UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

DISEÑO Y DESARROLLO DE UN PIPELINE PARA EL ANÁLISIS DE VARIACIONES GENÓMICAS ASOCIADAS A DISTROFIAS HEREDITARIA DE RETINA

Trabajo Fin de Máster

Máster Universitario en Ingeniería Biomédica

AUTOR/A: Pérez Martínez, María

Tutor/a: Costa Sánchez, Mireia

Cotutor/a: Pastor López, Oscar

Director/a Experimental: García Simón, Alberto

CURSO ACADÉMICO: 2023/2024

Diseño y desarrollo de un pipeline para el análisis de variaciones de distrofia hereditaria de retina

AGRADECIMIENTOS

Quiero aprovechar la ocasión para agradecer el apoyo de mi familia y de mis amigos por estar siempre presentes en mi vida y ayudándome en todas mis decisiones. Por otro lado agradecer también al grupo PROS en especial a Mireia y Adrián ya que sin ellos lo hubiera tenido un poquito más difícil.

RESUMEN

Las distrofias hereditarias de retina (DHR) son un grupo de trastornos genéticos que afectan la retina y que pueden causar pérdida de visión progresiva. Con el avance de las técnicas de secuenciación de nueva generación (NGS), se ha abierto la posibilidad de realizar una identificación más precisa de las variaciones presentes en el genoma de un individuo. Sin embargo, este avance tecnológico ha dado lugar a una gran cantidad de información, planteando así el desafío de determinar qué datos son relevantes y cómo evaluarlos para comprender el impacto clínico de las diferentes variaciones genómicas.

En este contexto, se plantea la necesidad de combinar el conocimiento genético con herramientas bioinformáticas para determinar qué variaciones genéticas son relevantes en un paciente con una determinada enfermedad. Para conseguir este objetivo, resulta fundamental el desarrollo de pipelines personalizados que tengan en cuenta las particularidades de la enfermedad genética de estudio. Este Trabajo Fin de Máster se centra en cubrir las necesidades detectadas en el dominio concreto de las DHR, siendo el objetivo principal el diseño y desarrollo de un pipeline bioinformático para el análisis de variaciones genómicas asociadas a las DHR, destacando la importancia de obtener un pipeline robusto y eficiente que pueda procesar grandes volúmenes de datos de secuenciación de manera sistemática y reproducible.

El diseño y desarrollo de este pipeline constará de varios pasos clave. Primero, se recopilarán datos de diversas bases de datos relevantes para la clasificación, seleccionadas en conjunto con expertos clínicos en DHR. Tras esto, se aplicará un filtrado para realizar la selección de las variantes más prometedoras, seguido de una clasificación según su relevancia clínica basado en la evaluación de múltiples criterios, como pueden ser la evidencia de patogenicidad o la concordancia con la literatura científica previa. Este enfoque busca ofrecer una estrategia integral y personalizada para la clasificación de variantes genéticas en DHR, con el objetivo de impulsar la implementación de la medicina de precisión en el manejo de enfermedades genéticas hereditarias.

Palabras clave: Distrofias hereditarias de retina, clasificación de variaciones, bioinformática,.

ABSTRACT

Hereditary retinal dystrophies (HRDs) are a group of genetic disorders that affect the retina and can cause progressive vision loss. With the development of next-generation sequencing (NGS) techniques, the possibility of more precise identification of the variations present in an individual's genome has been opened up. However, this technological advance has resulted in an extensive amount of information, thus posing the challenge of determining which data are relevant and how to evaluate them in order to understand the clinical impact of different genomic variations.

In this context, the need arises to combine genetic knowledge with bioinformatics tools to determine which genetic variations are relevant in a patient with a given disease. To achieve this goal, it is essential to develop personalized pipelines that consider the particularities of the genetic disease under study. This Master's Thesis focuses on covering the needs detected in the specific domain of HRD, the main objective being the design and development of a bioinformatics pipeline for the analysis of genomic variations associated with HRD, highlighting the importance of obtaining a sturdy and efficient pipeline that can process large volumes of sequencing data in a systematic and reproducible way.

