Diseño de una herramienta web para la priorización de variantes genómicas detectadas mediante secuenciación masiva.

Roser Martínez Rubio

Máster en Bioinformática y Bioestadística
TFM-Bioinformática y Bioestadística Área 1

Joan Maynou Fernández

Javier Luis Cánovas Izquierdo
Marc Maceira Duch

Junio de 2020

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

Copyright © 2020 Roser Martínez Rubio.

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Diseño de una herramienta web para la priorización de variantes genómicas detectadas mediante secuenciación masiva.*
Nombre del autor:	*Roser Martinez Rubio*
Nombre del consultor/a:	*Joan Maynou Fernández*
Nombre del PRA:	*Javier Luis Cánovas Izquierdo* *Marc Maceira Duch*
Fecha de entrega:	06/2020
Titulación::	*Máster en Bioinfomática y Bioestadística*
Área del Trabajo Final:	*TFM-Bioinformática y Bioestadística Área 1*
Idioma del trabajo:	*Castellano*
Palabras clave	*Priorización de variantes, HPO, NGS, patogenicidad, clasificación de variantes, relación genotipo-fenotipo*

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

Las nuevas tecnologías de secuenciación masiva permiten obtener de forma rápida y económica una gran cantidad de información en forma de secuencias de ácidos nucleicos. Este hecho, que *a priori* es una ventaja, conlleva a su vez el inconveniente del gran volumen de información a tratar. En este aspecto, la Bioinformática juega un papel fundamental realizando herramientas que permitan detectar los cambios respecto a un genoma de referencia y establecer cuáles de estos cambios tienen relación con la enfermedad. En este segundo punto es donde nos centraremos.

Por tanto, la finalidad de nuestro trabajo será realizar una aplicación que permita a los genetistas filtrar y priorizar de forma sencilla y amigable las variantes genómicas detectadas con el fin de determinar cuál o cuáles de ellas

son las responsables de los fenotipos observados.

Para realizarla trabajaremos con el lenguaje de programación R sobre archivos en formato .vcf que contengan la información anotada sobre las variantes de los individuos a estudiar.

El resultado que hemos obtenido es una aplicación que permite seleccionar aquellas variantes que se encuentran en determinados genes o relacionadas con determinados fenotipos (mediante el usa de términos HPO) y que cumplen unos requisitos elegidos por el usuario (genetista), como el tipo de herencia, la frecuencia alélica o el significado clínico de la variante.

Abstract (in English, 250 words or less):

New massive sequencing technologies allow a large amount of information to be obtained quickly and cheaply in the form of nucleic acid sequences. This fact, which a priori is an advantage, in turn entails the inconvenience of the large volume of information to be processed. In this aspect, Bioinformatics plays a fundamental role making tools that allow detecting changes with respect to a reference genome and establishing which of these changes are related to the disease. This second point is where we will focus.

Therefore, the purpose of our work will be to carry out an application that allows geneticists to filter and prioritize in a simple and friendly way the detected genomic variants in order to determine which one or which of them are responsible for the observed phenotypes.

To carry it out we will work with the R programming language on files in .vcf format that contain the annotated information on the variants of the individuals to be studied.

The result that we have obtained is an application that allows us to select those variants that are found in certain genes or related to certain phenotypes (using HPO terms) and that meet certain requirements chosen by the user (geneticist), such as the type of inheritance , the allelic frequency or the clinical significance of the variant.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

1.1.1 Descripción general:

Una enfermedad genética es un trastorno que se origina debido a una alteración en el material genético. Estas pueden ser monogénicas, donde la alteración en un solo gen produce una enfermedad concreta (por ejemplo, la enfermedad de huntington o la fibrosis quística) y siguen un patrón de herencia mendeliana, o poligénicas o complejas, donde intervienen más de un gen y además el ambiente puede influir en el desarrollo de la enfermedad. Este segundo tipo de enfermedades genéticas son las mayoritarias. Por tanto, las enfermedades genéticas engloban a un grupo de patologías muy amplio y heterogéneo. Respecto a las monogénicas podemos decir que según el catálogo OMIM (Online Mendelian Inheritance in Man), se conocen alrededor de 8.000 enfermedades genéticas relacionadas con todas las especialidades médicas [1]. Muchas de estas enfermedades son de baja prevalencia y en general son de difícil diagnóstico debido a que se desconoce la causa molecular exacta que la origina (es decir el gen exacto), por lo que, mediante el uso de métodos clásicos de secuenciación (Método Sanger) en muchos casos solía transcurrir un tiempo considerable desde la aparición de los primeros síntomas hasta la correcta identificación de la patología. Se estima que el tiempo para el diagnóstico molecular se encuentra entre 1 y 10 años usando secuenciación Sanger [2,3].

Con la aparición de las nuevas técnicas de secuenciación masiva (NGS-next generation sequencing) hemos pasado de obtener la secuenciación de unos pocos cientos de pares de bases por reacción (método Sanger) a más de un millón de bases por reacción (mediante NGS) [3,4]. De esta forma, se ha conseguido secuenciar toda la región codificante de un individuo o exoma (WES-whole exome sequencing) o incluso todo su genoma (WGS- whole genome sequencing) en muy poco tiempo y con un coste relativamente bajo [5].

Pero al mismo tiempo genera un gran número de datos que hay que saber tratar y analizar para que, efectivamente, esto sea útil.

Tras la secuenciación se obtienen millones de pequeñas lecturas (reads) (se trata de secuencias de entre 30 a 100 pares de bases, el número de lecturas y la longitud de las mismas dependen de la tecnología que se use al secuenciar) en un fichero en formato FastQ, estas se alinean con un genoma de referencia y se genera un fichero llamado sequence alingment map (SAM) que se transforma en binary alingment map (BAM), donde se tiene la posición genómica de cada lectura. De esta forma conseguimos localizar las variantes, tanto de nucleótidos simples (SNVs) o pequeñas deleciones inserciones (INDELS), comparando las secuencias obtenidas con un genoma de referencia, generándose un archivo de tipo VCF (variant calling format) [6].

Cuando secuenciamos un exoma se obtienen aproximadamente de 20000 a 40000 variantes entre SNVs e INDELS. Tras el filtrado de estas obtenemos aproximadamente de 400 a 700 variantes de utilidad clínica. Donde aproximadamente un 14% son falsos positivos, debidos a que se trata de regiones con pocas lecturas o alineaciones subóptimas, entre otras causas. Posteriormente se aplican criterios de clasificación y priorización para determinar el grado de patogenicidad de la variante y la implicación con la patología del individuo estudiado.

Actualmente existen multitud de aplicaciones para la priorización de variantes, cada una de ellas con sus determinadas características, es decir, distintos formatos de archivos aceptados, distintos criterios de priorización, etc. Nombramos algunos ejemplos; QueryOR [3], VPot [7], wANNOVAR [8], The Ensembl variant effect predictor (VEP) [9], BierApp [10], BrowseVCF [11], VariantRanker [12], OVA [13] y PhenIX [14], estas dos últimas tienen en cuenta el fenotipo del paciente a la hora de realizar la priorización y, concretamente, PhenIX usa los términos HPO para este propósito.

Con todo esto, este es el tema sobre el cual va a tratar el presente TFM, es decir, la priorización de variantes obtenidas mediante secuenciación masiva. Con el fin de intentar disminuir al mínimo el tiempo en el diagnóstico, lo cual es

de suma importancia, ya que este retraso impide que los pacientes recibieran medidas terapéuticas y de rehabilitación específicas y que sus familiares entren en programas preventivos y recibieran asesoramiento genético adecuado.

1.1.2. Justificación del TFM:

Aunque los criterios de priorización que se establecen para el filtraje de variantes se podrían aplicar directamente sobre el fichero obtenido tras la comparación con el genoma de referencia mediante un terminal. Para los genetistas clínicos, con muchos conocimientos en genética, pero pocos en informática, este sería un ambiente hostil. Por tanto, el problema que deseamos resolver mediante este TFM es la creación de un ambiente amigable para el genetista clínico donde poder aplicar los filtros y criterios de selección sobre las variantes a estudiar.

1.2 Objetivos del Trabajo

1.2.1. Objetivos generales:

- Diseñar una herramienta web capaz de priorizar entre las variantes detectada mediante secuenciación masiva.

1.2.2. Objetivos específicos:

- Establecer criterios de selección de las variantes, con el fin de poder diferenciarlas entre polimorfismos poblacionales y variantes patogénicas, mediante el uso de filtros (frecuencias alélicas, filtros de tipo de herencia, etc…).
- Establecer criterios de priorización de las variantes seleccionadas, con el fin de asociarlas a determinadas patologías.

