Universidad Europea

Máster en Bioinformática

DESARROLLO DE ALGORITMOS BIOINFORMÁTICOS PARA EL MANEJO AUTOMÁTICO DE HALLAZGOS SECUNDARIOS EN DATOS GENÓMICOS

Autor: Edurne Urrutia Lafuente

Tutor: Javier Pérez Florido

Curso 2022-23

AGRADECIMIENTOS

En primer lugar, quisiera agradecer a Javier Pérez Florido, por dedicarme su tiempo y paciencia infinita.

A David Gómez-Cabrero y Ana, consejeros en la sombra, siempre dispuestos a ayudar.

A Ana (sí, otra vez), Asier, Dani, Estefanía, Sara y Xabi. Porque habéis sido la mejor sorpresa de este año duro, gracias por sacarme una sonrisa cada día. ¡Con vosotros me atrevo a saltar de donde haga falta!

A Moni, porque los bizcochos del Culto sirven para celebrar y para consolar, y porque sé que como mujeres for(m)i(d)ables seguiremos haciéndolo.

A Leyre, Pablo, Adrián, Aida y David, porque a pesar de empezar como desconocidos y de la distancia, hemos estado muy presentes (¡cómo une el sufrimiento compartido!). Espero visita de vuelta.

A Aitor, Mai y Maite por estar ahí siempre, aunque me hagáis sudar la gota gorda.

Al Doc, porque en este camino con curvas y bifurcaciones hemos sabido reencontrarnos y acompañarnos una etapa más.

Y a mis aitas, por acompañarme y apoyarme siempre, tengo mucha suerte de teneros a mi lado.

Muchísimas gracias a todos por acompañarme en este camino de aprendizaje en tantos sentidos.

RESUMEN

Introducción: La secuenciación de nueva generación (NGS) es una herramienta fundamental en el estudio de enfermedades genéticas. Sus implicaciones van más allá de la enfermedad primaria, ya que permiten la identificación de hallazgos secundarios (de riesgo personal, reproductivo y farmacogenético), que pueden tener un impacto relevante en el manejo clínico del paciente. Por tanto, el desarrollo de algoritmos bioinformáticos dirigidos a su manejo es un paso importante hacia la implementación de la genómica en la práctica médica. Sin embargo, hasta ahora no se han descrito herramientas para el manejo específico de hallazgos secundarios.

Objetivos: El objetivo de este Trabajo Final de Máster consiste en el desarrollo de una herramienta bioinformática destinada al manejo automático de hallazgos secundarios en datos genómicos.

Material y métodos: La herramienta, desarrollada en Python y operable desde la línea de comandos, procesa archivos de variantes genómicas en formato *Variant Calling Format* (VCF) y permite seleccionar las categorías de hallazgos secundarios mencionadas. Además, se pueden configurar parámetros como el ensamblaje de referencia, el modo de ejecución, el nivel de evidencia para las interpretaciones y un archivo de términos de la ontología del fenotipo humano (HPO). La herramienta se apoya en InterVar y la base de datos ClinVar, y el flujo de trabajo incluye la preparación de la herramienta, la normalización del archivo VCF, la extracción de variantes genómicas en genes específicos de cada categoría mediante la intersección con archivos *Browser Extensible Data* (BED) y la ejecución de los módulos propios de cada categoría para la anotación e interpretación de variantes patogénicas o probablemente patogénicas. El último paso consiste en la generación de informes en formato Excel, separando los hallazgos por categorías. Además, la herramienta se validó utilizando datos genómicos con hallazgos secundarios previamente identificados en el proyecto NAGEN1000.

Resultados: La validación de la herramienta demostró su capacidad para identificar hallazgos secundarios en datos genómicos del proyecto NAGEN1000, respaldando su eficacia en las tres categorías de hallazgos secundarios.

Conclusiones: El desarrollo de esta herramienta representa un avance en la medicina genómica y la atención médica personalizada. Las limitaciones actuales identificadas se completan con líneas de mejora futuras, como la inclusión en un *docker*.

Palabras clave: secuenciación de nueva generación, hallazgos secundarios, medicina genómica, algoritmos bioinformáticos, Python

Índice general

Índice de figuras

Índice de tablas

Acrónimos/Abreviaturas

ACMG *American College of Medical Genetics and Genomics*

ADN ácido desoxirribonucleico

API interfaz de programación de aplicaciones

BED *Browser Extensible Data*

CPIC Consorcio para la Implementación de la Farmacogenética Clínica

CYP2C9 *Cytochrome P450 family 2 subfamily C member 9*

CYP2C19 *Cytochrome P450 family 2 subfamily C member 19*

DPYD *Dihydropyrimidine dehydrogenase*

ESHG Sociedad Europea de Genética Humana

FG riesgo farmacogenético

GPAP *Genome-Phenome Analysis Platform*

HPO Ontología de Fenotipos Humanos

JSON *JavaScript Object Notation*

NGS *Next Generation Sequencing*

NUDT15 *Nudix hydrolase 15*

PharmGKB Base de Conocimiento Farmacogenético

RP riesgo personal

RR riesgo reproductivo

SNV *Single Nucleotide Variant*

TPMT *Thiopurine S-methyltransferase*

TSV *Tab-Separated Values*

VCF *Variant Calling Format*

1. Introducción

1.1 Secuenciación de nueva generación y utilidad en diagnóstico clínico

En las últimas décadas, los avances tecnológicos en secuenciación de nueva generación (NGS; *Next Generation Sequencing*) revolucionaron el campo de la genética clínica y la medicina de precisión. La NGS superó las limitaciones de la secuenciación Sanger tradicional al permitir la obtención de datos genómicos a gran escala de manera más rápida, precisa y asequible. Esta tecnología facilitó nuevas perspectivas para el análisis integral de la información genética y permitió el estudio exhaustivo de la variabilidad genómica, revelando aspectos cruciales de la genética humana y su relación con la salud y las enfermedades (Pereira *et al.*, 2020; Qin, 2019).

La NGS ha demostrado ser una herramienta muy útil para el diagnóstico clínico de enfermedades genéticas y trastornos hereditarios a partir de datos genómicos, especialmente en enfermedades raras (Boycott *et al.*, 2013). La capacidad de secuenciar rápidamente el ácido desoxirribonucleico (ADN) de un individuo ha permitido la identificación de mutaciones genéticas responsables de diversas patologías, incluso en casos en los que el diagnóstico clínico convencional resultaba difícil o inconcluso. Asimismo, la NGS ha facilitado la identificación de variantes genéticas asociadas con un mayor riesgo de desarrollar ciertas enfermedades, lo que ha impulsado la medicina de precisión y el desarrollo de terapias personalizadas (Carrasco-Ramiro *et al.*, 2017; Rizzo & Buck, 2012).

Un ejemplo de la utilidad de la NGS en el diagnóstico clínico es el programa NAGEN, que tiene como objetivo la implementación de la secuenciación de genoma completo y el desarrollo de la medicina personalizada en la práctica clínica del sistema sanitario navarro. Mediante la aplicación de esta estrategia en más de 700 pacientes con condiciones de origen genético desconocidas, se logró un éxito diagnóstico del 33%. Además, el 2% de los participantes resultaron portadores de variantes de predisposición a enfermedades graves, el 4% presentó riesgos reproductivos y en el 100% de los casos se identificaron variantes farmacogenómicas accionables. Este enfoque demostró ser coste-efectivo y recibió una gran aceptación por parte de los pacientes, lo que subraya la utilidad de la NGS en el diagnóstico clínico y la medicina de precisión (Pasalodos *et al.*, 2020).

En este contexto, un factor clave que ha impulsado la adopción generalizada de la NGS en el diagnóstico clínico es el continuo abaratamiento de los costes asociados con la secuenciación del genoma y del exoma (Figura 1). Durante los últimos años, los avances tecnológicos y la optimización de los protocolos de secuenciación han contribuido a una reducción significativa en los costes de producción, de manera que se ha vuelto mucho más accesible y rentable para un mayor número de instituciones (van Dijk *et al.*, 2014).

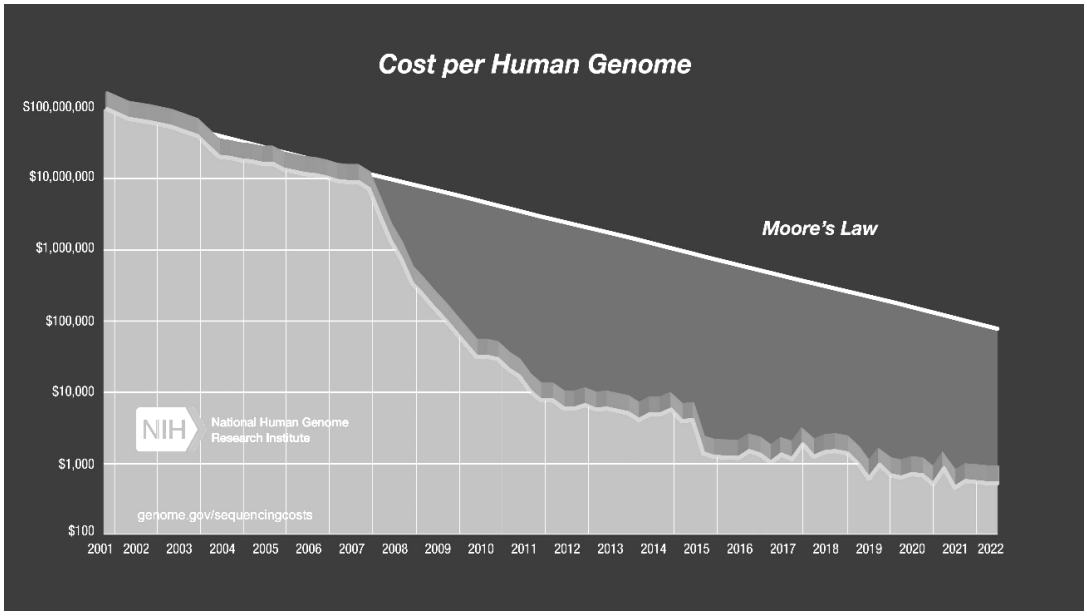


Figura 1 | Evolución a lo largo del tiempo del coste de secuenciación del genoma humano. El eje X muestra los años desde 2001 hasta 2022, y el eje Y indica el coste en dólares, en escala logarítmica. Se muestran también los datos hipotéticos que reflejan la Ley de Moore, que describe la tendencia a duplicar la potencia de cálculo cada dos años (NIH, 2023).

La disponibilidad de secuenciación de genoma y exoma a precios más bajos ha llevado a una mayor inclusión de estas tecnologías en el ámbito del diagnóstico clínico. En lugar de centrarse únicamente en análisis genéticos específicos y dirigidos, los profesionales de la salud pueden ahora aprovechar la información completa del genoma o exoma para obtener una visión más completa y detallada del perfil genético de un individuo (Di Resta *et al.*, 2018). El uso cada vez más frecuente del genoma y el exoma en el diagnóstico clínico, sobre todo en enfermedades raras, ha permitido, a su vez, la detección de un mayor número de variantes genéticas asociadas con enfermedades, lo que ha llevado a un aumento significativo en la comprensión de las bases genéticas de diversas condiciones médicas (Boycott *et al.*, 2013). Asimismo, el acceso a información genómica más completa ha facilitado la identificación de otros hallazgos no relacionados con el

propósito inicial de la secuenciación, los denominados hallazgos secundarios u otras variantes genómicas de significado clínico relevante. Todo ello influye de forma inequívoca en el manejo médico y en el asesoramiento genético de los pacientes (Horton & Lucassen, 2019; Katz et al., 2020; Rizzo & Buck, 2012).

En el siguiente apartado, se analizarán en detalle los hallazgos secundarios en datos genómicos y cómo el desarrollo de algoritmos bioinformáticos permitirá un manejo automático y eficiente de esta información, lo que potencialmente mejorará la interpretación clínica y el abordaje de enfermedades genéticas en el contexto del diagnóstico clínico.

1.2 Cribado oportunista: hallazgos secundarios

Como se ha mencionado en el apartado anterior, durante el análisis de la secuenciación genómica del paciente pueden identificarse diferentes tipos de hallazgos o variantes genéticas atendiendo a su relación con la clínica del paciente.

Los hallazgos primarios se refieren a las variantes directamente relacionadas con la indicación principal de la prueba, es decir, aquellas que se buscan inicialmente para confirmar o diagnosticar una enfermedad específica.

Por otro lado, también pueden encontrarse variantes no relacionadas con la indicación principal de la prueba, pero que podrían tener consecuencias significativas para la salud, como causar enfermedades, influir en las decisiones reproductivas del paciente o incluso personalizar su tratamiento. La identificación de estas variantes puede ocurrir de forma intencionada o no, de forma que se habla de hallazgos secundarios en el primer caso, y de hallazgos incidentales en el segundo caso (de Wert *et al.*, 2021).

