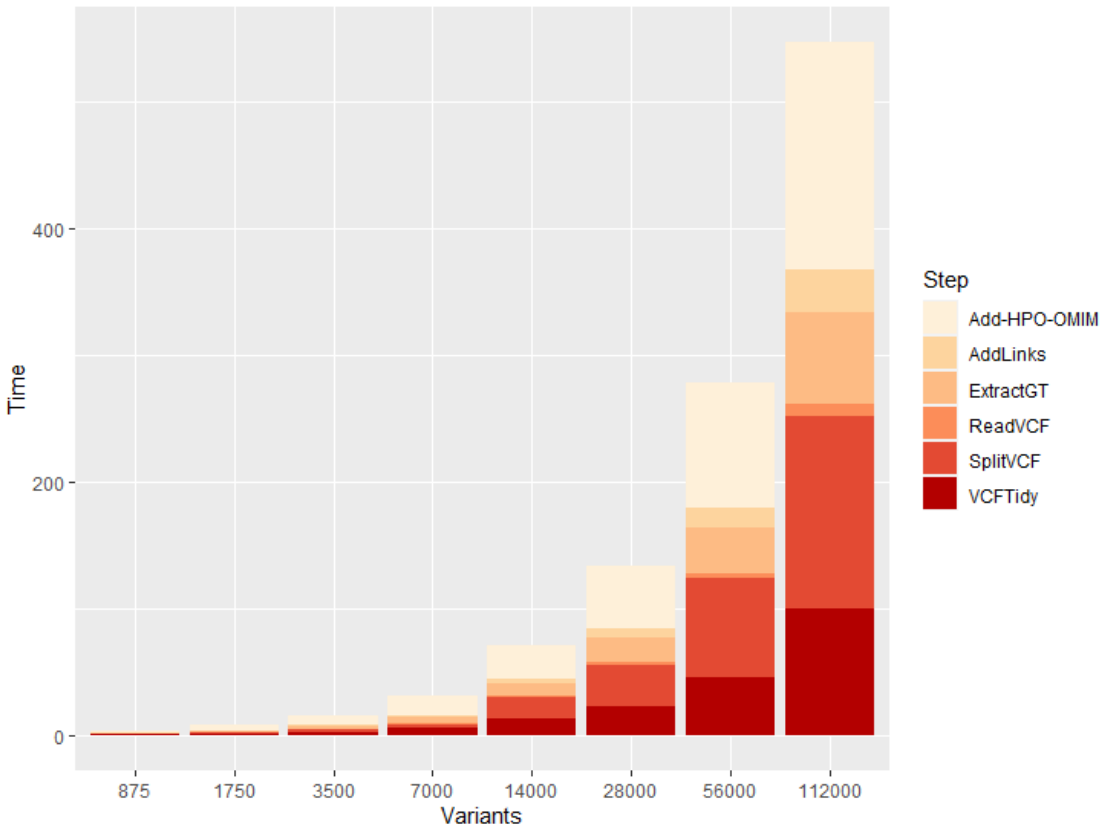


Figura 12. Representación del tiempo de ejecución (en segundos) frente al número de variantes del archivo VCF importado en la aplicación.

Observando la tendencia de los datos, se concluye que el tiempo de procesamiento se incrementa proporcionalmente con el número de variantes, a un ritmo de unas 11800 variantes por minuto, es decir, se observa una tendencia lineal en los datos. Los procesos o etapas que más tiempo consumen son la anotación de fenotipos mediante las bases de datos HPO y OMIM así como la incorporación de *links* a los distintos campos de la tabla de variantes.

9. Conclusiones

El desarrollo de este Trabajo de Fin de Máster me ha aportado, entre otras, las siguientes habilidades y capacidades:

- Adquirir conocimientos sobre el desarrollo de aplicaciones web y la complejidad que suponen, así como profundizar en áreas de la bioinformática desconocidas para mí.
- Valorar la importancia de definir unos objetivos específicos y una planificación adecuada al tiempo y recursos disponibles.
- Trabajar de forma independiente, tomar decisiones ante inconvenientes imprevistos y resolver los problemas presentados.