1.3 Enfoque y método seguido

En primer lugar, debemos obtener los datos con los que trabajar. Estos los obtendremos del proyecto “Genome in a bottle”, aquí encontramos ficheros VCF tanto para un caso índice (NA12878) como para un trio (Ashkenazim Trio).

Una vez obtenidos los datos debemos anotarlos, esto quiere decir, determinar qué tipos de variantes se incluyen dentro del fichero con formato VCF, para ello se utilizará el software de anotación snpEff y usando para ello, además de la anotación proporcionada por el programa, las bases de datos de ClinVar, dbSNP y dbNSFP.

Finalmente, implementaremos la aplicación con el paquete “shiny” de R que nos permite crear una interfaz de usuario amigable combinado con otros paquetes como “vcfR” y “dplyr” que nos permiten tratar de forma sencilla los datos a analizar.

1.4 Planificación del Trabajo

1.4.1. Tareas:

- Familiarización con los datos y metodología a usar (tanto para la anotación de variantes como para el diseño de la aplicación web) y evaluación el estado del arte, analizando otras aplicaciones web.
- Definir los filtros a usar para la selección de variantes.
- Definir los criterios a seguir para la priorización de las variantes seleccionadas.
- Diseño de la aplicación web
- Incluir en la aplicación web los criterios de para el filtraje y priorización de las variantes.
- Valoración de la aplicación.
- Escritura de la memoria.
- Elaboración de la presentación.

1.4.2 Calendario:

Se muestra el diagrama de Gantt con la planificación temporal de las tareas a realizar (Fig. 1).

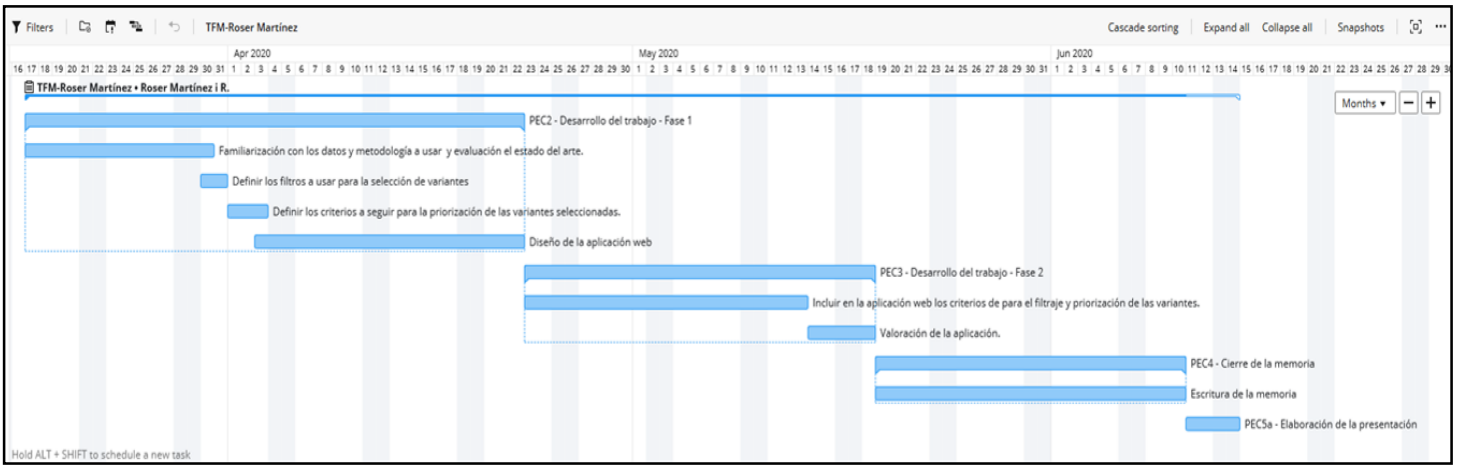


Fig 12: **Diagrama de Gantt.**

1.5 Breve sumario de productos obtenidos

- Plan de trabajo
- Memoria
- Producto: un software desarrollado.
- Presentación virtual
- Autoevaluación del proyecto

1.6 Breve descripción de los otros capítulos de la memoria

El resto de los capítulos explican cómo ha sido el proceso de creación de la aplicación. Consta de los siguientes puntos:

2.1 Resultados

2.1.1 Obtención de los recursos necesarios (Bases de datos y "muestras" usadas).

2.1.2 Anotación de los datos (Explica cómo se ha realizado la anotación de los archivos con la información de las muestras con las bases de datos seleccionadas).

2.1.3 Criterios de priorización y filtraje (Explica qué criterios se van a usar en la priorización de las variantes).

2.1.4 Implementación de la aplicación web (Explica el desarrollo de la aplicación y que paquetes de R se han usado).

2.1.5 Partes y uso de la aplicación (se describe la aplicación ya terminada).

2. Resto de capítulos

2.1 Resultados

2.1.1 Obtención de los recursos necesarios

2.1.1.1 Datos

Los datos que utilizaremos para probar el funcionamiento de nuestra aplicación los obtendremos del consorcio “Genome in a botlle” (GIAB)

Se trata de un consorcio público-privado-académico organizado por el NIST (National Institute of Standars and Technology) para desarrollar la infraestructura técnica (es decir, estándares, métodos y datos de referencia) que permita el paso de la secuenciación completa del genoma humano a la práctica clínica. La prioridad de GIAB es la caracterización de genomas humanos para su uso en validaciones analíticas y desarrollos, optimizaciones y demostraciones de tecnologías [15].

Actualmente, GIAB cuenta con la secuencia de un genoma piloto caracterizado (NA12878/HG001) proveniente del proyecto “HapMap”, y dos tríos hijo/padre/madre, uno de ellos de ascendencia judía Ashkenazi (HG002/HG003/HG004) y el otro China (HG005/HG006/HG007), ambos del proyecto “Personal Genome Project” (PGP). Nosotros utilizaremos los cuatro primeros genomas. La información de estos se muestra en la Tabla 1.

Genome	Coriell cell line ID	NIST ID	NCBI BioSample	PGP ID	Sex
CEPH Mother/Daughter	GM12878	HG001	SAMN03492678	Not PGP	Female
AJ Son	GM24385	HG002	SAMN03283347	huAA53E0	Male
AJ Father	GM24149	HG003	SAMN03283345	hu6E4515	Male
AJ Mother	GM24143	HG004	SAMN03283346	hu8E87A9	Female

Tabla 1. Identificadores asociados con los genomas caracterizados actualmente se por GIAB.

GIAB ha creado un método para la determinación de las variantes portadoras en estos genomas, produciendo así uno archivos en formato .vcf y .bed, para nuestro trabajo obtendremos la última versión de los ficheros con formato .vcf y alineados sobre la versión del genoma humano GRCh37.

2.1.1.2 Bases de datos (CinVar, bdSNP y l'altra)

Para la anotación de los archivos obtenidos hemos usado las siguientes bases de datos:

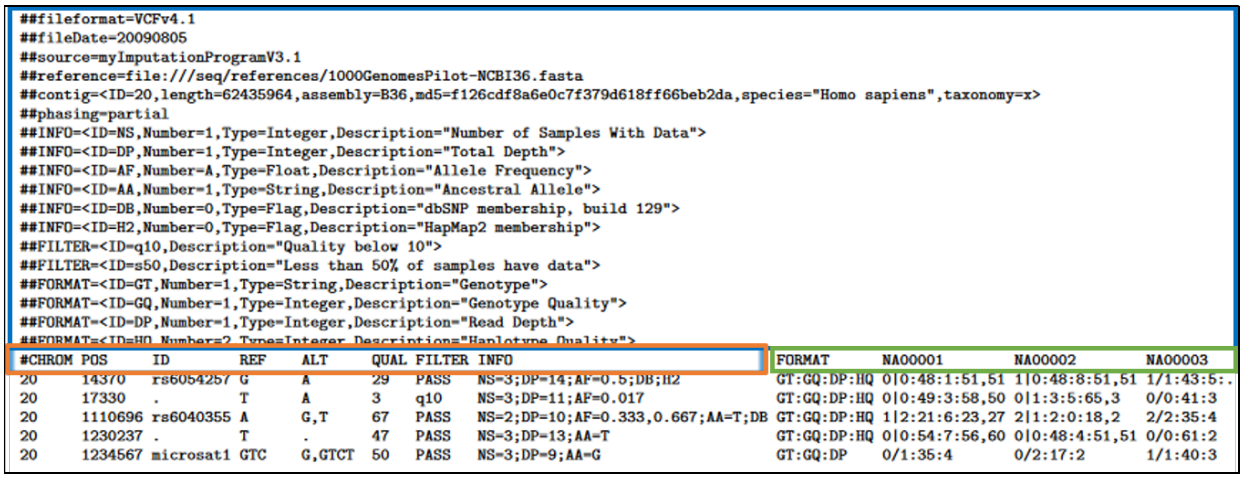
- ClinVar: base de datos pública y de acceso gratuito mantenida por el "National Institutes of Health" (NIH) donde se relacionan las variantes genómicas con su significado clínico, es decir con la capacidad de esta variante de producir o no enfermedad [16–18]. Las clasificaciones que encontramos en ella son: Patogénicas, Probablemente patogénicas, Variantes de significado incierto, Probablemente benignas, Benignas. Además, Factor de riesgo, Asociación, Protectora, Respuesta a drogas y Conflicto en la interpretación de la patogenicidad.