Es importante tener en cuenta que tanto los hallazgos secundarios como los incidentales pueden ser médicamente accionables, lo que implica que podrían tener un impacto relevante en el manejo clínico del paciente. Para abordar esta situación se han propuesto diversas orientaciones y recomendaciones por parte de organismos como el *American College of Medical Genetics and Genomics* (ACMG), ClinGen o la Sociedad Europea de Genética Humana (ESHG), entre otros (Austin-Tse *et al.*, 2022). El ACMG, en

particular, ha publicado una declaración de principios en secuenciación clínica que destaca la importancia de informar a los pacientes sobre la posibilidad de obtener este tipo de resultados, y ofrece recomendaciones para el manejo responsable de los hallazgos secundarios (Green *et al.*, 2013).

En función de sus implicaciones para la salud, estos hallazgos secundarios pueden clasificarse en tres categorías principales:

- **Hallazgos secundarios de riesgo personal**: se refieren a la identificación de variantes genéticas asociadas con un mayor riesgo de desarrollar ciertas enfermedades o afecciones. La detección temprana de estos riesgos permite la adopción de medidas preventivas, de cambios en el estilo de vida y un seguimiento médico más cercano para mitigar los efectos de estas afecciones (Miller *et al.*, 2023).
- **Hallazgos secundarios de riesgo reproductivo y riesgo en la descendencia:** se centran en la identificación de variantes genéticas con posibles implicaciones para la salud reproductiva de una pareja, tanto por suponer un riesgo en la descendencia, como por estar relacionadas con la fertilidad, abortos recurrentes o complicaciones durante el embarazo. En definitiva, los hallazgos de riesgo reproductivo pueden ser de gran importancia para la planificación familiar y la atención médica prenatal, ya que permiten a las parejas tomar decisiones informadas sobre la posibilidad de tener descendientes con riesgo de desarrollar ciertas enfermedades genéticas (Gregg *et al.*, 2021).
- **Hallazgos secundarios de riesgo farmacogenético**: estos hallazgos se centran en la identificación de variantes genéticas que influyen en la respuesta individual a ciertos medicamentos. Estas variantes pueden afectar a la farmacocinética (la absorción, distribución, metabolización o eliminación del fármaco), así como a la farmacodinamia del fármaco (modificando su diana o alterando las rutas biológicas que determinan la sensibilidad del paciente a sus efectos farmacológicos). Esto puede resultar en diferencias en la eficacia del tratamiento, así como en la tolerancia y los efectos secundarios de los medicamentos. La identificación de estas variantes es esencial para la medicina personalizada, ya que permite una selección más precisa de los medicamentos y dosis adecuadas para cada paciente, mejorando

así la eficacia del tratamiento y reduciendo el riesgo de reacciones adversas (Evans, 2015).

1.3 Software para el manejo automático de hallazgos secundarios

Como ya se ha comentado, la NGS ha emergido como una herramienta fundamental en el estudio de enfermedades genéticas, especialmente en enfermedades raras. Sin embargo, los datos obtenidos mediante NGS tienen implicaciones que van más allá de la enfermedad primaria de interés, ya que pueden revelar trastornos secundarios basados en la genética.

La detección y adecuado manejo de riesgos personales, riesgos reproductivos y farmacogenéticos puede proporcionar un panorama más completo de la salud del individuo y desempeñar un papel crucial en la mejora de la medicina personalizada y la toma de decisiones clínicas informadas. Por tanto, el desarrollo de algoritmos bioinformáticos que aborden estas cuestiones y brinden resultados precisos y significativos es un paso importante hacia una implementación efectiva de la genómica en la práctica médica.

En este sentido, existen numerosas herramientas software destinadas a la búsqueda y priorización de variantes genéticas a partir de datos de secuenciación de exoma o genoma completo. Algunos ejemplos incluyen Exomiser, IVA, ANNOVAR o SOPHiA DDM™ platform (SOPHiA GENETICS SA, Saint-Sulpice, Switzerland)(Medina, n.d.; Smedley *et al.*, 2015; Wang *et al.*, 2010). Aunque estas herramientas bioinformáticas han demostrado ser útiles en el análisis y la identificación de variantes diagnóstico de enfermedad (hallazgos primarios), no manejan de forma automática la identificación de hallazgos secundarios sobre un conjunto de genes definidos. Al margen de estas herramientas, también se han desarrollado otras para la anotación automática de haplotipos farmacogenéticos, como Stargazer o PharmCAT (Lee *et al.*, 2019; B. Li *et al.*, 2023).

Así, hasta la fecha no se han descrito herramientas bioinformáticas que aborden el manejo automático de hallazgos secundarios. Y ello a pesar de las importantes implicaciones que este tipo de hallazgos puede tener para la salud del individuo, en términos de riesgo personal, reproductivo o farmacogenético.

Por lo tanto, a pesar de su inmenso potencial, actualmente existe una notable brecha en el campo, pues no se dispone de soluciones de software orientadas al análisis automático de hallazgos secundarios a partir de los datos de NGS. Abordar esta necesidad crítica facilitaría el desarrollo de estrategias diagnósticas y terapéuticas personalizadas y su implementación en el ámbito clínico. En definitiva, el desarrollo de algoritmos que automaticen esta tarea permitirá un análisis más eficiente de los datos provenientes de la secuenciación masiva, y supondría un paso más hacia la implantación de la medicina personalizada.

2. Hipótesis y objetivos

La secuenciación genómica para el diagnóstico de enfermedades raras permite, además de la identificación del hallazgo primario o pertinente, la búsqueda intencionada de otros hallazgos secundarios en un conjunto de genes definido que pueden ser relevantes para las perspectivas de salud del paciente o sus familiares. El desarrollo de algoritmos para el manejo de hallazgos secundarios en datos genómicos mejorará la detección automatizada de variantes genéticas asociadas a enfermedades o características genéticas adicionales, lo que abrirá la posibilidad de mejorar los estudios genéticos del paciente de forma más personalizada y brindará una interpretación más precisa y rápida de los resultados genómicos.

Objetivo general:

1. Desarrollar algoritmos bioinformáticos para el manejo automático de hallazgos secundarios en datos genómicos.

Objetivos específicos:

1.1 Definir los catálogos de genes correspondientes a cada categoría de hallazgos secundarios: riesgo personal, riesgo reproductivo y riesgo farmacogenético. Esto permitirá identificar de manera adecuada las variantes genéticas relevantes asociadas con cada tipo de hallazgo secundario.

1.2 Definir e implementar la herramienta bioinformática que permitirá el manejo automático de hallazgos secundarios en los datos genómicos analizados.

1.3 Validar la herramienta bioinformática desarrollada empleando conjuntos de datos genómicos conocidos y bien caracterizados. La validación se realizará mediante la comparación de los resultados del algoritmo con los hallazgos secundarios obtenidos previamente a través de metodologías "*in house*" en un conjunto de datos de secuenciación de genoma completo del programa NAGEN.

1.4 Preparar una documentación completa de la herramienta bioinformática que describa su funcionamiento, requisitos y posibles limitaciones. Asimismo, incluir el código para su disponibilidad y uso en abierto, fomentando la colaboración y la mejora continua de la herramienta.

Mediante el logro de estos objetivos, se espera que la herramienta bioinformática desarrollada mejore significativamente la eficiencia y precisión en la detección de hallazgos secundarios en datos genómicos. Esto contribuirá al avance de la medicina personalizada y la implementación efectiva de la genómica en el campo clínico, permitiendo una atención médica más precisa y adaptada a las características genéticas de cada individuo, lo que se traduce en una mejora sustancial de los estudios genéticos y el cuidado del paciente de manera personalizada.

3. Material y métodos

En esta sección se describirá la metodología utilizada para alcanzar los objetivos establecidos en el apartado anterior. La metodología incluye el plan de trabajo, la elaboración de catálogos de genes para la búsqueda de los hallazgos secundarios de acuerdo con las tres categorías principales y el diseño, implementación y validación de la herramienta bioinformática, junto con el plan de gestión de datos.

3.1 Plan de trabajo

Con el fin de alcanzar los objetivos especificados en el apartado anterior, el trabajo se distribuyó en diferentes tareas con sus respectivas horas estimadas (**Tabla *1***). La Figura 2 muestra un diagrama de Gantt con la planificación temporal de las tareas de acuerdo con las fechas de entrega programadas en la asignatura.

Tabla 1 | Plan de trabajo desglosado por tareas. Se indican los objetivos específicos a los que corresponden cada tarea y las horas estimadas.

Objetivo	Tarea	Horas
1.1	1. Generación de un catálogo de genes para la identificación de hallazgos secundarios accionables de riesgo personal	4
	2. Generación de un catálogo de genes para la identificación de hallazgos secundarios accionables de riesgo reproductivo y riesgo en la descendencia	4
	3. Generación de un catálogo de genes para la identificación de hallazgos secundarios accionables de riesgo farmacogenético	4
1.2	4. Definición de la funcionalidad de la herramienta bioinformática para la identificación y el manejo automático de hallazgos secundarios	2
	5. Implementación de la herramienta bioinformática de acuerdo con la funcionalidad definida	81
1.3	6. Ejecución de tests para la comprobación del correcto funcionamiento de la herramienta con conjuntos de datos con hallazgos secundarios previamente identificados	4
1.4	7. Documentación de la herramienta y su inclusión en repositorios tipo Github	5
	8. Redacción de la memoria/documento TFM	40
	9. Defensa pública del TFM	2
TOTAL HORAS		150

Figura 2 | Diagrama de Gantt de la planificación temporal del TFM. En sombreado gris se muestran las tareas completadas y en rojo las pendientes. Los cuadrados indican los hitos entregables.

3.2 Catálogos de genes

3.2.1 Genes de riesgo personal

Los hallazgos de riesgo personal son variantes genéticas asociadas con un mayor riesgo de desarrollar ciertas enfermedades. Para facilitar su identificación y manejo, el ACMG publica anualmente una lista mínima de pares gen-fenotipo. Esta lista tiene como objetivo proporcionar una guía para el cribado oportunista de estas variantes y facilitar intervenciones establecidas que permitan prevenir o reducir significativamente la morbilidad y mortalidad asociadas a estos trastornos. Actualmente, en su versión v3.2, la lista incluye 81 genes relacionados con cáncer hereditario, enfermedades cardiovasculares, metabolopatías y fenotipos misceláneos como la hemocromatosis, la telangiectasia o la enfermedad de Wilson (Miller *et al.*, 2023). Estos 81 genes constituirán el catálogo de genes de riesgo personal.

3.2.2 Genes de riesgo reproductivo y riesgo en la descendencia

Los hallazgos de riesgo reproductivo y de riesgo en la descendencia consisten en variantes genéticas con posibles implicaciones para la salud reproductiva de una pareja, tanto por suponer un riesgo en la descendencia, como por estar relacionadas con la fertilidad, abortos recurrentes o complicaciones durante el embarazo. El ACMG publicó en 2021 un enfoque por niveles o *tiers* para el cribado de portadores, y enumera una serie de genes recomendados para su inclusión en el cribado. Estos *tiers* se definen en base a la

frecuencia de portadores. En concreto, el *tier* 4 incluye genes relacionados con patologías menos frecuentes (frecuencia de portadores <1/200), así como los incluidos en los niveles anteriores. Esto aumenta en gran medida el número de condiciones a cribar y permite, por tanto, identificar un mayor número de parejas en riesgo. Así, el nivel 4 comprende 112 genes, que serán los incluidos en el catálogo de genes de riesgo reproductivo (Gregg *et al.*, 2021).

3.2.3 Genes de riesgo farmacogenético

Los hallazgos farmacogenéticos son variantes genéticas que influyen en la respuesta individual a ciertos medicamentos. Una de las principales barreras de la implementación de la farmacogenómica radica en la dificultad para traducir las variantes genéticas en decisiones de prescripción farmacológica. Con el fin de facilitar esa traslación del conocimiento farmacogenómico a la práctica clínica, el Consorcio para la Implementación de la Farmacogenética Clínica (CPIC) (*CPIC*, n.d.), en colaboración con la Base de Conocimiento de Farmacogenómica (PharmGKB) (*PharmGKB*, n.d.), elabora periódicamente guías de parejas gen/fármaco. Estas guías revisadas por pares, actualizadas, basadas en la evidencia y accesibles gratuitamente, permiten trasladar los resultados del laboratorio a decisiones de prescripción accionables para fármacos específicos (Relling & Klein, 2009; Whirl-Carrillo *et al.*, 2021). El catálogo de genes de riesgo farmacogenético se basará en estas guías. Además de los genes, se establecerá también la correspondencia entre los diferentes diplotipos y fenotipos de cada gen, así como la puntuación de actividad correspondiente. Un haplotipo es una combinación de variantes génicas que tienden a heredarse juntas en la misma región cromosómica. El diplotipo, por lo tanto, consiste en la combinación específica de dos haplotipos (paterno y materno). Esta combinación puede tener un fenotipo asociado en relación con la respuesta a ciertos fármacos (Judson *et al.*, 2000). Este archivo se elaborará a partir de la información obtenida en PharmGKB.