The design and development of this pipeline will consist of several key steps. Firstly, data will be collected from classification-relevant databases, selected in conjunction with clinical HRD experts. After this, filtering will be applied to select the most promising variants, followed by classification according to clinical relevance based on the assessment of multiple criteria, such as evidence of pathogenicity or concordance with previous scientific literature. This approach aims to provide a comprehensive and personalized strategy for the classification of genetic variants in HRD, with the goal of advancing the implementation of precision medicine in the management of inherited genetic diseases.

Key words: Hereditary retinal dystrophies; genomic variations; variations classification; bioinformatics.

MEMORIA

Diseño y Desarrollo de un pipeline para el análisis de variaciones de distrofia hereditaria de retina

Documento I

Índice

Índice figuras

Índice Tablas

Documento I

Documento II

Lista de acrónimos

ACMG, American College of Medical Genetics and Genomics
AD, Autosómico Dominante
ADN, Ácido Desoxirribonucleico
AMP, Association for Molecular Pathology
AR, Autosómico Recesivo
ARN, Ácido Ribonucleico
DHR, Distrofias Hereditarias de Retina
DM, Distrofia Macular
ER, Enfermedad Rara
HSF, Human Splicing Finder
LCA, Leber Congenital Amaurosis
MAF, Minor Allele Frecuency
OCR, Open Chromatin Regions
ODS, Objetivo Desarrollo Sostenible
PI, Preguntas de Investigación
RP, Retinosis Pigmentaria
SNP, Single Nucleotide Polymorphism
TFBS, Transcription Factor Binding Site
TSS, Transcription Start Site
VUS, Variant of Uncertain Significance
WES, Whole Exome Secuencing
WGS, Whole Genome Secuencing

1. INTRODUCCIÓN

En este Trabajo Fin de Máster se diseñará y desarrollará un pipeline para el análisis de variaciones en las distrofias hereditarias de retina, facilitando la gestión de datos genómicos y su aplicación en el ámbito clínico. Por tanto, este proyecto se sitúa en el campo de la bioinformática, un área de gran relevancia en la ingeniería biomédica.

En este capítulo, se proporciona una visión general del trabajo. Se comienza con una introducción a las distrofias hereditarias de retina o DHR, sus características clínicas y genéticas y su impacto. Tras esto, se presenta el concepto de medicina de precisión y su relevancia en enfermedades genéticas como las DHR. Finalmente se discuten las motivaciones del proyecto.

1.1. Distrofias hereditarias de retina

Las distrofias hereditarias de retina (DHR) son un conjunto de enfermedades genéticas caracterizadas por la degeneración progresiva de los fotorreceptores que se encuentran en la retina, que provoca la pérdida de visión en los pacientes (Ayuso & Millan, 2010). Las DHR afectan a la retina y, aunque todas las partes del ojo son importantes para percibir la información visual, la retina es una parte vital del sistema (Figura 1). Esta enfermedad es considerada como enfermedad rara (ER) ya que tiene una prevalencia de 1:3000 personas (Pozo Valero, 2022).

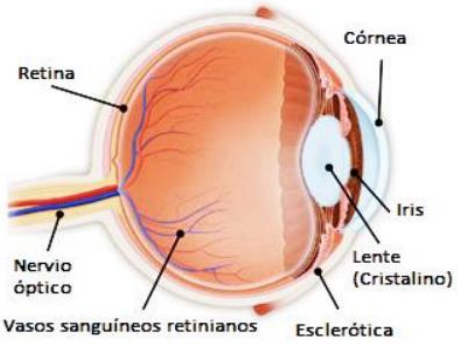


Figura 1. *Anatomía del ojo humano. Extraído de («¿Qué son las Distrofias Hereditarias de Retina?», 2019).*