- dbNSFP: base de datos desarrollada para la predicción funcional y la anotación de todas las variantes de un solo nucleótido no sinónimos (nsSNV) en el genoma humano. Recopila puntuaciones de predicción de 36 algoritmos de predicción (SIFT, SIFT4G, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster2, MutationAssessor, FATHMM, MetaSVM, MetaLR, CADD, VEST4, PROVEAN, FATHMM-MKL codificación, FATHMM-XF x 4, LINSIGHT, DANN, GenoCanyon, Eigen, Eigen-PC, M-CAP, REVEL, MutPred, MVP, MPC, PrimateAI, GEOGEN2, BayesDel_addAF, BayesDel_noAF, ClinPred, LIST-S2, ALoFT), 9 puntuaciones de conservación (PhyloP x 3, phastCons x 3, GERP ++, SiPhy y bStatistic) y otra información relacionada, incluidas las frecuencias alélicas [19,20].

2.1.2 Anotación de los datos

El programa utilizado para realizar la anotación de variantes ha sido snpEff [21].

Como hemos comentado anteriormente, los datos de los genomas que hemos obtenido para la realización de este proyecto se encuentran en formato .vcf. Este es un formato de archivo de texto. Contiene líneas de metainformación, un encabezado y luego líneas de datos que contienen información sobre una posición en el genoma (Figura 2). En el encabezado se muestra el nombre de las 8 columnas fijas que son obligatorias (CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO). Mediante el snpEff añadimos a la columna "INFO" la información que nos proporciona las bases de datos sobre cada una de las variantes, usando para esto determinados comandos en la terminal de Linux.



```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A      29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.
20     17330   .         T   A      3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T   .      47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC G,GTCT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Fig 13. Ejemplo de las primeras líneas de un archivo en formato .vcf. Encuadrado de azul se muestran la metainformación. De naranja los nombres de las columnas fijas y obligatorias que se muestran en el encabezado y en verde las columnas opcionales. El resto de las líneas nos dan información sobre una posición concreta del genoma.

Para realizar dicha anotación, seguimos el siguiente *pipeline* de comandos para cada uno de los ficheros a anotar, esto siempre desde la carpeta que contiene el programa:

1. Anotación mediante snpEff:

~/snpEff$ java -Xmx4g -jar snpEff.jar GRCh37.75 *ruta*/archivo.vcf > archivo_ann.vcf

2. Anotación mediante dbNsfp:

~/snpEff$ java -jar SnpSift.jar dbnsfp -v -db dbNSFP2.9.txt.gz *ruta*/archivo_ann.vcf > archivo_ann_dbNSFP.vcf

3. Anotación mediante ClinVar:

~/snpEff$ java -jar SnpSift.jar annotate -a *ruta*/clinvar.GRCh37.vcf *ruta*/ archivo_ann_dbNSFP.vcf > archivo_ann_dbNSFP_clinvar.vcf

Una vez tenemos la anotación realizada, el archivo resultante tiene toda la información necesaria para poder trabajar con él.

2.1.3 Criterios de priorización y filtraje:

2.1.3.1 Criterios de priorización:

Antes de empezar a implementar la aplicación y una vez tenemos los datos preparados para poder trabajar con ellos. Debemos establecer que criterios de priorización y filtraje vamos a desarrollar en nuestra aplicación.

Como priorización entendemos al hecho de establecer sobre cuáles de todas las variantes halladas vamos posteriormente a aplicar los filtros. Desde un punto de vista clínico, nos interesa seleccionar estas variantes teniendo en cuenta la patología de nuestro paciente con el fin de determinar cuál o cuáles son las variantes causantes de dicha patología. Por tanto, para realizar esta priorización usaremos principalmente dos estrategias. Estas serán, o bien seleccionar aquellas variantes que se encuentran en determinados genes asociados a la patología en concreto, introduciendo en nuestra aplicación directamente los genes que se desea estudiar, o bien mediante la selección de las variantes que se encuentran en genes asociados a rasgos fenotípicos mostrados en el paciente cuando no se puede establecer con seguridad que enfermedad es la que sufre el paciente. Esta segunda estrategia se realizará mediante la introducción de términos de la ontología del fenotipo humano (HPO, Human Phenotype Ontology). Esta proporciona un vocabulario estandarizado de anormalidades fenotípicas encontradas en las enfermedades humanas y se asocian a términos (Por ejemplo, el rasgo fenotípico de

Macrocefalia tiene asociado el termino HPO HP:0000256) . Por tanto, cada término en el HPO describe una anormalidad fenotípica que puede llevar asociados la alteración de varios genes [22].

Además, también incluiremos la posibilidad de priorizar por una serie de genes recomendados por la ACMG (American College of Medical Genetics and Genomics), esto es una lista mínima de genes que se informaran como hallazgos incidentales o secundarios, el objetivo de este reporte es identificar y gestionar los riesgos derivados de los trastornos genéticos que se presentan con alta penetrancia [23].