3.3 Esquema funcional de la herramienta bioinformática

En esta sección se hará una descripción general del funcionamiento de la herramienta, y en la sección **3.4 Implementación** se explicará en mayor detalle.

Como se ha comentado anteriormente, el objetivo principal de este TFM es el desarrollo de algoritmos bioinformáticos para el manejo automático (identificación y gestión) de hallazgos secundarios conforme a las tres categorías definidas: riesgo personal, riesgo reproductivo y riesgo farmacogenético.

Para ello, se requiere como entrada un archivo con variantes genéticas de un solo nucleótido (*Single Nucleotide Variant*, SNV) o pequeñas inserciones o deleciones con una longitud máxima de 50 pares de bases (*indels*) en formato *Variant Calling Format* (VCF), un fichero de texto que almacena de forma estandarizada este tipo de variantes. Permite además proporcionar un archivo de texto con los HPOs relacionados con la clínica del paciente, con el objetivo de relacionar los resultados con estos términos. Asimismo, se ofrece la posibilidad de seleccionar como entrada en el teclado las categorías de hallazgos secundarios a analizar (riesgo personal, reproductivo y farmacogenético), el genoma de referencia (GRCh37 o GRCh38), así como el modo de análisis: básico o avanzado. En el modo básico, las variantes se anotan e interpretan mediante la herramienta InterVar (Q. Li & Wang, 2017). En el modo avanzado, la herramienta utiliza además información de la base de datos ClinVar (Landrum *et al.*, 2020) con este fin, y el usuario puede especificar el nivel de evidencia deseado para dichas anotaciones (Sección 3.4.1.6 Obtención de la base de datos ClinVar). En las siguientes secciones se explicará en más detalle el proceso.

El flujo de trabajo (Figura 3) comienza con la normalización del archivo VCF de entrada, lo que garantiza un formato adecuado para su posterior procesamiento (Sección **3.4.2 Normalización**).

A continuación, la herramienta extrae del fichero VCF las variantes genómicas identificadas en los genes definidos en los catálogos para cada una de las categorías de hallazgos secundarios (riesgo personal, RP; riesgo reproductivo, RR; riesgo farmacogenético, FG). Para ello, se realiza la intersección del VCF con los listados de genes de las categorías seleccionadas por el usuario: RP, RR y/o FG (Sección **3.4.3 Intersección con archivos BED**). Para las categorías RP y RR, los genes se encuentran representados en un archivo en formato BED (Browser Extensible Data) que proporciona las regiones genómicas de los genes de interés. En el caso de los hallazgos farmacogenéticos, la

intersección se realiza con las posiciones genéticas concretas, también en formato BED. De este modo se obtienen tres VCFs filtrados diferentes: VCF_RP con el conjunto de variantes genómicas identificadas según el catálogo de genes de riesgo personal, VCF_RR con las correspondientes al catálogo de riesgo reproductivo y VCF_FG con las correspondientes al catálogo de riesgo farmacogenético.

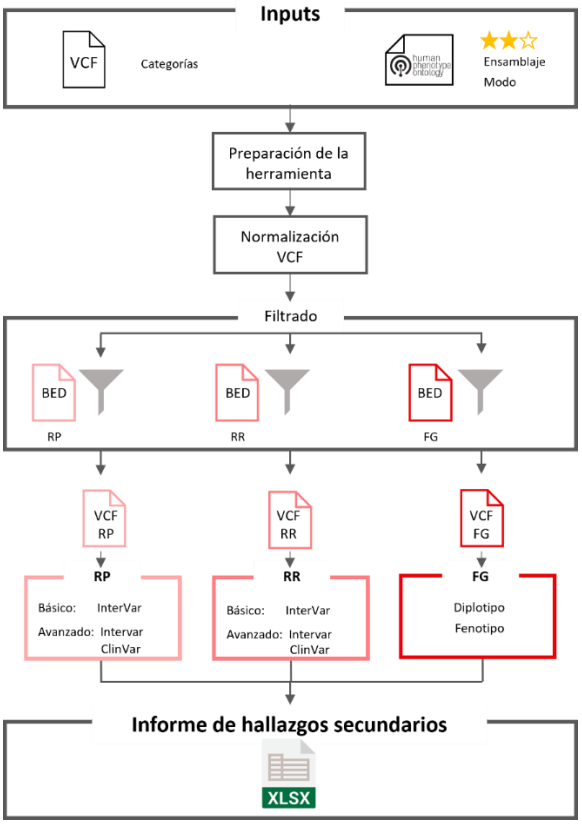


Figura 3 | Diagrama de flujo del diseño funcional de la herramienta. VCF y BED indican el formato de los archivos. Se muestra el proceso para las tres categorías de interés: riesgo personal (RP), riesgo reproductivo (RR) y farmacogenética (FG).

El siguiente paso consiste en ejecutar el módulo correspondiente a la categoría seleccionada por el usuario.

Para los módulos de riesgo personal y riesgo reproductivo, se utiliza la herramienta InterVar (Q. Li & Wang, 2017) para, a partir del VCF de variantes, obtener el listado de variantes patogénicas o probablemente patogénicas, según la clasificación del ACMG (Richards *et al.*, 2015). Adicionalmente, en el modo avanzado, a partir del VCF de variantes se obtienen aquellas catalogadas por ClinVar como patogénicas, probablemente patogénicas o conflictivas (etiquetadas como *conflicting*) con al menos una evidencia patogénica. ClinVar informa del nivel de revisión de estas interpretaciones mediante un

nivel de evidencia o "review status" del 0 al 4, siendo 0 las variantes sin criterios de afirmación, y 4 aquellas incluidas en guías de práctica clínica. En la **Tabla 2** se muestra el significado de cada uno de los niveles. La herramienta permite filtrar las variantes de acuerdo con este *review status*, según el nivel de evidencia seleccionado por el usuario. Los resultados se etiquetan según su origen: ACMG y/o ClinVar (Landrum *et al.*, 2020).

Tabla 2 | Niveles de evidencia de la base de datos ClinVar. Se indica el número de estrellas o nivel y su significado (*review status*) (Landrum & Kattman, 2018).

Número de estrellas	*Review status*
0	Sin criterios de afirmación y/o evidencia
1	Único remitente o conflictos de interpretación
2	Múltiples remitentes, sin conflictos entre sí.
3	Revisado por panel de expertos
4	Guía de práctica clínica

En el caso del módulo de riesgo farmacogenético, a modo de prueba de concepto se han seleccionado 5 genes por su relevancia terapéutica y menor complejidad de cara a la implementación (*TPMT*, *NUDT15*, *DPYD*, *CYP2C9* y *CYP2C19*). La herramienta determinará el diplotipo para cada uno de estos genes en base a las variantes identificadas y a las tablas de nomenclatura de alelos correspondientes. Estas tablas contienen la información sobre las variantes que definen cada haplotipo, y se encuentran disponibles en la página web de PharmGKB. A continuación, se asignará el fenotipo en función de la correspondencia diplotipo-fenotipo de PharmGKB-CPIC (Lee *et al.*, 2019; B. Li *et al.*, 2023). El resultado de este módulo incluye una lista de todas las variantes filtradas, así como el diplotipo y fenotipo de los 5 genes mencionados.

Por último y tras la ejecución de los módulos de las categorías seleccionadas por el usuario, la herramienta genera un informe de los hallazgos secundarios identificados en formato Excel, con una hoja por cada categoría de hallazgos. Además, se considera el modo de herencia del gen y el genotipo del paciente, así como posibles casos de heterocigosis compuesta (dos variantes distintas en un mismo gen) (Sección **3.4.5 Generación del informe de salida**). La Figura 3 presenta un diagrama de flujo que resume el diseño funcional de la herramienta, mostrando el proceso para las tres categorías de interés: riesgo personal (RP), riesgo reproductivo (RR) y farmacogenético (FG).

3.4 Implementación

La herramienta se desarrolló en un entorno de línea de comandos y se basa en diversos módulos que interactúan entre sí, empleando para ello el lenguaje de programación Python (versión 3.8). Las librerías y paquetes empleados se recogen en el material suplementario (requirements.txt). La **Figura *4*** resume el flujo de trabajo, junto con las funciones desarrolladas para su implementación.

3.4.1 Preparación de la herramienta

3.4.1.1 Verificación de dependencias

Esta herramienta requiere para su uso la instalación de InterVar. Este software bioinformático permite la interpretación clínica de variantes genéticas según 18 criterios de la guía ACMG (Q. Li & Wang, 2017). InterVar, a su vez, tiene como dependencia ANNOVAR, que anota variantes de un solo nucleótido (SNV) e inserciones/deleciones, además de examinar sus consecuencias funcionales en los genes (Wang *et al.*, 2010). Antes de realizar el análisis de hallazgos secundarios, el script principal verifica si InterVar está disponible.

3.4.1.2 Configuración y parámetros

La ejecución de la herramienta se inicia mediante el script principal denominado "secondary_findings.py", que importa una serie de funciones todas ellas desarrolladas específicamente para esta herramienta. Para su correcto funcionamiento, el usuario debe proporcionar una serie de parámetros de configuración. Estos parámetros se almacenan en un archivo de configuración llamado "config.json", y permiten definir la ubicación de una serie de directorios: principal, categorías de hallazgos secundarios (subdividido a su vez en personal, reproductivo y farmacogenético), temporal, de salida, y de ubicación de las dependencias InterVar y ANNOVAR. La primera parte del código, por lo tanto, implementa un proceso para leer y cargar estos parámetros desde el archivo de configuración. Esto permite al usuario personalizar el comportamiento de la herramienta y especificar rutas.

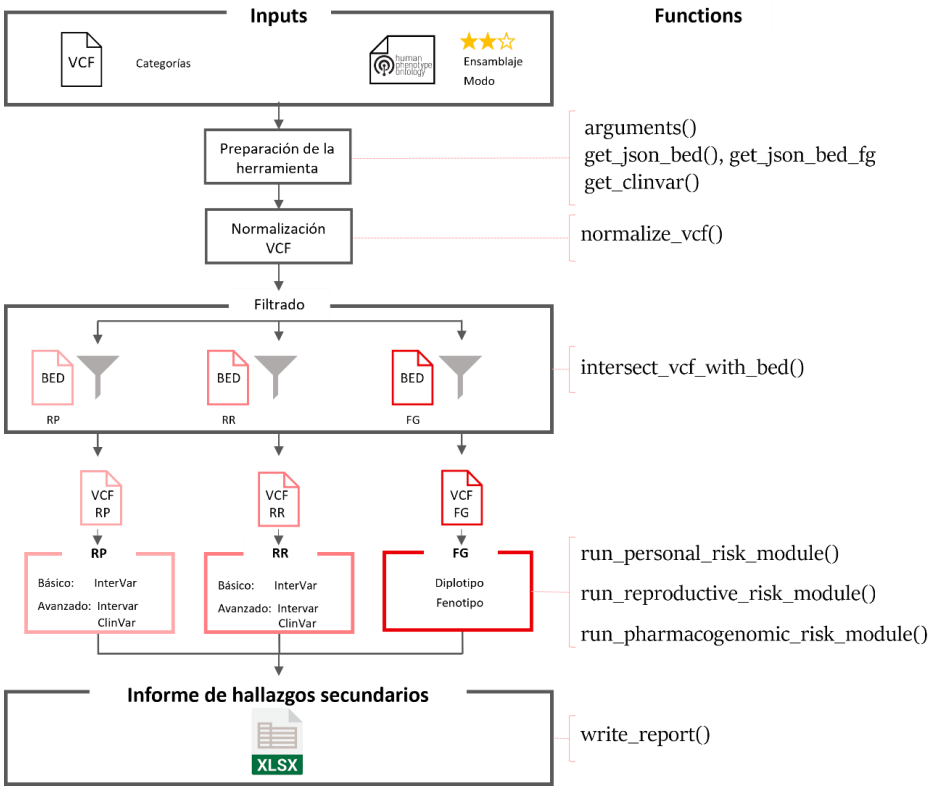


Figura 4 | Diagrama de flujo de la herramienta y funciones correspondientes a cada etapa. VCF y BED indican el formato de los archivos. Se muestra el proceso para las tres categorías de interés: riesgo personal (RP), riesgo reproductivo (RR) y farmacogenética (FG).