Los fotorreceptores son unas células específicas del organismo que son capaces de captar luz y convertirla en señales eléctricas que en el cerebro se procesan y las puede interpretar como imágenes (*Fotorreceptores*, 2017). Los dos tipos principales de fotorreceptores son los conos y los bastones. Los bastones son sensibles a la luz tenue, por lo que sobre todo sirven para poder ver de noche y para la visión periférica. Por otro lado, los conos son responsables de la visión central y de percibir los colores, por lo que funcionan mejor con una luz brillante, es decir por el día.

Las DHR tienen diferentes clasificaciones clínicas según los fotorreceptores afectados. Si los bastones (Figura 2, izquierda) son los afectados, será una forma periférica de DHR, dentro de estas se encuentra la retinosis pigmentaria (RP) que es la forma más frecuente (Ayuso & Millan, 2010). Si los conos (Figura 2, derecha) son los afectados, será una forma central de las DHR, estas afectan principalmente a la mácula y la principal es la distrofia macular (DM). Sin embargo, con las nuevas tecnologías y herramientas genéticas se ha visto que hay muchas características que solapan en todas las DHR por lo que los criterios de clasificación están cambiando (Pozo Valero, 2022).

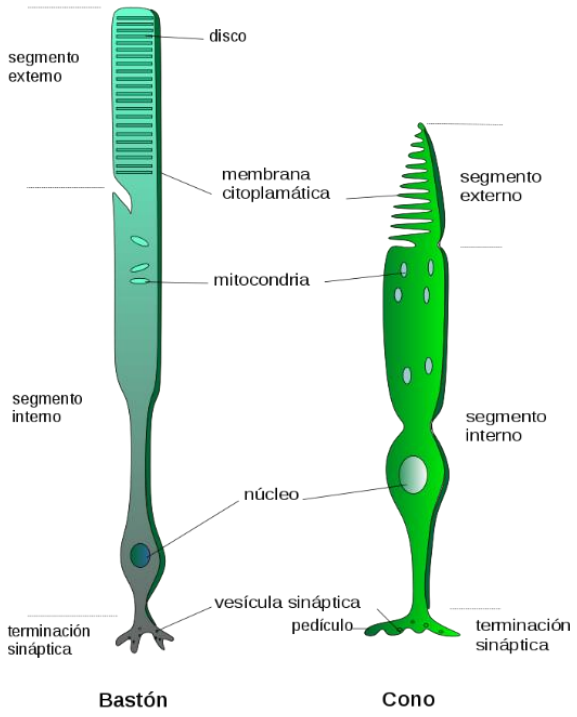


Figura 2. *Fotorreceptores de retina. Extraído de (Wikibooks. 2017).*

En estudios recientes se ha visto que estas enfermedades tienen una base genética muy importante causada por variaciones genéticas, estas variaciones son cambios en la secuencia de ADN que pueden afectar a la función de un gen. Son enfermedades heterogéneas con más de 284 genes descritos como asociados (*RetNet - Retinal Information Network*, 1996).

Algunos subtipos de DHR son poligénicos, ya que existen varios genes que están relacionados con estas enfermedades como RP o amaurosis congénita de Leber (LCA), mientras que otros subtipos de estas DHR son monogénicos y están causadas únicamente por una alteración en un gen como por ejemplo la retinosquisis, la coroideremia o la enfermedad de Stargardt (Pozo Valero, 2022).