El objetivo principal era desarrollar una aplicación que combinara una interfaz adaptada a la visualización de datos de variantes con sencillez de uso, junto a algoritmos de filtrado y priorización que facilitaran y enfocaran el análisis a variantes causantes de enfermedad. Aunque la aplicación tiene sus limitaciones en comparación con otras herramientas similares tanto comerciales como de libre uso, considero que la aplicación desarrollada es práctica y eficiente, obteniendo un resultado satisfactorio teniendo en cuenta las limitaciones de tiempo y recursos en el contexto de un Trabajo Fin de Máster, cumpliéndose así los objetivos generales y específicos de este Trabajo.

La planificación inicialmente propuesta se ha cumplido en términos generales. Sin embargo, debido a situaciones imprevistas, se han producido desviaciones que finalmente han podido ser compensadas. Estas desviaciones se han producido principalmente por la dificultad de estimar *a priori* el tiempo y el esfuerzo necesarios para diseñar e implementar algunos elementos de la aplicación, cuya complejidad no fue correctamente evaluada. Por otro lado, en alguna ocasión ha sido necesario ir hacia atrás en el desarrollo para corregir problemas del diseño que no estaban previstos e implementar alternativas. Por último, se ha dedicado tiempo no previsto en implementar mejoras recomendadas durante la evaluación de las distintas PECs.

Respecto a las líneas de trabajo futuras:

- Aunque las pruebas de funcionalidad realizadas mediante casos simulados han sido exitosas, es necesario un testeo mucho más riguroso con casos reales y usuarios con experiencia en este tipo de aplicaciones con el fin de afinar y mejorar la aplicación mediante su *feedback*.
- Tanto el proceso de anotación como el filtrado de calidad están dirigidos a secuenciadores *Illumina* y GATK como *Variant Caller*. Por ello, es necesario ampliar el abanico a otras plataformas y *pipelines* bioinformáticos, e incluir el proceso de anotación en la aplicación.
- La interfaz gráfica podría mejorar con la inclusión de un *browser* genómico para la visualización de variantes y archivos de alineamientos, permitiendo así mostrar el aspecto de las lecturas portadoras de las variantes, facilitando el filtrado de artefactos o variantes en regiones problemáticas. También sería interesante para observar la presencia de mutaciones *patogénicas* cercanas.
- Tras las pruebas de rendimiento, existen etapas que consumen una gran cantidad de tiempo en el procesamiento inicial. Se plantea la posibilidad de optimizar estas etapas.

10. Glosario

- NGS: Secuenciación de segunda generación (*Next-Generation Sequencing)*
- VCF: *Variant Call Format*
- ACMG: *American College of Medical Genetics and Genomics*
- SNP: *Single Nucleotide Polymorphism*
- SNV: *Single Nucleotide Variant*
- WES: Secuenciación Completa de Exoma (*Whole Exome Sequencing*)
- INDEL: Inserción-Deleción
- CNV's: Variación en el Número de Copia (*Copy Number Variation*)
- SV's: Variación Estructural (*Structural Variation*)
- GIAB: *Genome In A Bottle*
- GFF: *Generic Feature Format*
- dbSNP: *Single Nucleotide Polymorphism Database*
- dbSNFP: *Functional Predictions and Annotations for human Nonsynonymous and Splice-Site Database*
- gnomAD: *The Genome Aggregation Database*
- PGVWeb: *Priorization of Genetic Variants Web*
- OMIM: *Online Mendelian Inheritance in Man*
- HPO: *Human Phenotype Ontology*