3.4.1.3 Creación de directorios

La herramienta, por lo tanto, se organiza en varios directorios que derivan del directorio principal (Figura 5). Esto facilita la gestión y organización de los datos. Los directorios incluyen: "*modules*", con las funciones a las que llama el script principal; "*categories*", que contiene las listas de genes a analizar en cada categoría, así como la correspondencia diplotipo-fenotipo de cada gen en el caso de la farmacogenética; "*clinvar*", con la base de datos ClinVar parseada (3.4.1.6 Obtención de la base de datos ClinVar) y "*temp*", que alberga los archivos intermedios generados durante la ejecución del programa (resultado de la normalización, las intersecciones, la ejecución de InterVar y las combinaciones de resultados). Por último, "*final_output*" contendrá el fichero Excel con el resultado final de variantes a informar en cada categoría y/o diplotipos y fenotipos de farmacogenética.

Además, la herramienta requiere un directorio para las dependencias InterVar y AnnoVar. El siguiente paso del script principal, consiste, por lo tanto, en la comprobación y creación de los directorios necesarios.

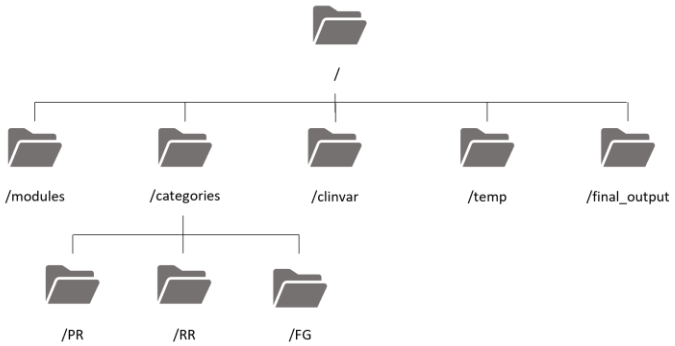


Figura 5 | Organización de los directorios de la herramienta.

3.4.1.4 Obtención de argumentos del usuario

A continuación, el script principal (secondary_findings.py) recibe parámetros directamente del usuario a través de la línea de comandos. Estos parámetros incluyen la ruta del archivo VCF a analizar y argumentos opcionales como el modo de ejecución (básico o avanzado, básico por defecto), el nivel de evidencia ClinVar (1 por defecto), la versión de ensamblaje del genoma (GRCh37 por defecto) y un fichero txt con términos HPO asociados a la clínica del paciente. Estos argumentos se manejan mediante la biblioteca *argparse* de Python (Davis, 2019).

El usuario también indica, como entrada en el teclado, las categorías que desea analizar (PR para riesgo personal, RR para riesgo reproductivo y FG para farmacogenética).

3.4.1.5 Generación del catálogo de genes de hallazgos secundarios en formato JSON y BED

Para realizar los análisis, la herramienta requiere unos archivos *JavaScript Object Notation* (JSON y BED para cada categoría (**Figura 6**). Estos archivos almacenan información sobre los genes de riesgo personal (PR), riesgo reproductivo (RR), y riesgo farmacogenético (FG). Concretamente, el archivo JSON contiene información sobre los genes de cada categoría (fenotipo y código OMIM asociados (McKusick-Nathans Institute of Genetic Medicine, n.d.), modo de herencia, tipo de variantes a informar, y versión del ACMG). El archivo BED, por su parte, contiene las regiones cromosómicas de cada gen (o variante, en el caso de la categoría de farmacogenética).

Archivo BED

1	17345217	17380665	SDHB
1	45794835	45806142	MUTYH
1	55505221	55530525	PCSK9
1	68894505	68915642	RPE65
1	116242628	116311402	CASQ2
1	156052364	156109880	LMNA
1	161284047	161332984	SDHC
1	201008642	201081694	CACNA1S
1	201328136	201346890	TNNT2
1	237205505	237997288	RYR2
2	21224301	21266945	APOB
2	47387221	47403740	CALM2
2	47630108	47789450	MSH2
2	47922669	48037240	MSH6
2	96914254	96931732	TMEM127
2	179390716	179695529	TTN
...	...	...	...

Archivo JSON

```
{
    "category": "Hallazgos secundarios de riesgo personal",
    "genes": [
        {
            "gene_symbol": "RET",
            "phenotype": "Familial medullary thyroid cancer",
            "ACMG_version": "1",
            "OMIM_disorder": "155240",
            "inheritance": "AD",
            "variants_to_report": "All P and LP"
        },
        {
            "gene_symbol": "BRCA1",
            "phenotype": "Hereditary breast and/or ovarian cancer",
            "ACMG_version": "1",
            "OMIM_disorder": "604370",
            "inheritance": "AD",
            "variants_to_report": "All P and LP"
        },
        ...
```

Figura 6 | Ejemplo de archivos BED y JSON. Se muestra a la izquierda un fragmento de archivo BED. Las columnas contienen, de izquierda a derecha, número de cromosoma, posición genómica de inicio, posición final y gen. A la derecha se muestra un fragmento de archivo JSON.

Por ello, antes de ejecutar el análisis, la herramienta comprueba la existencia de los archivos JSON y BED requeridos para cada categoría. Si estos archivos no existen en el directorio especificado, el script principal los generará mediante las funciones desarrolladas get_json_bed() y get_json_bed_fg(). Estas funciones generan un archivo de tipo JSON a partir de un csv con la lista de genes de cada categoría según los criterios del ACMG (y actualizable por el usuario), empleando la librería *json* de Python.

Para generar los archivos BED (específicos del ensamblaje genómico elegido como argumento), es necesario obtener las coordenadas cromosómicas de cada gen. Para ello, la herramienta emplea la Interfaz de Programación de Aplicaciones (API) para Python biomart (Briois, 2017), que permite acceder a BioMart (Smedley *et al.*, 2009). BioMart, a su vez, facilita el acceso a mapeos, relaciones y anotaciones de diversas bases de datos biológicas como Ensembl y Reactome. Cabe señalar que se han considerado las posibles variaciones de nombre del gen entre diferentes ensamblajes.

El proceso para el catálogo de farmacogenética es similar, aunque en este caso se manejan posiciones de variantes concretas, en lugar de genes completos.

3.4.1.6 Obtención de la base de datos ClinVar

A continuación, el script principal maneja la obtención o actualización de la base de datos ClinVar, empleada en el modo avanzado de la herramienta. ClinVar es un archivo público de libre acceso de variantes genéticas humanas e interpretaciones de sus relaciones con enfermedades, mantenido por los Institutos Nacionales de Salud de Estados Unidos

(NIH) (Landrum *et al.*, 2020). En caso de haber optado por el modo avanzado, la herramienta comprueba la disponibilidad de esta base de datos. Si los archivos no se encuentran en el directorio "*clinvar*", la herramienta descargará automáticamente la última versión disponible. En caso de encontrar archivos en el directorio, la herramienta muestra la última versión disponible y pregunta al usuario si desea actualizarla. La descarga y actualización de la base de datos ClinVar se lleva a cabo mediante la función get_clinvar(), que emplea para ello la librería urllib.request, un módulo que ayuda a abrir URLs (principalmente HTTP).

A continuación, este archivo se parsea, generando dos archivos txt (uno por cada ensamblaje disponible, GRCh37 y GRCh38) con la información de interés: significado clínico (interpretación de la variante), *review status* (nivel de evidencia de dicha variante), identificador de dbSNP (número único que la identifica) y posición VCF (sus coordenadas genómicas), entre otros.

3.4.2 Normalización

El siguiente paso consiste en la normalización del archivo VCF de entrada para garantizar que tenga un formato adecuado para su posterior procesamiento. La función normalize_vcf() emplea BCFtools, un programa utilizado en innumerables *pipelines* para procesar y analizar datos de secuenciación de alto rendimiento. Incluye herramientas para la conversión y manipulación de formatos de archivo, clasificación, consulta, estadística, llamada de variantes y análisis de efectos, entre otros (Danecek *et al.*, 2021). Dado que BCFtools se emplea en línea de comandos, el script emplea la librería subprocess, un módulo que permite lanzar procesos, en este caso mediante Windows Subsystem for Linux (WSL2). La normalización incluye la división de sitios multialélicos, así como el alineamiento a la izquierda de *indels*, empleando para ello el genoma de referencia hs37d5 o el GRCh38 de Ensembl (Martin *et al.*, 2023), en función del ensamblaje seleccionado. Además, se eliminan posibles duplicados y se comprueba la indexación de los VCFs comprimidos.

3.4.3 Intersección con archivos BED

Posteriormente, la herramienta realiza la intersección entre el archivo VCF de entrada normalizado y los archivos BED predefinidos mediante la función

intersect_vcf_with_bed(). Esto se realiza para cada categoría seleccionada por el usuario, de manera que se extraen del fichero VCF de variantes aquellas que coinciden con las regiones cromosómicas indicadas en los ficheros BED para cada una de las categorías seleccionadas por el usuario. Para ello se emplea pybedtools, una biblioteca de software Python que facilita la manipulación de conjuntos de datos genómicos en diversos formatos comunes (Dale *et al.*, 2011).

3.4.4 Ejecución de módulos

Finalmente, la herramienta ejecuta los módulos específicos de análisis de acuerdo con las categorías seleccionadas por el usuario. Estos módulos incluyen los módulos de riesgo personal (PR), riesgo reproductivo (RR) y riesgo farmacogenético (FG).

- **Módulo de riesgo personal**: emplea la función run_personal_risk_module(). En el modo básico, se utilizan ANNOVAR e InterVar para anotar e interpretar las variantes según los criterios del ACMG, respectivamente. A continuación, estos resultados se filtran, para obtener sólo las variantes patogénicas o probablemente patogénicas. En el modo avanzado, además de lanzar ANNOVAR e InterVar, las variantes procedentes de la intersección se buscan también en la base de datos ClinVar. A continuación, se retienen todas aquellas cuya clasificación en InterVar y/o en ClinVar sea patogénica, probablemente patogénica o *conflicting* (con alguna evidencia patogénica o probablemente patogénica), de manera que se combina la información de ambas fuentes. En este punto es importante considerar que ANNOVAR modifica la nomenclatura de las inserciones y deleciones. Además, en el filtrado se tiene en cuenta el nivel de evidencia seleccionado por el usuario. Estos resultados combinados se escriben en un archivo intermedio de tipo Tab-Separated Values (TSV), que contendrá todas las variantes patogénicas o probablemente patogénicas detectadas, independientemente de su modo de herencia.
- **Módulo de riesgo reproductivo:** funciona de manera similar al módulo de riesgo personal, pero con la función run_reproductive_risk_module().
- **Módulo de riesgo farmacogenético**: emplea la función run_pharmacogenomic_risk_module(). En este caso, el primer paso consiste en la anotación de las variantes procedentes de la intersección, asignando el gen y el

identificador dbSNP correspondiente. A tal fin se emplea el archivo JSON de variantes farmacogenéticas. Así mismo, a modo de prueba de concepto, como se ha mencionado previamente, se va a inferir el diplotipo y fenotipo de 5 genes (*TPMT*, *NUDT15*, *DPYD*, *CYP2C9* y *CYP2C19*). Para ello, en primer lugar, se obtiene un diccionario a partir del archivo con las correspondencias diplotipo-fenotipo. A continuación, se infiere el diplotipo de cada uno de los genes a partir de las variantes procedentes de la intersección, su genotipo, y una serie de reglas incorporadas en el algoritmo, deducidas a partir de las tablas de asignación de haplotipos de PharmGKB (Whirl-Carrillo *et al.,* 2021). Una vez determinado el diplotipo, se asigna el fenotipo y la puntuación de actividad correspondientes, empleando el diccionario de correspondencias diplotipo-fenotipo obtenido previamente.

3.4.5 Generación del informe de salida

Tras completar la ejecución de los módulos correspondientes, se dispone de la información de variantes patogénicas o probablemente patogénicas para los módulos de riesgo personal y reproductivo, así como de las variantes farmacogenéticas presentes y los diplotipos y fenotipos asociados a 5 de los genes.

Sin embargo, atendiendo a los criterios del ACMG (Green *et al.,* 2013), no es adecuado informar todas las variantes. En el caso de la categoría de riesgo personal, se informan todas las variantes patogénicas o probablemente patogénicas cuyo modo de herencia sea autosómico dominante, semidominante o ligado a X. Por el contrario, si el modo de herencia es recesivo, sólo se informan las variantes en homocigosis y los heterocigotos compuestos (2 variantes diferentes en el mismo gen). En la categoría de riesgo reproductivo, en cambio, se consideran los portadores, por lo que se informarán todas las variantes, independientemente de su heterocigosidad o modo de herencia (Ram & Klugman, 2010).