Dentro de los patrones de herencia en las DHR se puede encontrar el autosómico recesivo (AR), que ocurre cuando una persona necesita dos copias del gen mutadas para presentar la enfermedad (una de cada progenitor). Las personas con una sola copia mutada solo serán portadores, es decir no tendrán síntomas. También encontramos el patrón autosómico dominante (AD), basta solo con heredar una copia mutada del gen de uno de los progenitores para que se desarrolle la patología. El ligado al cromosoma X (XL), se refiere a genes que se localizan en el cromosoma X, por tanto, en el caso de los hombres al solo tener un cromosoma X van a desarrollar la enfermedad si heredan una copia mutada y en el caso de las mujeres como tienen dos cromosomas X pueden ser solo portadoras. Por último, también se ha descrito herencia mitocondrial, se refiere a genes que están ubicados en el ADN mitocondrial, que se hereda solo de la madre (Ferrari et al., 2011).

En este tipo de enfermedades también son muy importantes las regiones reguladoras del ADN, zonas que como en su nombre se indica regulan la transcripción de genes, por lo que las variaciones en estas regiones son muy relevantes y se deben tener en cuenta ya que pueden causar la enfermedad de forma indirecta.

Las distrofias hereditarias de retina no solo representan un desafío médico debido a su complejidad genética, también tienen un profundo impacto en la calidad de vida de los pacientes. La pérdida progresiva de visión puede afectar significativamente a la calidad de vida, limitar la independencia y causar estrés emocional. Sin embargo, los avances en la investigación genética y en tecnologías como la edición de genes y la terapia génica ofrecen nuevas esperanzas.

1.2. Medicina de precisión

La medicina de precisión representa un paradigma revolucionario en la atención médica, donde los tratamientos y medidas preventivas se seleccionan teniendo en cuenta las complejas interacciones entre los factores genéticos, ambientales y de estilo de vida que distinguen a distintos grupos de individuos. Este enfoque pionero implica una cuidadosa utilización de la información genética y molecular de los pacientes para desarrollar terapias más efectivas, específicas y adaptadas a las necesidades individuales.

En esencia, la medicina de precisión busca personalizar la atención médica, de esta manera se generan tratamientos únicos y medicamentos adaptados a las características de cada paciente. Por lo que se abre una puerta a una medicina más eficiente y con menos efectos secundarios.

Es importante destacar que la medicina de precisión es un proceso que está en continuo desarrollo. Los avances tecnológicos y la disponibilidad cada vez mayor de datos genómicos y moleculares permiten esta continua evolución dando lugar a mejores tratamientos. Esta disciplina se adaptará según la cantidad de conocimiento genético y molecular que se tenga.

Las tecnologías actuales como la secuenciación masiva o de *next generation*, permiten secuenciar muchos fragmentos de ADN al mismo tiempo, esto sumado a programas bioinformáticos que están constantemente recopilando información y generando grandes bases de datos que ayudan a interpretar el significado de las variaciones genéticas, permiten que día a día se identifiquen nuevos genes responsables de enfermedad o cuyas alteraciones determinen enfermedades específicas (Hurtado, 2022).

La medicina de precisión puede ser muy importante en el tratamiento de las DHR. En el caso de las distrofias hereditarias de retina como se ha visto están generadas por variaciones en distintos genes, la secuenciación permite identificar las variaciones exactas y guiar a la elección de terapias. Como por ejemplo en los últimos años, se han desarrollado tratamientos innovadores, como la terapia génica para la LCA, que ha mostrado resultados prometedores en pacientes específicos (Russell et al., 2017).

La secuenciación del genoma permite identificar muchas variaciones en el ADN. Pero no todas supondrán la aparición de una enfermedad por lo que no todas serán relevantes en los tratamientos. Uno de los objetivos de la medicina de precisión es identificar las variaciones que causan la enfermedad, a este proceso se le conoce como clasificación de variaciones.

Para poder hacer esta clasificación de las variaciones hay diferentes guías. La guía más usada es la de ACMG/AMP o *American College of Medical Genetics and Genomics* y la *Association for Molecular Pathology* (Richards et al., 2015). Mediante estas guías se van a clasificar las variaciones en cinco grupos diferentes los cuales son: *Pathogenic, Likely Pathogenic, Benign, Likely Benign* y *VUS*. Las variaciones consideradas como relevantes a nivel clínico son las patogénicas o con alta probabilidad de serlo.