2.1.3.2 Criterios de filtraje:

Una vez ya tenemos seleccionadas las variantes en los genes que clínicamente nos interesan pasaremos a aplicarles filtros para establecer cuáles de ellas son, efectivamente, las causantes de la enfermedad y no son polimorfismos. En este sentido, planteamos dos tipos de filtrajes, por un lado, que se pueda aplicar un filtro de herencia a los datos y por otro lado que se le apliquen determinados filtros relacionados con la naturaleza de la variante en sí.

Los filtros de herencia consistirán en buscar variantes en nuestros datos que cumplan determinados criterios relacionados con los patrones de herencia mendelianos, es decir, si siguen un patrón autosómico (no asociado a los cromosomas sexuales) dominante o recesivo o si esta herencia está ligada al cromosoma X. Al realizar estos filtros también deberemos tener en cuenta si algunos de los progenitores son afecto o si ninguno de ellos lo es.

A continuación, se detallan los criterios que deberán seguir las variantes para cada tipo de herencia:

Autosómica dominante

-Si ninguno de los progenitores es afecto: buscaremos variantes *de novo* en el Caso Índice.

-Si uno de los padres es afecto: buscaremos variantes en heterocigosis que también estén en el progenitor afecto.

#Autosómica recesiva:

-Si ningún padre es afecto: buscaremos variantes en homocigosis en el hijo que se encuentren en heterocigosis en los dos progenitores.

-Si uno de los padres es afecto, buscaremos variantes en homocigosis en hijo que se encuentren también en homocigosis en el progenitor afecto y en heterocigosis en el no afecto.

#Ligada al X

-Buscaremos variantes en el cromosoma X que también se encuentren en la madre en hetero.

Los filtros que hemos establecido para cuando se desea filtrar por la naturaleza de la variante en si son los siguientes:

- Frecuencia Alélica: Para realizar este filtraje usaremos los datos de provenientes del Consorcio de Agregación de Exomas (ExAC, en sus siglas en inglés). Los resultados de este describen el análisis de los exomas de más de 60.000 individuos de diversas poblaciones [24].

- Significado Clínico: Así el usuario podrá filtrar las variantes en función de la anotación de esta según ClinVar, que, como ya hemos explicado, clasifica las variantes en función de la capacidad de estas de causar enfermedad o no (Significado clínico).

- Impacto de la variante:

2.1.4 Implementación de la aplicación web:

En la ingeniería de software se denomina aplicación web a aquellas herramientas que los usuarios pueden utilizar accediendo a un servidor web a través de internet o de una intranet mediante un navegador [25].

2.1.4.1 Lenguaje R:

En la realización de la aplicación hemos usado el lenguaje de programación R. Este es un lenguaje dirigido principalmente al análisis estadístico pero que contiene en su repositorio una gran numero de paquetes que nos van a ser útiles tanto en la creación de la interfaz de la aplicación como el la lectura y tratamiento de los datos a estudiar. A continuación, describiremos los paquetes que hemos usado en la implementación de la aplicación.

2.1.4.1.1 Paquete “Shiny”:

Las Aplicaciones Web Shiny funcionan de la misma forma que una Aplicación Web hecha en otras tecnologías, con la ventaja de que R y el paquete Shiny se encargan de generar todo el código necesario para facilitar la creación de una Aplicación Web, sin necesidad de entender en detalle el funcionamiento de las tecnologías Web. Shiny ofrece muchas funciones que generan código HTML para mostrar elementos en una página Web y al mismo tiempo ofrece un modelo de interacción con los componentes de la página para hacerla interactiva y dinámica.

Una aplicación Shiny está conformada por un archivo app.R o dos archivos ui.R y server.R. Es decir, se puede hacer una aplicación Shiny partiendo de un solo archivo con todo el código o se puede partir de dos archivos que separan los aspectos de la interfaz de los aspectos centralizados (servidor).

- ui.R es el archivo donde se especifican los elementos de la interfaz y la disposición de dichos elementos en la pantalla.

- server.R es el archivo donde se programa la lógica del servidor y se genera el contenido dinámico que depende de las interacciones con la pantalla.

Shiny se encarga de generar el HTML/JavaScript y además ofrece toda la lógica de manejo de los eventos que se producen en pantalla (clicks a botones, cambios de menú, etc), cuando cambia un parámetro en pantalla, Shiny también se encarga de comunicarse con el servidor enviando el nuevo valor del