Por ello, el siguiente paso en la herramienta consiste en comprobar, de manera adecuada a cada categoría, el modo de herencia, el genotipo y la presencia de una o más variantes en el mismo gen. Para determinar la herencia de cada gen se basa en el archivo JSON de genes. En el caso de que las variantes cumplan los requisitos, se combina la

información propia de la variante (procedente de InterVar y ClinVar) y la del gen (del archivo JSON).

A continuación, en el supuesto de que se proporcione un archivo con los HPOs del paciente, se verifica si alguno de los hallazgos de riesgo personal o reproductivo está relacionado con dichos términos, empleando para ello un archivo con la correspondencia gen-fenotipo disponible en la página web de HPO (*HPO*, n.d.). Si alguno de los HPOs asociados a los hallazgos coincide con los HPOs de la entrada de la herramienta (es decir, están relacionados con la clínica del paciente), se añade a la información de la variante.

Finalmente, la herramienta genera un informe de resultados de tipo Excel, con una hoja para cada categoría, y una adicional para las combinaciones diplotipo-fenotipo detectadas en los hallazgos secundarios de riesgo farmacogenético. Este informe, generado mediante la función write_report(), resume los hallazgos y resultados del análisis atendiendo a los criterios mencionados del ACMG.

3.5 Validación

Para garantizar la precisión y confiabilidad de la herramienta, el algoritmo se validó utilizando casos reales con hallazgos secundarios conocidos y anotados. Estos casos clínicos forman parte del proyecto NAGEN1000, una iniciativa estratégica que tiene como objetivo la implementación de la secuenciación de genoma completo en el sistema público de salud navarro, y por ende el desarrollo de la medicina personalizada en la práctica clínica (Pasalodos *et al.*, 2020). Consisten en 12 casos de secuenciación de genoma completo, con hallazgos en alguna de las categorías mencionadas, o sin hallazgos pero con VCF modificados a tal fin. Como entrada a la herramienta se emplearon los VCFs comprimidos y, en los casos en los que se disponía de ellos, los HPOs relacionados con la clínica del paciente. Este proceso permitió comparar los resultados del algoritmo con las anotaciones previas para evaluar su precisión y capacidad para identificar hallazgos secundarios relevantes. Así, se aseguró que la herramienta produzca resultados coherentes y confiables.

A continuación, se describen brevemente cada uno de los casos utilizados en la validación. Todos ellos fueron secuenciados mediante HiSeq Illumina con una profundidad media de 30X, a partir de DNA extraído de sangre periférica:

- **Caso 1:** paciente con discapacidad intelectual moderada (HP: 0002342), clinodactilia (HP:0004209) y comportamiento autista (HP:0000729), sin antecedentes familiares. Sin hallazgo primario.
- **Caso 2:** familiar no afecto de paciente con epilepsia hereditaria, sin antecedentes familiares.
- **Caso 3:** paciente con discapacidad intelectual moderada (HP:0002342), macrocefalia (HP:0005490), insensibilidad al dolor (HP:0007021) y parálisis facial (HP:0007209), sin antecedentes familiares. Hallazgo pertinente reportado: 2 variantes en heterocigosis, asociadas a síndrome autosómico recesivo.
- **Caso 4:** paciente con cardiomiopatía dilatada (HP:0001644), metabolismo lipídico anormal (HP:0003119) y creatina fosfoquinasa sérica elevada (HP:0003236), sin antecedentes familiares, pero con consanguinidad. Sin hallazgo primario.
- **Caso 5:** paciente con discapacidad intelectual moderada (HP: 0002342), trastorno por déficit de atención con hiperactividad (HP:0007018) y obesidad (HP:0001513), sin antecedentes familiares. Sin hallazgo primario.
- **Caso 6:** paciente con cardiomiopatía hipertrófica (HP:0001639) e hipertrofia septal asimétrica (HP:0001670), y madre afecta. Sin hallazgo primario.
- **Caso 7:** paciente con anormalidad morfológica del sistema nervioso central (HP:0002011) y agenesia del cuerpo calloso (HP:0001274), sin antecedentes familiares. Sin hallazgo primario.
- **Caso 8:** paciente con discapacidad intelectual moderada (HP:0002342) y retraso en el desarrollo del habla y del lenguaje (HP:0000750), y antecedentes familiares desconocidos. Sin hallazgo primario.
- **Caso 9:** paciente con encefalopatía epiléptica y discapacidad moderada, sin antecedentes familiares. Sin hallazgo primario.
- **Caso 10:** paciente con hipoplasia del maxilar superior (HP:0000327), deficiencia visual grave (HP:0001141) y anomalía del sistema respiratorio (HP:0002086), y antecedentes familiares desconocidos. Sin hallazgo primario.

- **Caso 11:** paciente con discapacidad intelectual severa, dolicocefalia (HP:0000268) y prognatia mandibular (HP:0000303), y sin antecedentes familiares. Sin hallazgo primario.
- **Caso 12:** familiar no afecto de paciente con cáncer de mama, con antecedentes familiares de cánceres diversos. Sin hallazgo primario.

3.5.1 Plan de gestión de datos

Estos datos requeridos para la validación se encuentran almacenados en la *Genome-Phenome Analysis Platform* (GPAP), un entorno colaborativo que facilita la recopilación, descubrimiento, intercambio y análisis de datos genómico-fenómicos estandarizados. En ella se almacenan perfiles fenotípicos pseudonimizados, codificados mediante la Ontología de Fenotipos Humanos (HPO, un estándar mundial para describir y analizar computacionalmente las anomalías fenotípicas que se encuentran en las enfermedades humanas); datos genómicos procesados de manera estandarizada (a partir de ficheros FASTQ, BAM o CRAM) y metadatos (tipo de experimento, preparación de la librería genómica o estrategia de secuenciación, entre otros) (Köhler *et al.*, 2021; Laurie *et al.*, 2022). Sin embargo, al tratarse de datos altamente sensibles, el acceso es controlado, de manera que sólo el grupo de investigación que los generó puede acceder a ellos (el grupo de Medicina Genómica de Navarrabiomed). En el caso de usuarios sin licencia, los datos de las variantes pueden consultarse de forma limitada vía Beacon.

Para la ejecución de los tests de comprobación de la herramienta fueron necesarios, además de los datos y metadatos procedentes de la GPAP (incluyendo los resultados de los hallazgos secundarios identificados), los archivos Variant Calling Format (VCF) almacenados en el clúster de supercomputación de Nasertic (empresa pública Navarra de Servicios y Tecnologías).

4. Resultados

A continuación, se muestran los principales resultados obtenidos durante el desarrollo de los algoritmos bioinformáticos.

4.1 Catálogos de genes

Los catálogos de genes elaborados para cada categoría se recogen en el material suplementario (Catálogos_genes.xlsx), así como la correspondencia entre diplotipo y fenotipo para cada gen (diplotipo_fenotipo.csv).

4.2 Disponibilidad de la herramienta y documentación

La herramienta desarrollada está disponible en un repositorio de GitHub (https://github.com/edurlaf/tfm_secondaryfindings), lo que facilita el control de versiones, y permite a otros investigadores y desarrolladores acceder al código, colaborar y realizar contribuciones. La documentación de la herramienta es una parte integral de su desarrollo. Se creó una guía de usuario o README, disponible en GitHub, que explica cómo instalar, configurar y utilizar la herramienta de manera efectiva. La documentación incluye una descripción de los archivos generados.

4.3 Casos de prueba: evaluación de rendimiento

Para evaluar el correcto funcionamiento de la herramienta, se emplearon casos de investigación reales, con hallazgos de riesgo personal, reproductivo y farmacogenéticos conocidos y validados mediante técnica ortogonal (secuenciación Sanger en todos los casos), así como casos sin hallazgos secundarios, pero cuyos VCFs han sido modificados manualmente. Para ello, se emplearon VCFs comprimidos de secuenciación de genoma completo procedentes del proyecto NAGEN1000 (Pasalodos *et al.,* 2020). Estos VCFs se utilizaron como entrada para la herramienta, con el fin de identificar variantes genéticas de interés y compararlas con las obtenidas previamente. A continuación, se indican los resultados de estas pruebas, que se recogen asimismo en las tablas 3-5. Además, la ***Figura 7*** muestra una captura de pantalla con ejemplos de estos informes de resultados.

- **Caso 1:** en la categoría de riesgo personal, la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar, es decir, en

ninguno de los dos modos de ejecución (básico y avanzado). Esto concuerda con los resultados del análisis original.

En cuanto a la categoría de riesgo reproductivo, la herramienta detectó una variante en heterocigosis en el gen *CFTR* clasificada como patogénica en ClinVar según las guías de práctica clínica. Esta deleción *in-frame* de tres nucleótidos (c.1521_1523delCTT p.Phe508delPhe) está asociada con fibrosis quística, y su modo de herencia es recesivo, por lo que el paciente es un portador no afecto. Esta misma variante fue detectada en el análisis previo, y confirmada por secuenciación Sanger.

La herramienta también detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *CYP3A4*, *SLCO1B1*, *VKORC1*, *CYP2B6* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para los 5 genes analizados.

- **Caso 2:** en la categoría de riesgo personal, la herramienta detectó dos variantes. Una de ellas se encontraba en el gen *APC* (c.3920T>A p.Ile1307Lys rs1801155), y fue clasificada como variante con conflictos de interpretación en ClinVar, tal como se había informado en el análisis original. InterVar, sin embargo, la clasifica como probablemente benigna. Se requerirían ensayos funcionales para proporcionar una clasificación definitiva de esta variante. Además, la herramienta detectó otra variante clasificada como probablemente patogénica por InterVar, y como VUS en ClinVar, por lo que, aunque se detectó en el análisis original, no se informó. Se trata de una variante en el gen *MYHY* (c.4954G>T p.Asp1652Tyr rs397516233). Las variantes en este gen, de herencia dominante, están relacionadas con cardiomiopatía hipertrófica. En este caso, al no tratarse de un caso índice, no se disponen de los HPOs del individuo, por lo que no ha sido posible confirmar su interpretación.

En cuanto a la categoría de riesgo reproductivo, la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar, lo que concuerda con los resultados del análisis original.

La herramienta también detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *UGT1A1*, *CYP3A4*, *CYP2C19*, *SLCO1B1*, *VKORC1*, *CYP4F2*, *CYP2B6* y *CYP2D6*). Además, el análisis de

diplotipos mostró que el paciente era metabolizador normal para 4 de los 5 genes analizados, y metabolizador ultrarrápido para *CYP2C19*, lo que supondría un cambio en la pauta terapéutica de fármacos como citalopram, escitalopram o dexlansoprazol, entre otros. Estos resultados cuadran con el análisis manual de diplotipos.

- **Caso 3**: en la categoría de riesgo personal, la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar. Esto concuerda con los resultados del análisis original.

  En cuanto a la categoría de riesgo reproductivo, la herramienta detectó una variante en heterocigosis en el gen *CHRNE* clasificada como patogénica en ClinVar. Esta deleción de un nucleótido (c.131dupG p.Glu44Gly rs762368691) está asociada con síndrome de miastenia congénita, y su modo de herencia es recesivo, por lo que el paciente es un portador no afecto. Esta misma variante fue detectada en el análisis previo.

  La herramienta también detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *UGT1A1, CYP3A4*, *SLCO1B1*, *VKORC1*, *CYP2B6* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para los 5 genes analizados.

- **Caso 4**: en la categoría de riesgo personal, la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar. Esto concuerda con los resultados del análisis original.

  En cuanto a la categoría de riesgo reproductivo, al igual que en el caso 1, la herramienta detectó una variante en heterocigosis en el gen *CFTR* clasificada como patogénica en ClinVar según las guías de práctica clínica. Esta deleción *in-frame* de tres nucleótidos (c.1521_1523delCTT p.Phe508delPhe rs113993960) está asociada con fibrosis quística, y su modo de herencia es recesivo, por lo que el paciente es un portador no afecto. Esta misma variante fue detectada en el análisis original, y confirmada por secuenciación Sanger.

  La herramienta también detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *CYP3A4*, *CYP4F2*, *DPYD, CYP2B6* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era

metabolizador normal para 4 de los 5 genes analizados, y metabolizador intermedio para *DPYD*, lo que concuerda con el análisis manual. Esto supone un cambio en la pauta terapéutica de fármacos como la capecitabina.

- **Caso 5**: en la categoría de riesgo personal, la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar. Esto concuerda con los resultados del análisis original.

En cuanto a la categoría de riesgo reproductivo, la herramienta detectó dos variantes. Una de ellas en el gen *MMUT* clasificada como patogénica en ClinVar. Este cambio de nucleótido (c.970G>A p.Ala324Thr rs780387525) está asociado con aciduria metilmalónica, y su modo de herencia es autosómico recesivo, por lo que el paciente es un portador no afecto. Además, también se detectó una variante en *BTD* (c.1330G>C p.Asp444His rs13078881), clasificada como conflictos de interpretación en ClinVar y como VUS en InterVar, y asociada con deficiencia de biotinidasa (herencia recesiva). Ambas variantes fueron detectadas en el análisis previo.