De esta manera, la medicina de precisión puede ofrecer una esperanza significativa a los pacientes con DHR al proporcionar tratamientos y diagnósticos personalizados basados en el perfil genético individual. Realizar la clasificación no es un proceso sencillo, como se explica en la siguiente sección.

1.3. Motivación

Un desafío en el campo de la medicina de precisión es el estudio de las distrofias hereditarias de retina. Este desafío se debe a que, aunque se consideran enfermedades raras, afectan a muchas personas y se asocian con una pérdida gradual de la visión que puede afectar tanto en lo físico como en lo psicológico de una persona. Con una prevalencia de 1:3000, si la población mundial es de alrededor de 8 billones de personas, serían unos 3 millones los afectados por esta enfermedad.

La complejidad que presentan estas patologías se encuentra tanto en su variabilidad clínica como en su heterogeneidad genética, lo que genera preguntas sin resolver sobre su diagnóstico y tratamiento clínico. Actualmente, hay herramientas para clasificar variaciones de diferentes enfermedades. Sin embargo, aún no hay nada que sirva para la correcta clasificación de las distrofias hereditarias de retina.

Para realizar esta clasificación se hace uso de guías de clasificación que sirven para determinar si una enfermedad es patogénica o no, pero el uso de estas guías es complejo. Esta dificultad se debe a que es necesaria la realización de estudios detallados sobre los genes afectados de cada enfermedad y se necesita hacer una interpretación muy meticulosa. En las próximas secciones se explicarán estas guías con más detalle.

De esta manera, la motivación principal de este trabajo de fin de Máster se centra en la necesidad de diseñar y desarrollar un pipeline para el análisis de variaciones genéticas en DHR. Este pipeline tiene como objetivo principal integrar y aplicar herramientas bioinformáticas para abordar aspectos relacionados con estas enfermedades, como la identificación y clasificación de variaciones específicas. Para automatizar la clasificación de variaciones de las DHR, identificando las relevantes.

Al comprender mejor la base genética de estas enfermedades y su variabilidad, se puede avanzar hacia una medicina personalizada más desarrollada. El pipeline propuesto servirá como herramienta útil en la práctica clínica y la investigación, para el desarrollo de diagnósticos personalizados.

2. OBJETIVOS

El objetivo principal de este Trabajo de Fin de Máster es diseñar y desarrollar un pipeline para el análisis de variaciones en el contexto de las Distrofias Hereditarias de retina. Para lograr este objetivo global se plantearán objetivos específicos ligados a preguntas de investigación.

- Objetivo 1. Investigación del problema: Realizar una investigación de la literatura para comprender las DHR y las técnicas y herramientas actuales de análisis de variaciones genéticas.
- Objetivo 2. Diseño y desarrollo: Diseñar y desarrollar un pipeline para el análisis de variaciones genéticas en las DHR.
- Objetivo 3. Aplicación clínica y futuras direcciones: Explorar las posibles aplicaciones clínicas del pipeline desarrollado y discutir las futuras direcciones para su mejora y expansión.

3. METODOLOGÍA

Este Trabajo de Fin de Máster se ha desarrollado siguiendo la metodología de investigación *Design Science* (Wieringa, 2014), propuesta por Wieringa, esta metodología consiste en diseñar e investigar artefactos en un contexto específico para resolver un problema determinado.

El artefacto de este trabajo es diseñar y desarrollar un pipeline para la clasificación de variaciones, en un contexto genómico, específicamente relacionado con las Distrofias Hereditarias de Retina (DHR).