parámetro y el servidor procesa la llamada, genera el resultado y lo envía al cliente. Toda esta lógica la ofrece Shiny y el analista sólo se debe enfocar en lo que va a aparecer en pantalla y lo que se debe generar cada vez que el usuario hace algún cambio. El analista se concentra en programar en R para generar su Aplicación Web interactiva y dinámica [26].

2.1.4.1.2 Paquete “vcfR”:

Se trata de un conjunto de herramientas diseñadas para leer, escribir, manipular y analizar datos en formato VCF.

Una vez que hemos leído los archivos en formato VCF, anotados previamente, mediante la función “read.vcfR()” se genera un objeto de R. Este es un objeto de clase “vcfR” con tres ranuras, cada una de ellas contiene el contenido de las distintas partes del fichero VCF que hemos descrito en el capítulo “Anotación de los datos”, es decir; de la Metainformación (“meta”), de las 8 primeras columnas fijas (“fix”) y del resto de columnas variables (“gt”) (Fig.3)

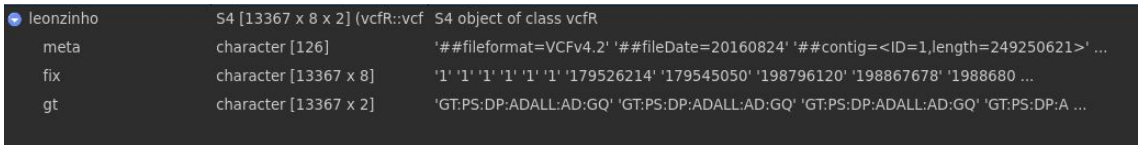


Fig 14. Ejemplo de objeto vcfR. Se muestra la información de las tres ranuras del objeto.

Mediante la manipulación de estas tres ranuras podemos acceder a los datos de cada una de las variantes generando una tabla sobre la cual posteriormente aplicaremos los criterios de priorización y filtraje.

2.1.4.1.3 Paquete “dplyr”:

El paquete “dplyr” es una versión optimizada del paquete “plyr”. El paquete “dplyr” no proporciona ninguna nueva funcionalidad a R per se, en el sentido que todo aquello que podemos hacer con “dplyr” lo podríamos hacer con la sintaxis básica de R [27].

La principal contribución del paquete “dplyr” es que proporciona una "gramática" sencilla para la manipulación de marcos de datos.

2.1.5 Partes y uso de la aplicación:

La aplicación está compuesta por cuatro pestañas (“Portada”, “Ayuda”, “Análisis de datos” y “Resultados”), a continuación, se detalla cada una de ellas.

2.1.5.1 Portada:

Es la primera de las pestañas y la que se muestra al ejecutar la aplicación. En ella encontramos información sobre la aplicación como el nombre, el desarrollador, etc. (Fig.4).

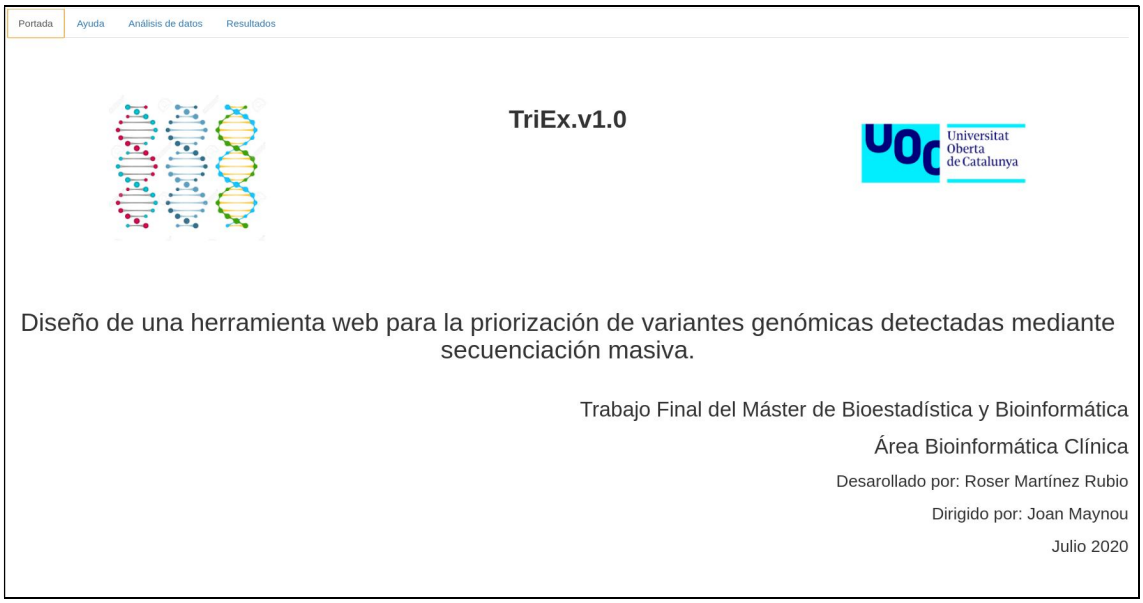


Fig 15. Portada.

2.1.5.2 Pestaña de ayuda:

Se trata de la segunda pestaña de la aplicación y contiene unas breves instrucciones sobre el uso de la aplicación (Fig.5).

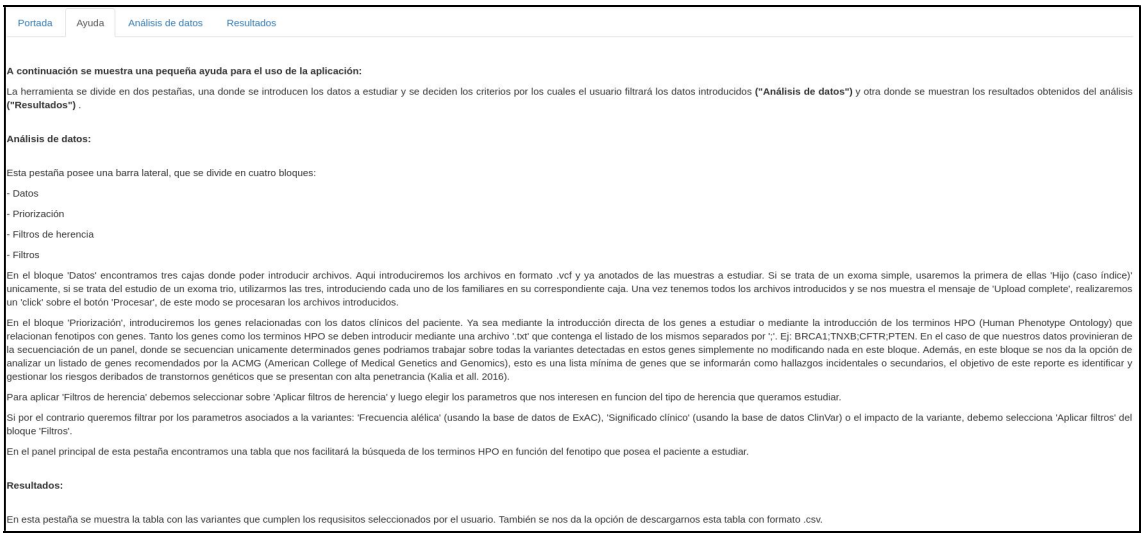


Fig 16. Pestaña de ayuda.

2.1.5.3 Análisis de datos:

A través de esta pestaña el usuario introducirá los datos que desee analizar y seleccionará los parámetros para la priorización y filtraje de las variantes (Fig 6).

Fig 17. Visión general de la pestaña "Análisis de datos".

Esta pestaña está compuesta de una barra lateral y un panel principal.

La barra lateral, a su vez, está dividida en cuatro bloques:

- Datos

Para la introducción de los datos hemos habilitado tres cajas donde se pueden introducir archivos (Fig 7). La primera de ellas se usará para el individuo a estudiar cuando se trata de un único caso y para el caso índice, cuando lo que se desea es realizar un estudio “trio”, donde se estudia a un individuo probando y a sus dos progenitores a la vez. El resto de los progenitores se introducirán en las siguientes cajas, tal y como se indica. En este bloque, también encontramos el botón “Procesar”, el cual deberemos clicar una vez los archivos se hayan cargado en nuestra aplicación. Esta acción permitirá que la parte del servidor de la aplicación lea los datos del archivo VCF y los transforme en una tabla sobre la cual trabajará posteriormente.

Fig 18. Cajas para introducir los datos.

- Priorización

Si seguimos bajando en la barra lateral, encontramos el bloque de Priorización. En este el usuario deberá elegir de qué forma realizará esta acción, es decir, si mediante la entrada de términos HPO o la entrada de genes. En ambos casos, el usuario deberá ingresar un archivo con formato .txt

que contenga una lista o bien de los términos HPO o bien de los genes que quiera analizar, separados mediante ";". En función de que es lo que desea estudiar deberá introducir el archivo en la caja correspondiente (Fig 8).

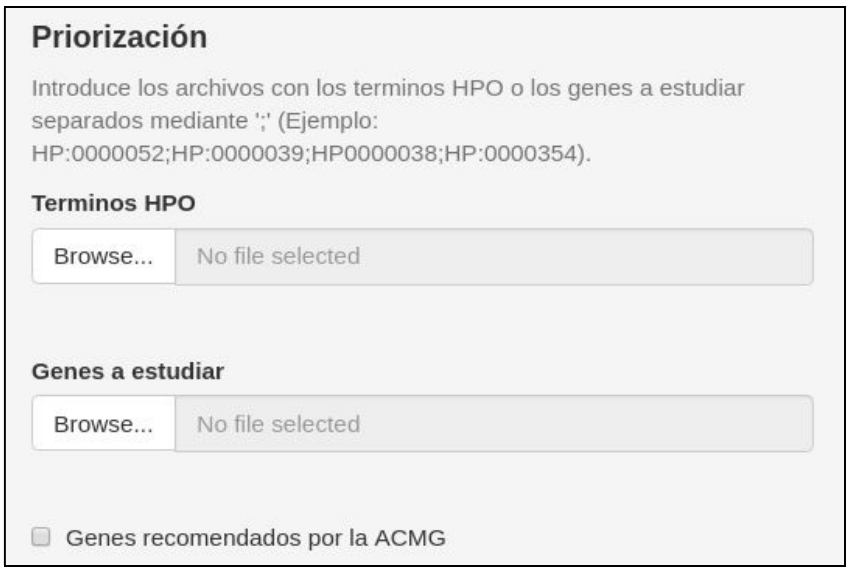


Fig 19. Bloque de Priorización.

Al final de este bloque, encontramos una casilla de verificación. Esta casilla deberá ser seleccionada por el usuario si desea analizar los 59 genes recomendados por la ACMG, los cuales ya hemos explicado en el capítulo "2.1.3.1 Criterios de priorización".

- Filtros de herencia

El siguiente bloque que nos encontramos es el de filtros de herencia (Fig 9). El usuario deberá validar la casilla de verificación si desea realizar este tipo de filtraje. Como es de esperar, sólo se puede aplicar en el caso de que se trate de el estudio de un trio. Los parámetros para seleccionar por el usuario serán: 1.- Tipo de herencia; es decir, si desea buscar variantes que sigan un patrón de herencia autosómico dominante, recesiva o ligado al cromosoma X. 2.- Progenitor afecto; en el caso de que alguno de los dos progenitores sea afecto de la misma patología que el caso índice. 3.- Sexo del caso índice; que se tendrá en cuenta cuando se realice el estudio de enfermedad ligada al cromosoma X.