La herramienta también detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *UGT1A1, CYP3A4*, *CYP2C19*, *VKORC1*, *CYP4F2* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para 4 de los 5 genes analizados, y metabolizador ultrarrápido para *CYP2C19*. Estos resultados concuerdan con el análisis manual.

- **Caso 6**: en la categoría de riesgo personal, la herramienta detectó una deleción en el gen *MSH6* clasificada como patogénica tanto en ClinVar como en InterVar. Esta variante (c.1168delG p.Asp390fs rs753796271) está asociada al síndrome de Lynch, y su modo de herencia es autosómico dominante. Esta misma variante fue detectada en el análisis original, y confirmada por secuenciación Sanger.

En cuanto a la categoría de riesgo reproductivo, la herramienta detectó una variante en heterocigosis clasificada como probablemente patogénica por InterVar, y como VUS en ClinVar, por lo que no había sido informada previamente. Se trata de una variante en la región 3'UTR del gen *CHRNE* (g.4804840A>G c.*1012A>G rs146931108). Las variantes en este gen, de herencia recesiva, están relacionadas con miastenia

congénita, por lo que el individuo, de confirmarse su patogenicidad, sería un portador no afecto. En este caso se requerirían ensayos funcionales para proporcionar una clasificación definitiva de esta variante.

La herramienta también detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *UGT1A1*, *CYP3A4*, *CYP2C19*, *VKORC1* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para 4 de los 5 genes analizados, y metabolizador intermedio para *CYP2C19,* lo que supondría un cambio en la pauta terapéutica de fármacos como citalopram, escitalopram o dexlansoprazol, entre otros. Estos resultados concuerdan con el análisis manual.

- **Caso 7**: la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar en las categorías de riesgo personal y riesgo reproductivo. Esto concuerda con los resultados del análisis original.

La herramienta detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *UGT1A1*, *CYP3A4*, *NUDT15*, *VKORC1, CYP4F2, CYP2B6* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para 4 de los 5 genes analizados, y metabolizador intermedio para *NUDT15 (*1/*2)*, lo que supondría un cambio en la pauta terapéutica de fármacos como la azatioprina. Estos resultados concuerdan con el análisis manual.

- **Caso 8**: en la categoría de riesgo personal, la herramienta detectó una variante en el gen *TTN* clasificada como patogénica en InterVar, pero no informada en ClinVar. Esta variante (c.13774G>T p.Glu4592Ter rs754572853) está asociada a cardiomiopatía dilatada, y su modo de herencia es autosómico dominante, por lo que si la variante fuera patogénica el paciente sería afecto. Las plataformas Franklin y Varsome clasifican esta variante como probablemente patogénica (Genoox, n.d.; Kopanos *et al.*, 2019). Sin embargo, la patología asociada no coincide con los HPOs observados, por lo que sería necesario un estudio más exhaustivo.

La herramienta detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *UGT1A1*, *TPMT*, *CYP3A4*,

CYP3A5, *CYP2C19, CYP2C9, VKORC1* y *SLCO1B1*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para 3 de los 5 genes analizados, y metabolizador intermedio para *CYP2C9* (*1/*3) y *TPMT* (*1/*3C), lo que supondría un cambio en la pauta terapéutica de fármacos como meloxicam o la azatioprina, respectivamente. Estos resultados concuerdan con el análisis manual.

- **Caso 9**: en la categoría de riesgo personal, la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar. Esto concuerda con los resultados del análisis original.

En cuanto a la categoría de riesgo reproductivo, la herramienta detectó una variante en heterocigosis clasificada como patogénica en ClinVar e InterVar. Se trata de una variante en el sitio aceptor de *splicing* del gen *FAH* (g.80465355G>A c.707-1G>A rs149052294). Las variantes en este gen, de herencia recesiva, están relacionadas con tirosinemia de tipo I, por lo que se trata de un portador no afecto. Esta variante concuerda con el análisis previo.

La herramienta detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *UGT1A1*, *CYP3A4*, *CYP3A5*, *CYP2C19, CYP2C9, VKORC1, CYP4F2, CYP2B6* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para 3 de los 5 genes analizados, metabolizador intermedio para *CYP2C9* (*1/*3) y metabolizador rápido para *CYP2C19* (*1/*17), lo que supondría un cambio en la pauta terapéutica de fármacos como los mencionados en casos anteriores. Estos resultados concuerdan con el análisis manual.

- **Caso 10**: la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar en las categorías de riesgo personal y riesgo reproductivo. Esto concuerda con los resultados del análisis original.

La herramienta detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *UGT1A1*, *CYP3A4*, *CYP2C19, CYP2C9, SLCO1B1, CYP4F2* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para 3 de los 5 genes analizados, metabolizador intermedio para *CYP2C9* (*1/*3) y metabolizador rápido para *CYP2C19* (*1/*17), lo que

supondría un cambio en la pauta terapéutica de fármacos como los mencionados en casos anteriores. Estos resultados concuerdan con el análisis manual.

- **Caso 11**: la herramienta no detectó ninguna variante patogénica o probablemente patogénica en ClinVar ni InterVar en las categorías de riesgo personal y riesgo reproductivo. Esto concuerda con los resultados del análisis original.

  La herramienta detectó variantes farmacogenéticas que podrían influir en la respuesta del paciente a ciertos medicamentos (en los genes *CYP3A4*, *CYP2C19, CYP2C9, SLCO1B1, VKORC1, CYP4F2, CYP2B6* y *CYP2D6*). Además, el análisis de diplotipos mostró que el paciente era metabolizador normal para 3 de los 5 genes analizados y metabolizador intermedio para *CYP2C19* (*2/*17) y rápido para *CYP2C19* (*1/*17), lo que supondría un cambio en la pauta terapéutica de fármacos como los mencionados en casos anteriores. En el caso de *CYP2C9* no fue capaz de asignar un diplotipo, ya que esta funcionalidad sólo está implementada para las variantes recomendadas como imprescindibles por la Sociedad Española de Farmacogenética y Farmacogenómica. En este caso, la variante rs28371685 no entra en esa categoría. Estos resultados concuerdan con el análisis manual.

- **Caso 12:** se modificó manualmente un archivo VCF de entrada, introduciendo variantes que no estaban presentes en el VCF original. En concreto, se introdujeron 2 variantes distintas en *BRCA1*. La herramienta logró detectar estas variantes y fue capaz de informarlas. Esto destaca la capacidad de la herramienta para identificar heterocigotos compuestos.

Tabla 3 | Hallazgos de riesgo personal. Se muestran los hallazgos patogénicos (P) y probablemente patogénicos (LP).

Casos	Variante	Gen	rs	GT	InterVar	ClinVar	Herencia	Fenotipo
1, 3, 4, 5, 7, 9, 10, 11	-	-	-	-	-	-	-	-
2	Chr14:23885041:C:A	*MYH7*	rs397516233	Het	LP	VUS	AD	Dilated cardiomyopathy
6	Chr2:48026289:CG:C	*MSH6*	rs753796271	Het	P	P	AD	Lynch syndrome
8	Chr2:179613353:C:A	*TTN*	rs754572853	Het	P	NA	AD	Dilated cardiomyopathy

Abreviaturas: GT, genotipo; Het, heterocigosis; VUS, variant of uncertain significance; AD, autosómica dominante; NA, not available.

Tabla 4 | Hallazgos de riesgo reproductivo. Se muestran los hallazgos patogénicos (P) y probablemente patogénicos (LP).

Casos	Variante	Gen	rs	GT	InterVar	ClinVar	Herencia	Fenotipo
1, 4	Chr7:117199644:ATCT:A	*CFTR*	rs113993960	Het	VUS	P	AR	Cystic fibrosis
2, 7, 8, 10, 11	-	-	-	-	-	-	-	-
3	Chr17:4805974:T:TC	*CHRNE*	rs762368691	Het	P	P	AR	Myasthenic syndrome
5	Chr6:49421411:C:T	*MMUT*	rs780387525	Het	LP	P	AR	Methylmalonic aciduria
6	Chr17:4804840:A:G	*CHRNE*	rs146931108	Het	LP	VUS	AR	Myasthenic syndrome
9	Chr15:80465355:G:A	*FAH*	rs149052294	Het	P	P	AR	Tyrosinemia

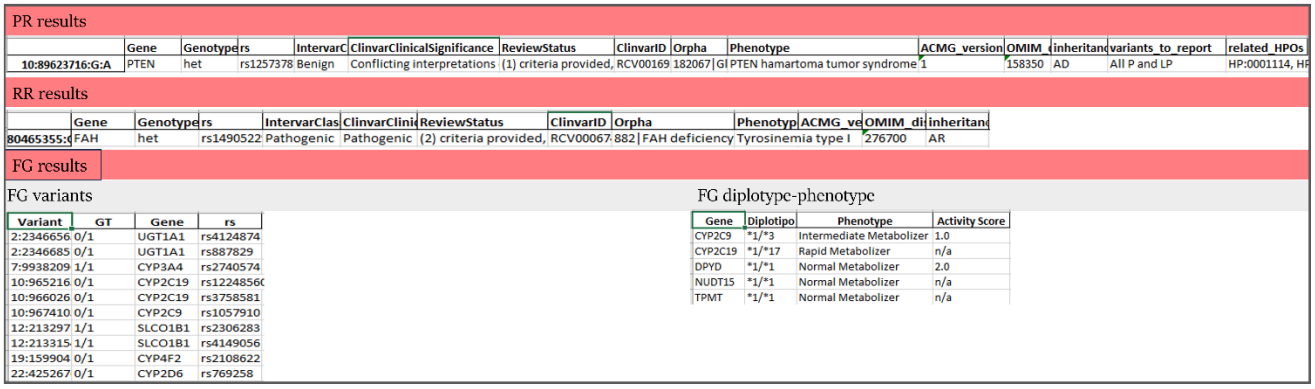
Abreviaturas: GT, genotipo; Het, heterocigosis; VUS, variant of uncertain significance; AR, autosómica recesiva; NA, not available.

Tabla 5 | Hallazgos farmacogenéticos. Se muestran los diplotipos y fenotipos de 5 genes. * simboliza el haplotipo.

Casos	*CYP2C9*	*CYP2C19*	*DPYD*	*NUDT15*	*TPMT*
1, 3	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal
2, 5	*1/*1 Metabolizador normal	*17/*17 Metabolizador ultrarrápido	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal
4	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/HapB3 Metabolizador intermedio	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal
6	*1/*1 Metabolizador normal	*2/*17 Metabolizador intermedio	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal
7	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*2 Metabolizador intermedio	*1/*1 Metabolizador normal
8	*1/*3 Metabolizador intermedio	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*3C Metabolizador intermedio
9, 10	*1/*3 Metabolizador intermedio	*1/*17 Metabolizador rápido	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal
11	NA	*2/*17 Metabolizador intermedio	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal	*1/*1 Metabolizador normal

En definitiva, la validación ha demostrado que la herramienta es capaz de detectar de manera automática variantes genéticas relacionadas con el riesgo personal,

reproductivo y farmacogenético. Ha sido capaz de detectar tanto cambios de nucleótido, como inserciones y deleciones, y las ha informado con éxito en función del modo de herencia, detectando también heterocigotos compuestos. Se han validado el 100% de las variantes halladas originalmente (los informes de la herramienta están disponibles en el material suplementario), y se han detectado variantes candidatas que no habían sido previamente informadas. Estos resultados respaldan su eficacia y utilidad en la genómica clínica.



PR results

	Gene	Genotype	rs	IntervarC	ClinvarClinicalSignificance	ReviewStatus	ClinvarID	Orpha	Phenotype	ACMG_version	OMIM_(	inheritanc	variants_to_report	related_HPOs
10:89623716:G:A	PTEN	het	rs1257378	Benign	Conflicting interpretations	(1) criteria provided,	RCV00169	182067	GI	PTEN hamartoma tumor syndrome	1	158350	AD	All P and LP	HP:0001114, HP

RR results

	Gene	Genotype	rs	IntervarClas	ClinvarClini	ReviewStatus	ClinvarID	Orpha	Phenotyp	ACMG_ve	OMIM_di	inheritanc
80465355:G	FAH	het	rs1490522	Pathogenic	Pathogenic	(2) criteria provided,	RCV00067	882	FAH deficiency	Tyrosinemia type I	276700	AR

FG results

FG variants

Variant	GT	Gene	rs
2:2346656	0/1	UGT1A1	rs4124874
2:2346685	0/1	UGT1A1	rs887829
7:9938209	1/1	CYP3A4	rs2740574
10:965216	0/1	CYP2C19	rs12248560
10:966026	0/1	CYP2C19	rs3758581
10:967410	0/1	CYP2C9	rs1057910
12:213297	1/1	SLCO1B1	rs2306283
12:213315	1/1	SLCO1B1	rs4149056
19:159904	0/1	CYP4F2	rs2108622
22:425267	0/1	CYP2D6	rs769258

FG diplotype-phenotype

Gene	Diplotipo	Phenotype	Activity Score
CYP2C9	*1/*3	Intermediate Metabolizer	1.0
CYP2C19	*1/*17	Rapid Metabolizer	n/a
DPYD	*1/*1	Normal Metabolizer	2.0
NUDT15	*1/*1	Normal Metabolizer	n/a
TPMT	*1/*1	Normal Metabolizer	n/a

Figura 7 | Captura de pantalla de resultados de riesgo personal (PR), reproductivo (RR) y farmacogenético (FG) de diferentes casos.