La metodología *Design Science* contempla dos tipos de problemas. En primer lugar, los problemas prácticos o de diseño, estos buscan un cambio en el mundo real, requieren un análisis de los objetivos y la solución diseñada debe considerar las necesidades de todos los *stakeholders*. Por otro lado, están los problemas experimentales o de conocimiento, al contrario de los anteriores estos no buscan un cambio en el mundo real, sino que buscan adquirir conocimiento. La respuesta a estos problemas es una proposición y se asume que solo hay una respuesta correcta con grados de incertidumbre.

Ya que el objetivo principal es diseñar y desarrollar un pipeline para la clasificación de variaciones de origen genómico de las DHR, este trabajo aborda un problema práctico. Por lo tanto, se debe aplicar un ciclo regulativo que consta de 5 etapas:

1. Investigación del problema. En esta etapa se identifica y comprende el problema que se debe resolver. Para hacerlo se pueden responder algunas preguntas como: ¿Cuáles son las necesidades de los usuarios?, ¿Qué se quiere mejorar con el diseño?, ¿Cuál es el marco conceptual?
2. Diseño de la solución. Comprendido el problema se diseña la solución para esto se deben considerar las mejores prácticas y tecnologías, con el objetivo de obtener una solución.
3. Validación de la solución. Se estudia si la solución propuesta puede resolver el problema identificado.
4. Implementación de la solución. Esta etapa implica llevar a cabo el desarrollo del pipeline según el diseño establecido.
5. Evaluación de la implementación. Se hace una evaluación para determinar si se ha resuelto el problema de forma efectiva.

Es importante mencionar que este TFM cubrirá las dos primeras etapas que corresponden a los objetivos expuestos en el apartado anterior, ya que las restantes corresponden a la transferencia a entornos industriales y la evaluación para garantizar el correcto funcionamiento del producto. Por otro lado, a pesar de que no se hayan abarcado estas últimas etapas se ha hecho un caso de estudio como una fase preliminar de validación.

Para realizar la etapa de investigación del problema, en primer lugar, se profundizó en el conocimiento sobre estas patologías mediante una búsqueda bibliográfica. Para ello, se utilizaron herramientas académicas como *Google Scholar* y *PubMed*, donde se revisaron artículos y estudios relevantes.

De forma continua, se investigaron bases de datos de elementos esenciales en la regulación de la expresión genética. La identificación de estos elementos y sus predictores fue también un proceso basado en la revisión bibliográfica. Se seleccionaron bases de datos confiables y se evaluaron los predictores disponibles.

Con esta información, se estudió cómo se diseñan y construyen los pipelines de clasificación de variaciones genéticas. Para esto se hizo otra búsqueda en literatura científica de la anotación, el filtrado y la clasificación de variaciones. Esta revisión permitió entender las metodologías y herramientas empleadas en estos procesos, de esta manera se obtuvo una guía para la implementación en el contexto de las DHR.

Tras la revisión bibliográfica general, se lleva a cabo la segunda etapa de diseño de la solución. Para esto se aplicaron los conocimientos obtenidos al caso específico de las DHR. Esto implicó buscar predictores, bases de datos, frecuencias de filtrado y criterios de clasificación específicos. Para la implementación de estas bases de datos se hizo uso del lenguaje de programación Python, una herramienta fundamental para la manipulación y análisis de grandes volúmenes de datos.

El uso de Python fue esencial en la implementación de las bases de datos y en la automatización de los procesos de anotación, filtrado y clasificación. Se emplearon librerías como Pandas que está diseñada para la manipulación y el análisis de datos, permite cargar, alinear, manipular e incluso fusionar datos.

Se realizó una transformación y mejora de los datos para asegurar una correcta anotación. Tras esto, se llevó a cabo la anotación y después el filtrado haciendo uso de los filtros definidos por el grupo de investigación PROS, estos se adaptaron a las DHR. Para la clasificación de las variaciones, se aplicaron también los criterios implementados y desarrollados por el grupo de las guías de ACMG/AMP también adaptados a las DHR.