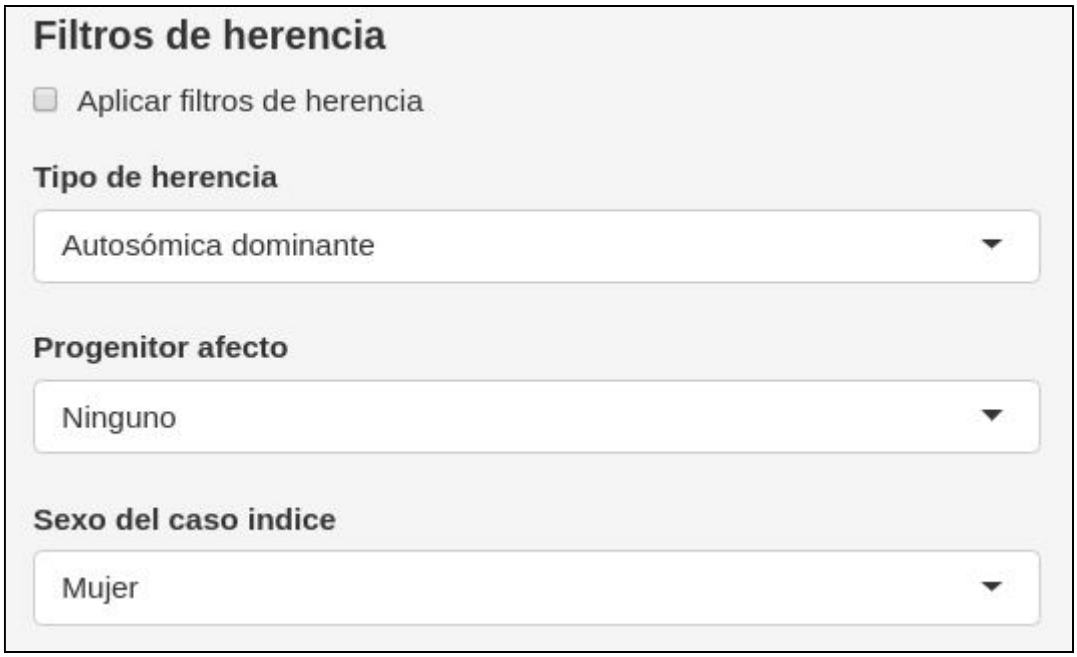


Fig 20. Bloque de Filtros de herencia.

- Filtros

Se trata del último bloque que encontramos en la barra lateral (Fig. 10). Al igual que ocurre con los filtros de herencia, el usuario debe validar la casilla de verificación para que se apliquen estos tipos de filtro. Este tipo de filtros podrán ser aplicados tanto para el estudio individual como para el estudio de un trio. Y los parámetros que podrá elegir el usuario serán: 1.- Frecuencia Alélica; basado en las frecuencias anotadas en ExAC. 2.- Significado clínico; basado en la anotación de la base de datos ClinVar. 3.- Impacto de la variante; que puede ser: Alta, Moderado y Modificador.

Por último, encontramos el botón “Priorizar”, este debe clicarse cuando el usuario ya ha establecido los criterios de Priorización y Filtraje. En el momento en que el usuario clica sobre este, la herramienta procesa la información que el usuario ha introducido en el formulario y devuelve la tabla con los resultados (esta tabla se explicará más adelante).

Fig 21. Bloque de Filtros.

- Panel principal

En el panel principal de esta pestaña, encontramos una tabla donde se relacionan los términos HPO con los fenotipos. Así, el facultativo tendrá la posibilidad de realizar la búsqueda de los términos HPO que se corresponden con los fenotipos mostrados por su paciente de una forma sencilla (Fig. 11).

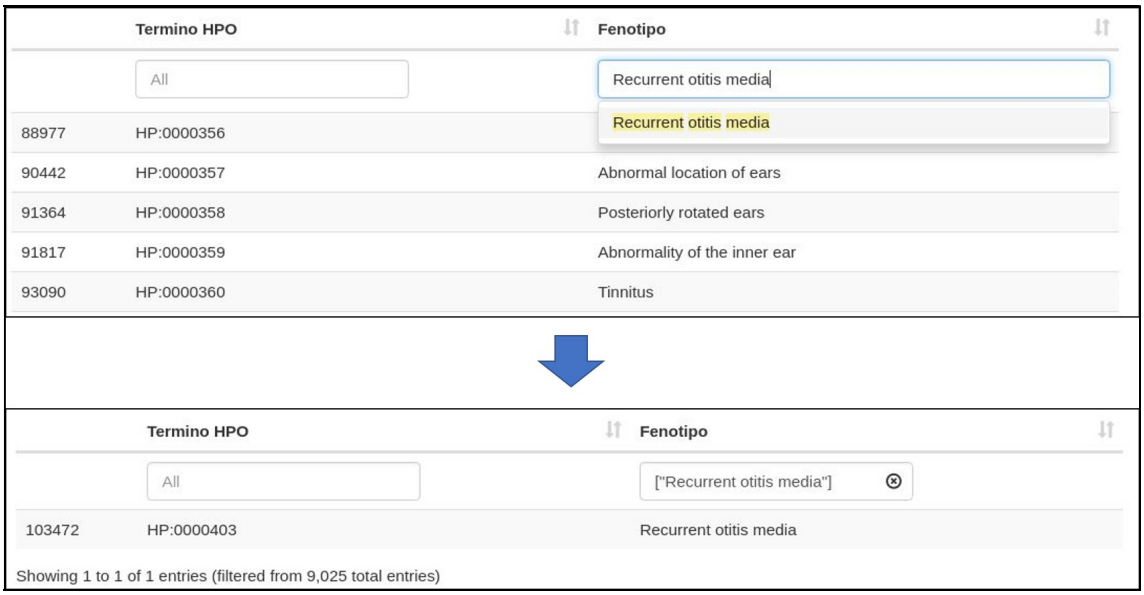


Fig 22. Tabla de términos HPO. Se muestra un ejemplo de cómo realizar la búsqueda de los términos en función del fenotipo mostrado por el paciente. En este caso, se ha realizado la búsqueda del fenotipo "Otitis media recurrente" y se observa que el termino correspondiente a ese fenotipo es HP:0000403.

2.1.5.4 Pestaña de resultados:

Es la última de las pestañas de la aplicación. En ella se muestra la tabla con los resultados obtenidos (Fig 12).

Los campos de la tabla son los siguientes:

1. **CHROM:** Cromosoma donde se encuentra la variante.
2. **POS:** Posición dentro del genoma (GRCh37).
3. **REF:** Nucleótido de referencia en esa posición.
4. **ALT:** Nucleótido que encontramos en nuestro caso concreto.
5. **ID_rs:** Identificador de la variante en dbSNP
6. **Cigosidad:** Heterocigoto/Homocigoto
7. **Gen:** Gen donde se localiza la variante
8. **Exón:** Exón donde se localiza dentro del gen
9. **Transcrito:** Identificador del transcrito en ENSEMBL
10. **Cambio_cDNA:** Nombre del cambio nucleotídico en el transcrito anterior
11. **Cambio_Prot:** Nombre del cambio proteico
12. **Tipo:** Tipo de variante
13. **Impacto:** Impacto de la variante
14. **CLNSIG:** Significado clínico
15. **dbNSFP_ExAC_AF:** Frecuencia de la variante en ExAC.
16. **dbNSFP_1000Gp1_AF:** Frecuencia de la variante en el proyecto 1000G.

Detrás de la tabla encontramos dos enlaces. El primero de ellos es para descargarse la tabla que estamos observando, el segundo, para descargarse un archivo de texto con los parámetros elegidos por el usuario.

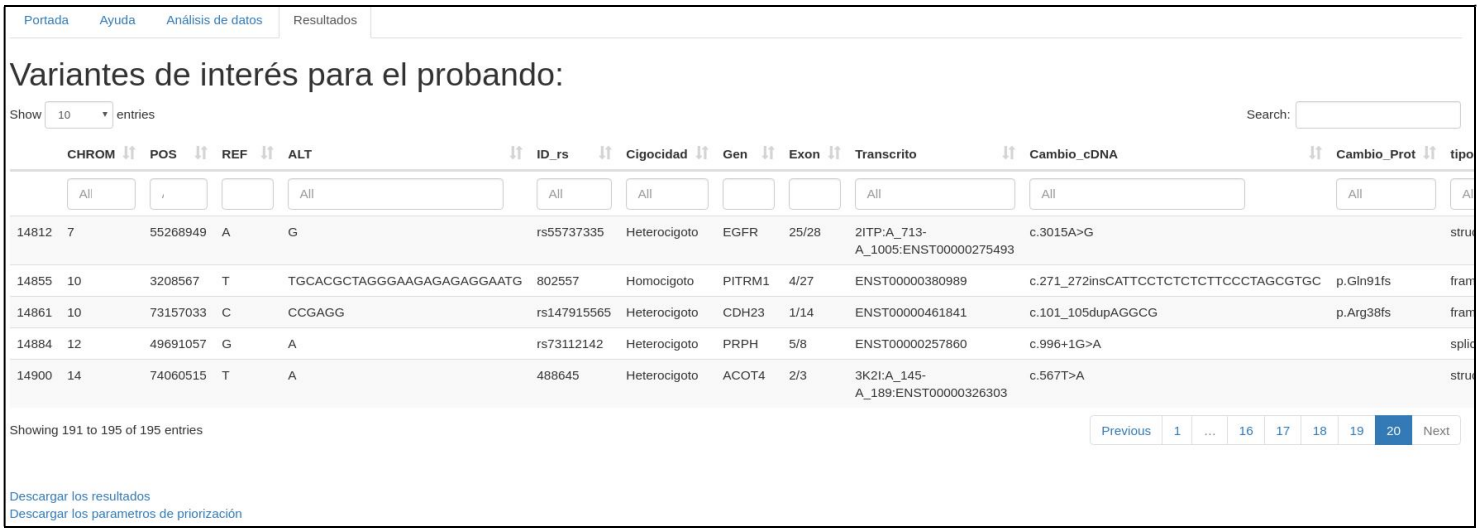
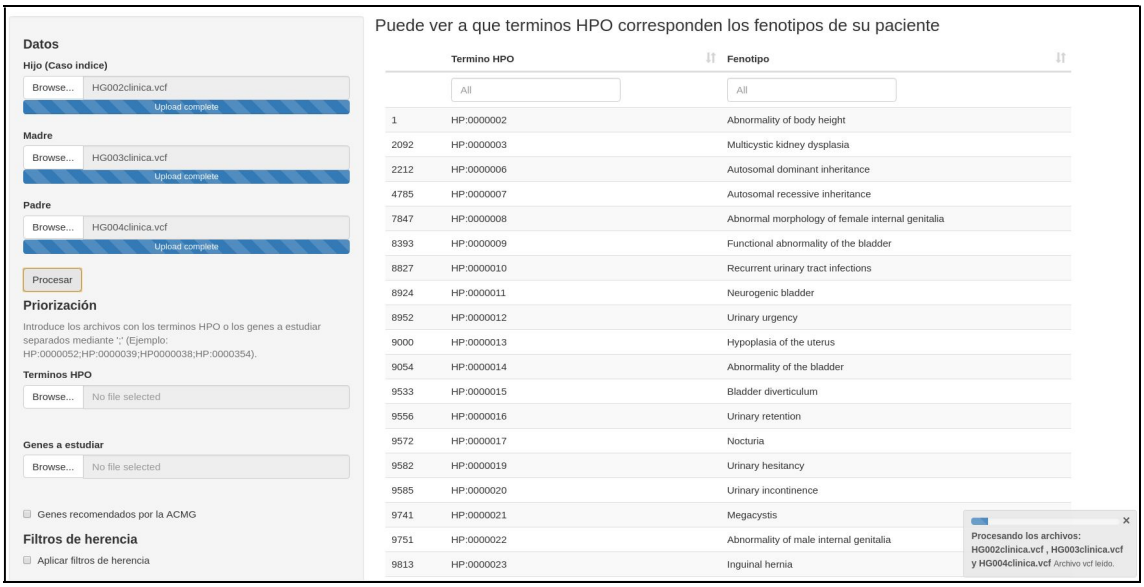


Fig 23. Visión general de la pestaña "Resultados". Este es un ejemplo donde se muestra una tabla obtenida de filtrar todas las variantas halladas en un estudio simple que tienen un impacto elevado. Se muestran solamente las 5 últimas, ya que se han obtenido un total de 191 variantes que cumplen este requisito. Debajo de la tabla se observan los dos enlaces (en azul) que nos permiten la descarga de los archivos con los resultados y con los criterios de priorización, respectivamente.

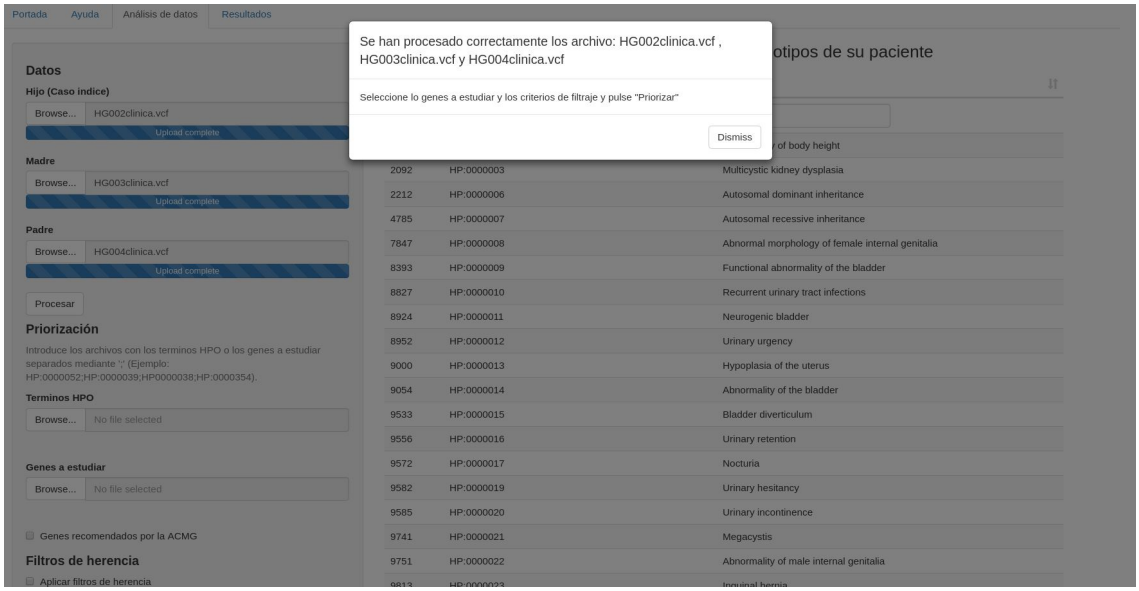
2.1.5.5 Ejemplo:

En este apartado vamos a realizar un ejemplo de cómo funciona la aplicación:

Introducimos los datos y cuando la descarga se haya completado, clicamos sobre "Procesar":

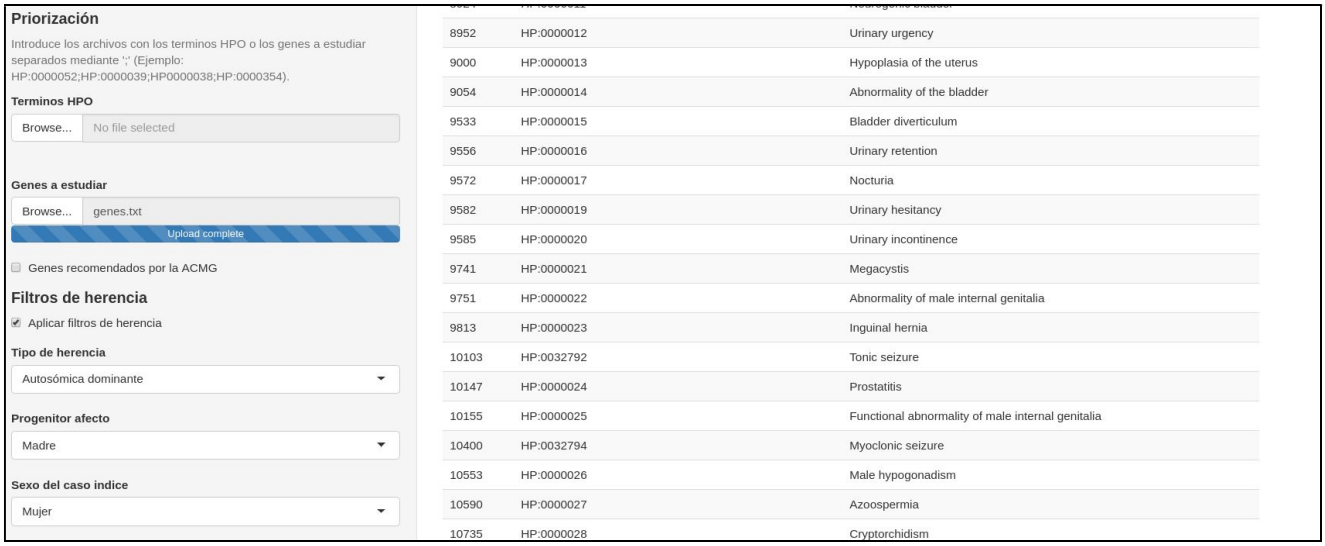


Cuando la aplicación termina de procesar los archivos VCF, aparece una ventana emergente avisando al usuario que este proceso se ha realizado satisfactoriamente y que debe introducir los parámetros de priorización.

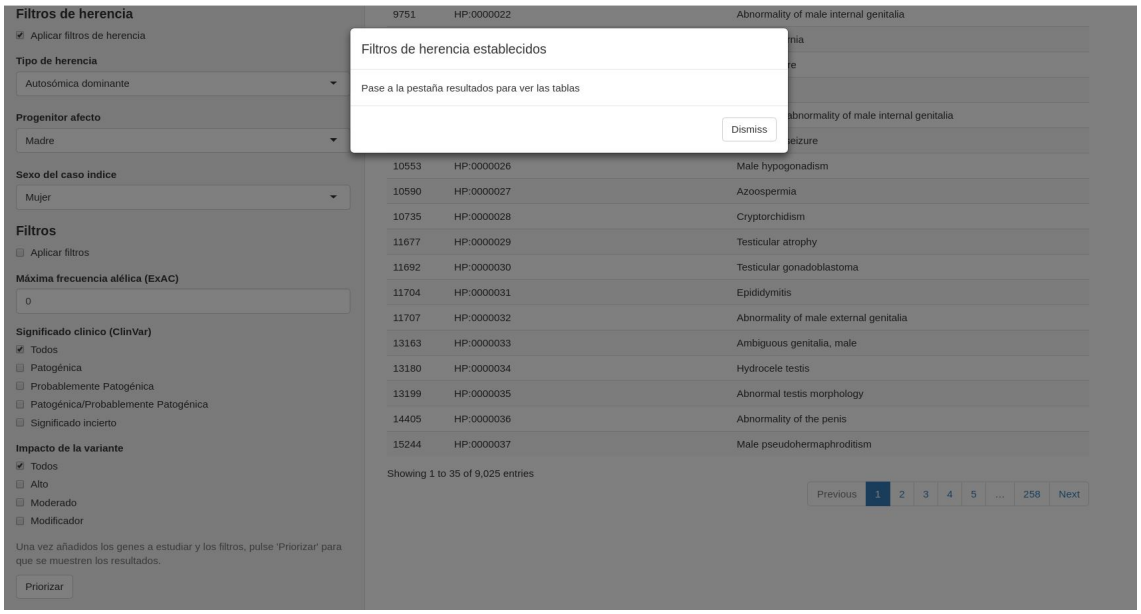


Introducimos los parámetros deseados, en este ejemplo, introduciremos en la caja habilitada para introducir archivos con un listado de genes, el archivo “genes.txt” que contiene una lista con los genes que deseamos estudiar, estos son:

Como criterios de filtraje, usaremos los filtros de herencia, y dentro de estos elegiremos tipo de herencia autosómico dominante y que la madre es afecta.