5. Discusión

El desarrollo de la herramienta bioinformática presentada en este Trabajo de Fin de Máster (TFM) representa un avance significativo en el manejo automático de hallazgos secundarios en estudios genómicos. Esta sección evalúa los resultados alcanzados, resumiendo los objetivos logrados y proporcionando una visión general de las aplicaciones y limitaciones de esta herramienta, así como futuras áreas de mejora.

5.1 Resumen de los objetivos

El objetivo principal de este TFM consistía en el desarrollo de algoritmos bioinformáticos para el manejo automático de hallazgos secundarios en datos genómicos. En este sentido, la herramienta ha demostrado cumplir con éxito este objetivo. Esta herramienta, basada en una arquitectura modular, permite identificar y analizar variantes genéticas asociadas con enfermedades de manera automatizada, tanto de riesgo personal y reproductivo como farmacogenético. Antes del desarrollo de esta herramienta, la identificación de variantes genéticas asociadas con hallazgos secundarios se realizaba principalmente de manera manual, ya que no existía ninguna herramienta específicamente destinada a ello. Esto era un proceso laborioso y propenso a errores. La herramienta desarrollada permite la detección automática y precisa de estas variantes, acelerando significativamente el proceso de análisis genómico.

Además, la herramienta va más allá de la detección de variantes genéticas al incorporar la comprobación del modo de herencia y el genotipo. Evalúa si las variantes cumplen con los criterios de herencia especificados y si están presentes en homocigosis o en heterocigosis compuesta. Esto asegura que solo se informen las variantes de acuerdo con los estándares del ACMG, lo que mejora la interpretación.

Entre los objetivos específicos se encontraba la definición de catálogos de genes correspondientes a cada categoría de hallazgos secundarios. Estos catálogos, que abarcan riesgo personal, riesgo reproductivo y riesgo farmacogenético, son fundamentales para identificar las variantes genéticas relevantes. Esta tarea se ha completado en base a las recomendaciones del ACMG, PharmGKB y del CPIC (Miller *et al.*, 2023; Relling & Klein, 2009).

En este sentido, cabe destacar que la herramienta ha logrado una gran flexibilidad. Por un lado, permite la actualización de las listas de genes de cada categoría de manera sencilla. Los usuarios pueden agregar tantos genes como deseen para ser evaluados en el análisis de hallazgos secundarios, sin necesidad de realizar cambios en el código. La herramienta automáticamente buscará la posición cromosómica de estos genes mediante la API Biomart (Smedley *et al.*, 2009), lo que facilita la personalización del análisis genómico según las necesidades específicas de cada estudio. Por otro, el archivo de configuración y los argumentos proporcionados por el usuario permiten ajustar el diseño del análisis, lo que otorga versatilidad y adaptabilidad.

También, y a diferencia de enfoques estáticos, esta herramienta permite la actualización dinámica de la base de datos ClinVar. En el modo avanzado, verifica la última versión disponible en local y descarga una nueva versión automáticamente si el usuario lo desea. Esta característica garantiza que la herramienta utilice la información clínica más actualizada, lo que es fundamental para la interpretación precisa de las variantes genéticas.

La validación de la herramienta bioinformática desarrollada era otro de los objetivos específicos. Este es un paso crítico para garantizar la fiabilidad y precisión de la herramienta en la detección de hallazgos secundarios en datos genómicos. Para llevarla a cabo, se utilizaron conjuntos de datos genómicos conocidos y bien caracterizados, concretamente, datos del programa NAGEN (Pasalodos *et al.,* 2020). Los resultados se compararon con los hallazgos secundarios obtenidos previamente, lo que permitió evaluar la capacidad de la herramienta para identificar y clasificar correctamente las variantes genéticas. Los resultados de esta validación confirmaron la eficacia de la herramienta y su capacidad para generar resultados coherentes y confiables. Además, permitió identificar otras variantes candidatas no reportadas inicialmente.

El cuarto objetivo específico implicaba la creación de una documentación completa de la herramienta bioinformática. Esta documentación describe de manera detallada el funcionamiento de la herramienta, sus requisitos y su uso. Además, se proporciona el código fuente para su disponibilidad en abierto, lo que fomenta la colaboración y la mejora continua de la herramienta. La disponibilidad en abierto de la herramienta es un paso

significativo para su adopción y uso por parte de la comunidad científica y clínica, ya que permite a otros investigadores y profesionales acceder, entender y utilizar la herramienta de manera efectiva.

Otra funcionalidad a destacar es la capacidad de relacionar los hallazgos con los HPOs del paciente. De este modo, si alguna de las variantes identificadas explica alguna de las características fenotípicas proporcionadas como entrada, la herramienta lo indica en el informe final.

Por último, la herramienta ha logrado un equilibrio entre la generación de informes sencillos y la disponibilidad de datos detallados. Produce informes fáciles de interpretar que resumen los hallazgos secundarios de manera clara. Además, permite a los usuarios acceder a información detallada sobre todas las variantes, no solo las informadas, mediante un archivo temporal. Esta funcionalidad proporciona una visión completa de los resultados y facilita un análisis más profundo si es necesario.

En resumen, la herramienta desarrollada en este TFM ha alcanzado con éxito los objetivos establecidos. Ha demostrado su capacidad para identificar hallazgos secundarios en datos genómicos de manera automatizada y ha sido validada mediante conjuntos de datos conocidos. Además, la documentación completa y la disponibilidad en abierto fomentan su uso y colaboración futura. Por ello, tiene un gran potencial para mejorar la medicina personalizada y la genómica clínica al agilizar el proceso de detección de hallazgos secundarios y ofrecer una interpretación más precisa y rápida de los resultados, lo que se traduce en una atención médica más efectiva y personalizada.

5.2. Aplicaciones potenciales

Esta herramienta puede beneficiar tanto a la comunidad médica como a la investigación genómica:

Medicina personalizada mejorada: La herramienta permite una interpretación más precisa y rápida de los resultados genómicos. Al identificar automáticamente variantes genéticas asociadas a enfermedades o de respuesta a fármacos, los médicos pueden identificar a individuos en riesgo, así como ofrecer tratamientos más personalizados,

basados en la información genética específica de cada paciente. Esto contribuye a la medicina personalizada y al cuidado más efectivo del paciente.

Investigación genómica: La herramienta es una adición valiosa para la investigación genética. Facilita la identificación y el análisis automatizado de variantes genéticas de interés en grandes conjuntos de datos genómicos. Esto acelera el proceso de investigación y permite a los científicos centrarse en la interpretación de resultados y descubrimientos más significativos.

En resumen, la herramienta desarrollada no solo mejora la interpretación de datos genómicos, sino que también abre nuevas oportunidades en medicina personalizada, investigación genética y prevención de enfermedades. Su versatilidad y capacidad para adaptarse a diferentes aplicaciones hacen de esta herramienta una contribución significativa al campo de la genómica y la atención médica personalizada.

5.3. Limitaciones de la herramienta

A pesar de las capacidades y aplicaciones potenciales de la herramienta, también presenta ciertas limitaciones.

En primer lugar, la precisión de los resultados de la herramienta depende de la calidad de los datos de entrada, en particular de la calidad de los archivos VCF. La presencia de errores en estos datos puede dar lugar a interpretaciones incorrectas.

En segundo lugar, la herramienta emplea la base de datos ClinVar para la interpretación de variantes. ClinVar es un archivo público de libre acceso de variantes genéticas humanas e interpretaciones de sus relaciones con enfermedades, mantenido por los Institutos Nacionales de Salud (NIH). Aunque proporciona un recurso inestimable para la interpretación de variantes, no incluye registros de todas las variantes que se han identificado en un genoma humano. Además, acepta interpretaciones de diversas organizaciones, y la actualización de los registros depende del remitente. ClinVar, por tanto, facilita el acceso a las interpretaciones proporcionadas por los miembros de la comunidad de genética clínica, y es el usuario quien debe combinar esos datos con su juicio y experiencia para hacer su propia interpretación (Landrum & Kattman, 2018). Por lo

tanto, si la base de datos contiene información desactualizada o inexacta, la herramienta puede proporcionar resultados incorrectos.

En cuanto a la anotación e interpretación de variantes, la herramienta emplea ANNOVAR e InterVar. Aunque se han descrito otras herramientas, como VEP, en los últimos años ANNOVAR se ha adoptado ampliamente en diversos estudios de investigación y clínicos sobre genomas humanos. Sin embargo, tiene limitaciones en relación con las variantes estructurales y traslocaciones a gran escala (Yang & Wang, 2015). En cuanto a InterVar, interpreta las variantes en base a 18 de los 28 criterios del ACMG, aquellos automatizables. En este sentido, algunos estudios indican que la unión de resultados de distintas herramientas orientadas a la interpretación como CharGer o TAPS podría proporcionar mejores resultados (Scott *et al.*, 2019; Shil *et al.*, 2023; Xavier *et al.*, 2019).

Otras limitaciones a considerar son la incapacidad de la herramienta para diferenciar entre variantes en *cis* y *trans*, lo que puede dificultar el diagnóstico de heterocigotos compuestos o la asignación de los diplotipos. Además, la herramienta se basa en anotaciones genéticas conocidas, lo que limita su capacidad para interpretar variantes genéticas previamente no descritas. La genómica es un campo en constante evolución, y la herramienta debe actualizarse y mantenerse regularmente. Aunque se realizaron pruebas de validación utilizando conjuntos de datos conocidos, la validación continua en diversos contextos clínicos y de investigación es esencial para garantizar su precisión. Cabe señalar que esta validación se realizó únicamente sobre datos de genoma completo, no exomas.

Por último, es importante destacar que, para utilizar eficazmente la herramienta, los usuarios deben poseer ciertos conocimientos técnicos en bioinformática y genética. Esto limita su accesibilidad para médicos o investigadores sin experiencia en estos campos. La interpretación clínica de los resultados requiere experiencia, y las decisiones clínicas basadas en los resultados deben ser tomadas por profesionales capacitados.

En resumen, aunque la herramienta presenta capacidades prometedoras, es fundamental comprender y abordar sus limitaciones para su uso efectivo en los ámbitos clínico y de investigación. La mejora continua y la validación son esenciales para garantizar su precisión y utilidad a largo plazo.

5.4. Futuras mejoras

En línea con las limitaciones mencionadas, algunas de las posibles mejoras que se podrían realizar en el futuro incluyen:

- Mejora de la interpretación de variantes: A pesar de su capacidad para identificar variantes, la herramienta podría beneficiarse de una mejora en su interpretación. Por ejemplo, se podrían utilizar herramientas adicionales como CharGer o TAPES y realizar un *metascore* de ellas (Shil *et al.*, 2023; Xavier *et al.*, 2019).
- Ampliación de la asignación de diplotipos y fenotipos: esta funcionalidad se ha implementado exclusivamente en cinco genes como prueba de concepto, pero podría ampliarse a otros genes.
- Inclusión de más de un VCF: la posibilidad de incorporar como entrada más de un VCF permitiría, por ejemplo, el análisis conjunto de parejas de cara al asesoramiento genético en reproducción.
- Adición de otras bases de datos: A pesar de la capacidad de actualizar automáticamente ClinVar, se podría considerar la incorporación de otras bases de datos relevantes para garantizar la precisión de las anotaciones genómicas, como por ejemplo Franklin (Genoox, n.d.).
- Mejoras en la interfaz de usuario: Una interfaz de usuario más amigable podría hacer que la herramienta sea más accesible para un público más amplio.
- Validación en cohortes más grandes: Aunque se ha realizado una validación inicial, la herramienta podría someterse a pruebas adicionales en cohortes más grandes y con exomas para evaluar su precisión en un conjunto más diverso de casos clínicos.
- Empaquetado en un Docker o Nextflow: La herramienta podría incluirse en un contenedor Docker o un sistema de flujo de trabajo como Nextflow para facilitar su uso y evitar problemas con las dependencias y actualizaciones de software.