11. Bibliografía

1. Roca I, Fernández-Marmiesse A, Gouveia S, Segovia M, Couce ML. Prioritization of Variants Detected by Next Generation Sequencing According to the Mutation Tolerance and Mutational Architecture of the Corresponding Genes. Int J Mol Sci [Internet]. 2018 May 27;19(6). Available from: http://www.ncbi.nlm.nih.gov/pubmed/29861492
2. Dashti MJS, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. Biotechniques [Internet]. 2017;62(1):18–30. Available from: http://dx.doi.org/10.2144/000114492
3. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature [Internet]. 2015;526(7571):68–74. Available from: http://dx.doi.org/10.1038/nature15393
4. Laboratories KD, Genetics M, Health O, Road P, Molecular C, Children N, et al. Standards and Guidelines for the Interpretation of Sequence Variants (ACMG, 2015). 2015;17(5):405–24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25741868
5. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res [Internet]. 2010 Sep;20(9):1297–303. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20644199
6. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
7. https://samtools.github.io/hts-specs/, Octubre 2020.
8. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) [Internet]. 6(2):80–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22728672

9. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res [Internet]. 2010 Sep;38(16):e164. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20601685
10. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016;17(1):1–14. Available from: http://dx.doi.org/10.1186/s13059-016-0974-4
11. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature [Internet]. 2020;581(7809):434–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32461654
12. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res [Internet]. 2018;46(D1):D1062–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29165669
13. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res [Internet]. 2001 Jan 1;29(1):308–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11125122
14. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res [Internet]. 2019;47(D1):D1018–27. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30476213
15. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res [Internet]. 2005 Jan 1;33(Database issue):D514-7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15608251
16. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol [Internet]. 2008 Oct 9;26(10):1135–45. Available from: http://www.nature.com/articles/nbt1486
17. Niguidula N, Alamillo C, Shahmirzadi Mowlavi L, Powis Z, Cohen JS, Farwell Hagman KD. Clinical whole-exome sequencing results impact

medical management. Mol Genet genomic Med [Internet]. 2018;6(6):1068–78. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30318729

18. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet [Internet]. 2014 Jan 7;59(1):5–15. Available from: http://www.nature.com/articles/jhg2013114
19. Robinson PN, Kohler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res [Internet]. 2014 Feb 1;24(2):340–8. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.160325.113
20. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med [Internet]. 2014 Sep 3;6(252):252ra123-252ra123. Available from: https://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.3009262
21. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods [Internet]. 2015 Sep 20;12(9):841–3. Available from: http://www.nature.com/articles/nmeth.3484
22. James RA, Campbell IM, Chen ES, Boone PM, Rao MA, Bainbridge MN, et al. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. Genome Med [Internet]. 2016 Dec 2;8(1):13. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0261-8
23. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol [Internet]. 2014 Mar 16;32(3):246–51. Available from: http://www.nature.com/articles/nbt.2835
24. Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, et al. A standard variation file format for human genome sequences. Genome Biol [Internet]. 2010;11(8):R88. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20796305
25. Picard Toolkit. Broad Institute. 2019. p. GitHub Repository.

http://broadinstitute.github.io.

26. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie JM. Shiny: Web Application Framework for R. 2020. p. R package version 1.5.0. https://CRAN.R-project.or.
27. Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour [Internet]. 2017 Jan;17(1):44–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27401132
28. Kirill Müller ORCID iD, Hadley Wickham, David A. James, Seth Falcon, SQLite Authors, Liam Healy, R Consortium Rs. RSQLite: “SQLite” Interface for R. 2020. p. R package version 2.2.1. https://CRAN.R-project.or.
29. Yihui Xie, Joe Cheng, Xianying Tan, JJ Allaire, Maximilian Girlich, Greg Freedman Ellis, Johannes Rauh, jQuery contributors, SpryMedia Limited, Brian Reavis, Leon Gersen, Bartek Szopka (jquery.highlight.js in htmlwidgets/lib), RStudio P [cph]. DT: A Wrapper of the JavaScript Library “DataTables.” 2020. p. R package version 0.16. https://CRAN.R-project.org.
30. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res [Internet]. 2017;45(D1):D865–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27899602