Una vez tenemos establecidos los criterios de priorización clicamos sobre el botón “Priorizar” y nos aparece otra ventana emergente.



En este momento, debemos pasar a la cuarta ventana de “Resultados” para poder observar la tabla.

Variantes en heterocigosis, tanto en la madre como en el hijo, que se encuentran en genes con patrón de herencia Autosómico Dominante:

Show 10 entries — Search:

	CHROM	POS	REF	ALT	ID_rs	Cigocidad	Gen	Exon	Transcrito	Cambio_cDNA	Cambio_Prot	tipo	Impacto	CLNSIG
1	1	55517940	G	A	rs11800231	Heterocigoto	PCSK9	3/11	ENST00000302118	c.524-11G>A		intron_variant	MODIFIER	Benign/Like
2	1	55518093	G	A	rs11800243	Heterocigoto	PCSK9	4/11	ENST00000302118	c.657+9G>A		intron_variant	MODIFIER	Conflicting_
3	1	55518160	C	A	rs11806638	Heterocigoto	PCSK9	4/11	ENST00000302118	c.657+76C>A		intron_variant	MODIFIER	Benign
4	1	55518467	A	G	rs2495477	Heterocigoto	PCSK9	5/11	ENST00000302118	c.799+3A>G		splice_region_variant&intron_variant	LOW	Benign/Like
5	1	55525399	C	T	rs45439391	Heterocigoto	PCSK9	10/11	ENST00000302118	c.1681+63C>T		intron_variant	MODIFIER	Benign
6	1	55525400	G	A	rs483462	Heterocigoto	PCSK9	10/11	ENST00000302118	c.1681+64G>A		intron_variant	MODIFIER	Benign
7	1	156105417	A	G	rs76017998	Heterocigoto	LMNA		ENST00000498722	n.-107A>G		upstream_gene_variant	MODIFIER	Benign
8	1	201037962	T	C	rs1272740	Heterocigoto	CACNA1S	19/43	ENST00000362061	c.2550+303A>G		intron_variant	MODIFIER	Benign
9	1	201046388	A	G	rs6672094	Heterocigoto	CACNA1S	11/43	ENST00000362061	c.1620-133T>C		intron_variant	MODIFIER	Benign
10	1	201047062	G	A	rs4915476	Heterocigoto	CACNA1S	11/44	ENST00000362061	c.1564C>T	p.Leu522Leu	synonymous_variant	LOW	Benign

Showing 1 to 10 of 189 entries — Previous 1 2 3 4 5 … 19 Next

Descargar los resultados
Descargar los parametros de priorización

El encabezamiento de la tabla nos indica que criterios deben seguir las variantes que se muestran en la tabla.

Por último, si el usuario quisiera descargarse la tabla con los resultados obtenidos en formato CSV o los criterios de priorización debería clicar sobre los enlaces de descarga y guardar los archivos en el directorio deseado.

3. Conclusiones

En cada paso que he dado para llegar al objetivo final, crear una herramienta capaz de priorizar variantes provenientes de la secuenciación masiva del genoma de un individuo enfermo, he ido aprendiendo distintas cosas. Primero, como obtener información de la paginas web de las distintas bases de datos para poder trabajar con ellas y la información de los genomas de los pacientes a estudiar. El segundo punto ha sido aprender a usar un programa de anotación como es snpEff mediante el uso de comandos en la terminal de Linux. El tercer punto clave en la realización de este TFM ha sido la creación de la aplicación mediante el paquete “shiny” de R, del cual, hasta la fecha, desconocía de su existencia y por tanto el aprendizaje ha sido desde cero. Por último, también he aprendido a usar el paquete “dplyr” de R para la manipulación de marcos de datos.

Los objetivos planteados al inicio de este trabajo han sido alcanzados, tanto el general como los específicos. La aplicación ha sido desarrollada y funciona. Y los parámetros de priorización y filtraje han sido establecidos.

Si bien, a pesar de que la metodología usada si ha sido la propuesta al inicio del TFM y ha sido la adecuada para alcanzar el objetivo, la planificación ha sufrido alguna alteración. Debido, en parte, a fallos en la estimación del tiempo requerido en unos de los primeros pasos, que era la familiarización con toda esta metodología a usar. Creo que me ha costado un poco entender cómo funcionaba la reactividad de “shiny”.

Como líneas de trabajo futuro, mejoraría, en primer lugar, la anotación de las variantes. Por ejemplo, usaría para anotar la frecuencia alélica la base de datos de GenomAD, que es más amplia que la que he usado en este trabajo (ExAC) y esta basada en un mayor número de genomas y exomas estudiados. En la anotación, también añadiría los transcritos con identificador NM_ del NCBI, ya que su uso esta mucho más expandido que los que he usado de Ensembl.

Respecto a la aplicación, los cambios o mejoras que realizaría serian: En primer lugar, implementar que la anotación se pudiera realizar a través de la propia aplicación. En la pestaña de “Análisis de datos”, permitiría que también

se pudieran introducir genes o términos HPO de forma manual, de este modo el facultativo podría ir buscando los términos HPO en la tabla que relaciona estos con el fenotipo e ir introduciéndolos en el momento. Por último, en esta pestaña permitir que se pueda realizar el filtro de herencia a la vez que los otros tipos de filtros.

En general la sensación es buena, he conseguido alcanzar el objetivo real de este TFM y del máster qu

e era el de adquirir nuevos conocimientos y habilidades que me puedan ser útiles en otros ámbitos.