Estas futuras mejoras pueden fortalecer aún más la utilidad de la herramienta y su capacidad para contribuir a la genómica clínica y la medicina personalizada. La evolución continua y la adaptación de la herramienta son esenciales para mantenerse al día con los avances en la genómica y las necesidades cambiantes de la comunidad científica y médica.

6. Conclusiones

1. Se ha logrado desarrollar una herramienta bioinformática efectiva y versátil que permite el manejo automático de hallazgos secundarios en datos genómicos. Esto agiliza y mejora significativamente los análisis genómicos. La automatización de esta tarea representa un avance importante en la medicina genómica.
2. La definición de catálogos de genes correspondientes a tres categorías de hallazgos secundarios (riesgo personal, riesgo reproductivo y riesgo farmacogenético) es un componente esencial de la herramienta. Estos catálogos de genes, definidos atendiendo a las recomendaciones del ACMG y CPIC, son esenciales para identificar las variantes genéticas relevantes asociadas con cada tipo de hallazgo secundario.
3. La herramienta ha sido sometida a una validación rigurosa, que incluyó la comparación de sus resultados con hallazgos secundarios obtenidos con metodologías *in house*. Esta validación confirmó su precisión y confiabilidad en la identificación de variantes genéticas relevantes, lo que respalda su utilidad en investigaciones y entornos clínicos.
4. La documentación completa y la disponibilidad del código en un repositorio público fomentan la accesibilidad y la colaboración en la comunidad científica y médica. Esto es esencial para impulsar el uso de la herramienta y permitir mejoras continuas por parte de otros investigadores y profesionales.
5. Este trabajo contribuye al avance de la medicina genómica y la implementación efectiva de la genómica en el campo clínico, lo que permite una atención médica más precisa y adaptada a las características genéticas de cada paciente.
6. Es importante también reconocer las limitaciones de la herramienta, como su dependencia de la interpretación realizada por InterVar. Esta consciencia es esencial para garantizar su uso efectivo en el ámbito clínico y de investigación.
7. Este trabajo ha identificado varias áreas para futuras mejoras, como la inclusión en un Docker y la ampliación de la asignación de diplotipos a otros genes o la mejora de la interpretación de las variantes empleando *metascores*.

7. Referencias

Austin-Tse, C. A., Jobanputra, V., Perry, D. L., Bick, D., Taft, R. J., Venner, E., Gibbs, R. A., Young, T., Barnett, S., Belmont, J. W., Boczek, N., Chowdhury, S., Ellsworth, K. A., Guha, S., Kulkarni, S., Marcou, C., Meng, L., Murdock, D. R., Rehman, A. U., … Rehm, H. L. (2022). Best practices for the interpretation and reporting of clinical whole genome sequencing. *Npj Genomic Medicine*, *7*(1). https://doi.org/10.1038/s41525-022-00295-z

Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nature Reviews Genetics*, *14*(10), 681–691. https://doi.org/10.1038/nrg3555

Briois, S. (2017). *biomart 0.9.2* (0.9.2). https://pypi.org/project/biomart/

Carrasco-Ramiro, F., Peiró-Pastor, R., & Aguado, B. (2017). Human genomics projects and precision medicine. *Gene Therapy*, *24*(9), 551–561. https://doi.org/10.1038/gt.2017.77

CPIC. (n.d.). cpicpgx.org

Dale, R. K., Pedersen, B. S., & Quinlan, A. R. (2011). Pybedtools : a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, *27*(24), 3423–3424. https://doi.org/10.1093/bioinformatics/btr539

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., Mccarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/gigascience/giab008

Davis, T. (2019). *argparse: Command Line Optional and Positional Argument Parser* (Python package version, 2019).

de Wert, G., Dondorp, W., Clarke, A., Dequeker, E. M. C., Cordier, C., Deans, Z., van El, C. G., Fellmann, F., Hastings, R., Hentze, S., Howard, H., Macek, M., Mendes, A., Patch, C., Rial-Sebbag, E., Stefansdottir, V., Cornel, M. C., & Forzano, F. (2021). Opportunistic genomic screening. Recommendations of the European Society of Human Genetics. *European Journal of Human Genetics*, *29*(3), 365–377. https://doi.org/10.1038/s41431-020-00758-w

Di Resta, C., Galbiati, S., Carrera, P., & Ferrari, M. (2018). Next-generation sequencing approach for the diagnosis of human diseases: Open challenges and new opportunities. *Electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine*, *29*(1), 4–14.

Evans, W. E. (2015). *Pharmacogenomics in the clinic*. 3–10. https://doi.org/10.1038/nature15817

Genoox. (n.d.). *Franklin*. https://franklin.genoox.com

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., McGuire, A. L., Nussbaum, R. L., O'Daniel, J. M., Ormond, K. E., Rehm, H. L., Watson, M. S., Williams, M. S., & Biesecker, L. G. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine*, *15*(7), 565–574. https://doi.org/10.1038/gim.2013.73

Gregg, A. R., Aarabi, M., Klugman, S., Leach, N. T., Bashford, M. T., Goldwaser, T., & Chen, E. (2021). Screening for autosomal recessive and X-linked conditions during pregnancy and preconception: a practice resource of the American College of Medical Genetics and Genomics (ACMG). *GENETICS in MEDICINE*, *April*. https://doi.org/10.1038/s41436-021-01203-z

Horton, R. H., & Lucassen, A. M. (2019). *Recent developments in genetic / genomic medicine*. *133*(March), 697–708.

HPO. (n.d.). https://hpo.jax.org/

Judson, R., Stephens, J. C., & Windemuth, A. (2000). The predictive power of haplotypes in clinical response. *Pharmacogenomics*, *1*(1), 15–26. https://doi.org/10.1517/14622416.1.1.15

Katz, A. E., Nussbaum, R. L., Solomon, B. D., Rehm, H. L., Williams, M. S., & Biesecker, L. G. (2020). Management of Secondary Genomic Findings. *American Journal of Human Genetics*, *107*(1), 3–14. https://doi.org/10.1016/j.ajhg.2020.05.002

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., … Robinson, P. N. (2021). The human phenotype ontology in 2021. *Nucleic Acids Research*, *49*(D1), D1207–D1217. https://doi.org/10.1093/nar/gkaa1043

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Aguilera, M. A., Meyer, R., Massouras, A., Saphetor, S. A., & C, E. I. P. (2019). *VarSome : the human genomic variant search engine*. *35*(October 2018), 1978–1980. https://doi.org/10.1093/bioinformatics/bty897

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., Lyoshin, V., Maddipatla, Z., Maiti, R., Mitchell, J., O'Leary, N., Riley, G. R., Shi, W., Zhou, G., Schneider, V., … Kattman, B. L. (2020). ClinVar: Improvements to accessing data. *Nucleic Acids Research*, *48*(D1), D835–D844. https://doi.org/10.1093/nar/gkz972

Landrum, M. J., & Kattman, B. L. (2018). ClinVar at five years: Delivering on the promise. *Human Mutation*, *39*(11), 1623–1630. https://doi.org/10.1002/humu.23641

Laurie, S., Piscia, D., Matalonga, L., Corvó, A., Fernández-Callejo, M., Garcia-Linares, C., Hernandez-Ferrer, C., Luengo, C., Martínez, I., Papakonstantinou, A., Picó-Amador, D., Protasio, J., Thompson, R., Tonda, R., Bayés, M., Bullich, G., Camps-Puchadas, J., Paramonov, I., Trotta, J. R., … Beltran, S. (2022). The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases. *Human Mutation*, *43*(6), 717–733. https://doi.org/10.1002/humu.24353

Lee, S., Wheeler, M. M., Thummel, K. E., & Nickerson, D. A. (2019). *Calling Star Alleles With Stargazer in 28 Pharmacogenes With Whole Genome Sequences*. *106*(6). https://doi.org/10.1002/cpt.1552

Li, B., Sangkuhl, K., Keat, K., Dudek, S., Tuteja, S., Verma, A., Klein, T. E., Whaley, R. M., Woon, M., Verma, S., Carrillo, M. W.-, & Ritchie, M. D. (2023). *How to Run the Pharmacogenomics Clinical Annotation Tool (PharmCAT)*. *113*(5). https://doi.org/10.1002/cpt.2790

Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American Journal of Human Genetics*, *100*(2), 267–280. https://doi.org/10.1016/j.ajhg.2017.01.004

Martin, F. J., Amode, M. R., Aneja, A., Austine-orimoloye, O., Azov, A. G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Bignell, A., Boddu, S., Lins, P. R. B., Brooks, L., Ramaraju, S. B., Charkhchi, M., Cockburn, A., Da, L., … Flicek, P. (2023). *Ensembl 2023*. *51*(November 2022), 933–941.

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, M. (n.d.).

Online Mendelian Inheritance in Man, OMIM®. https://omim.org/

Medina, N. (n.d.). *IVA, Interactive Variant Analysis*. Retrieved July 24, 2023, from http://docs.opencb.org/display/iva/Welcome+to+IVA

Miller, D. T., Lee, K., Abul-husn, N. S., Amendola, L. M., Brothers, K., Chung, W. K., Gollob, M. H., Gordon, A. S., Harrison, S. M., Hershberger, R. E., Klein, T. E., Richards, C. S., & Stewart, D. R. (2023). ACMG STATEMENT ACMG SF v3 . 2 list for reporting of secondary fi ndings in clinical exome and genome sequencing : A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*, 100866. https://doi.org/10.1016/j.gim.2023.100866

NIH. (2023). *DNA Sequencing Costs: Data*. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Pasalodos, S., Salgado, J., Miranda, M., Maillo, A., Matalonga, L., Beltrán, S., Carmona, R., Etayo, G., Lasheras, G., Bernad, T., Pinillos, I., Lasa, I., & Alonso, A. (2020). The Navarra 1000 Genomes Project (NAGEN 1000): Benefits for Predictive, Preventive and Personalized Medicine. *EPMA Journal (2020)*, *11*, 8–11.

Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *Journal of Clinical Medicine*, *9*(1). https://doi.org/10.3390/jcm9010132

PharmGKB. (n.d.). https://www.pharmgkb.org/

Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer Biology and Medicine*, *16*(1), 4–10. https://doi.org/10.20892/j.issn.2095-3941.2018.0055

Ram, K. T., & Klugman, S. D. (2010). *Best practices : antenatal screening for common genetic conditions other than aneuploidy*. https://doi.org/10.1097/GCO.0b013e3283372379

Relling, M. V, & Klein, T. E. (2009). CPIC : Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clinical Pharmacology & Therapeutics*, *89*(3), 464–467. https://doi.org/10.1038/clpt.2010.279

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Rizzo, J. M., & Buck, M. J. (2012). Key principles and clinical applications of "next-generation" DNA sequencing. *Cancer Prevention Research*, *5*(7), 887–900. https://doi.org/10.1158/1940-6207.CAPR-11-0432

Scott, A. D., Huang, K. L., Weerasinghe, A., Mashl, R. J., Gao, Q., Martins Rodrigues, F., Wyczalkowski, M. A., & Ding, L. (2019). CharGer: Clinical Characterization of Germline variants. *Bioinformatics*, *35*(5), 865–867. https://doi.org/10.1093/bioinformatics/bty649

Shil, A., Levin, L., Golan, H., Meiri, G., Michaelovski, A., Tsadaka, Y., Aran, A., Dinstein, I., & Menashe, I. (2023). *Comparison of three bioinformatics tools in the detection of ASD candidate variants from whole exome sequencing data*.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009).

BioMart – biological queries made easy. *BMC Genomics*, *22*, 1–12. https://doi.org/10.1186/1471-2164-10-22

Smedley, D., Jacobsen, J. O. B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., Bone, W. P., Haendel, M. A., & Robinson, P. N. (2015). *Next-generation diagnostics and disease-gene discovery with the Exomiser*. 17–19. https://doi.org/10.1038/nprot.2015.124

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics : TIG*, *30*(9), 418–426. https://doi.org/10.1016/j.tig.2014.07.001

Wang, K., Li, M., & Hakonarson, H. (2010). *ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data*. *38*(16), 1–7. https://doi.org/10.1093/nar/gkq603

Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F., Whaley, R., & Klein, T. E. (2021). An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology and Therapeutics*, *110*(3), 563–572. https://doi.org/10.1002/cpt.2350

Xavier, A., Scott, R. J., & Talseth-Palmer, B. A. (2019). TAPES: A tool for assessment and prioritisation in exome studies. *PLoS Computational Biology*, *15*(10), 1–9. https://doi.org/10.1371/journal.pcbi.1007453

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols*, *10*(10), 1556–1566. https://doi.org/10.1038/nprot.2015.105