4. Glosario

pipeline: conjunto de comandos que permiten analizar los datos y obtener los resultados

Variant-calling: proceso por el cual se obtienen las variantes alélicas propias de una muestra determinada en comparación con el genoma de referencia

Alineamiento: proceso por el cual los reads son mapeados en un genoma de referencia

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

Anotación de variantes: proceso por el que se identifican el posible impacto funcional de una variante en un gen, usando normalmente diferentes bases de datos

Cigosidad: es el grado de similitud de los alelos para un rasgo genético en un organismo.

Fenotipo: expresión del genotipo en función del ambiente.

Homocigoto: Un organismo es homocigótico respecto a un gen cuando los dos alelos codifican la misma información para un carácter

Heterocigoto: Un organismo es heterocigótico respecto a un gen cuando los dos alelos codifican distinta información para un carácter

Variante genética: cambio en la secuencia de nucleótidos respecto aun genoma de referencia.

SNP (single nucleotide polimorfism): Son variantes no asociados a patología que confieren la variabilidad poblacional.

5. Bibliografía

1. OMIM. https://omim.org/.
2. Santillán-Garzón, S. *et al.* Diagnóstico Molecular De Enfermedades Genéticas: Molecular Diagnosis of Genetic Diseases: From Genetic To Genomic Diagnosis Using Next Generation Sequencing. *Rev. Clínica Las Condes* **26**, 458–469 (2015).
3. Bertoldi, L. *et al.* QueryOR: A comprehensive web platform for genetic variant analysis and prioritization. *BMC Bioinformatics* **18**, 1–11 (2017).
4. Yohe, S. & Thyagarajan, B. Review of clinical next-generation sequencing. *Archives of Pathology and Laboratory Medicine* vol. 141 (2017).
5. Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. Clinical sequencing: is WGS the better WES? *Hum. Genet.* **135**, 359–362 (2016).
6. Meena, N., Mathur, P., Medicherla, K. & Suravajhala, P. A Bioinformatics Pipeline for Whole Exome Sequencing: Overview of the Processing and Steps from Raw Data to Downstream Analysis. *Bio-Protocol* **8**, 1–15 (2018).
7. Ip, E., Chapman, G., Winlaw, D., Dunwoodie, S. L. & Giannoulatou, E. VPOT: A Customizable Variant Prioritization Ordering Tool for Annotated Variants. *Genomics, Proteomics Bioinforma.* (2019) doi:10.1016/j.gpb.2019.11.001.
8. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* (2015) doi:10.1038/nprot.2015.105.
9. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *bioRxiv* (2016) doi:10.1101/042374.
10. Alemán, A., Garcia-Garcia, F., Salavert, F., Medina, I. & Dopazo, J. A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gku407.
11. Salatino, S. & Ramraj, V. BrowseVCF: a web-based application and workflow to quickly prioritize disease-causative variants in VCF files. *Brief.*

Bioinform. (2017) doi:10.1093/bib/bbw054.

12. Alexander, J., Mantzaris, D., Georgitsi, M., Drineas, P. & Paschou, P. Variant Ranker: A web-tool to rank genomic data according to functional significance. *BMC Bioinformatics* (2017) doi:10.1186/s12859-017-1752-3.
13. Antanaviciute, A. *et al.* OVA: Integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btv473.
14. Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.* (2014) doi:10.1126/scitranslmed.3009262.
15. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* (2014) doi:10.1038/nbt.2835.
16. Landrum, M. J. *et al.* ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, (2016).
17. Landrum, M. J. *et al.* ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, (2018).
18. Landrum, M. J. & Kattman, B. L. ClinVar at five years: Delivering on the promise. *Hum. Mutat.* (2018) doi:10.1002/humu.23641.
19. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* (2013) doi:10.1002/humu.22376.
20. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* (2011) doi:10.1002/humu.21517.
21. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**, 80–92 (2012).
22. Trujillano, D. *et al.* Clinical exome sequencing: Results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* **25**, (2017).
23. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* (2017) doi:10.1038/gim.2016.190.

24. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* (2016) doi:10.1038/nature19057.
25. Aplicación web. Recuperado en Junio de 2020 de https://es.wikipedia.org/wiki/Aplicación_web.
26. Ramírez, A. A. Aplicaciones Web en R con Shiny. https://synergy.vision/corpus/shiny/2017-08-15-shiny.html (2017).
27. Peng, R. D. *R Programming for Data Science*. (2019).

6. Anexos

Listado de apartados que son demasiado extensos para incluir dentro de la memoria y tienen un carácter autocontienido (por ejemplo, manuales de usuario, manuales de instalación, etc.)

Dependiente del tipo de trabajo, es posible que no haya que añadir ningún anexo.