Por último, para la validación preliminar se hace uso de un caso de estudio sin embargo la etapa tres no queda terminada ya que para hacer una validación completa es necesario contactar con los expertos y obtener más datos, esto es lo que se está en proceso de hacer.

4. INVESTIGACIÓN DEL PROBLEMA

En esta sección se van a presentar conceptos para comprender los objetivos y el pipeline a desarrollar abordando así el primer objetivo del trabajo. En primer lugar, se revisan los conceptos básicos de biología como ADN, genes, genoma y variaciones. Tras esto se explica el flujo general de análisis de las variaciones del genoma, que incluye las etapas de anotación, filtrado y clasificación. Finalmente, se presenta el Oráculo Genómico de DELFOS, una herramienta diseñada para el análisis de variaciones genómicos, que se tomará como base para este trabajo.

4.1. Fundamentos conceptuales

En el contexto en el que se desarrollará el trabajo, es fundamental entender diferentes conceptos relacionados con la biología genómica como ADN, gen, genoma humano, variaciones, mutaciones y fenotipo.

4.1.1. Conceptos básicos

El ácido desoxirribonucleico (ADN) es la molécula que transporta información genética esencial para el desarrollo y el funcionamiento de los organismos. Estas moléculas son el medio de transmisión de la información genética de una generación a otra. Está compuesto por una doble hélice de nucleótidos: Adenina (A) que se empareja con Timina (T) y Guanina (G) con Citosina (C) (*Definición de ADN - Diccionario de genética del NCI - NCI*, 2012).

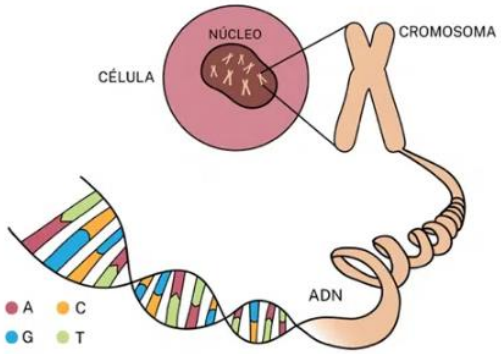


Figura 3. *Esquema de la organización del ADN en el núcleo celular. Extraído de (Ferreiro, 2023).*

A partir del ADN, se generan los genes, que son secuencias específicas responsables de la síntesis de proteínas y la regulación de diferentes funciones celulares. En la Figura 3 se puede ver un esquema de la organización del ADN. Este ADN a su vez se divide en ADN codificante y no codificante. Los genes forman parte del ADN codificante, y están organizados en exones que son las regiones que codifican a proteína y en intrones que son las regiones que no codifican a proteína. Además, a pesar de que el ADN no codificante no produce proteínas específicamente, puede tener otras funciones fundamentales como la regulación de expresión génica. Todos los genes juntos forman el genoma humano (Figura 4) que es el conjunto completo del material genético en una célula.

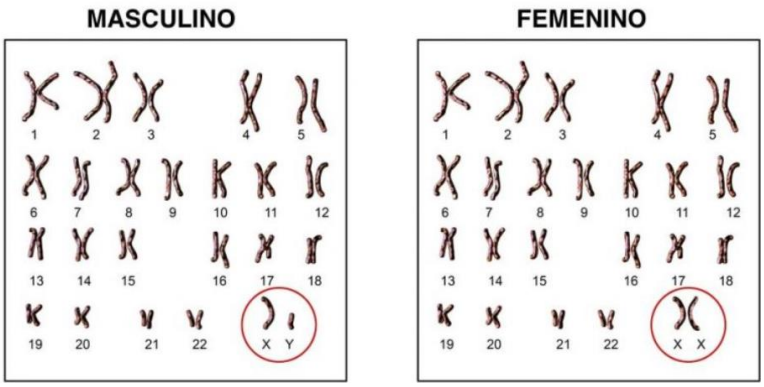


Figura 4. *Cromosomas humanos. Extraído de (*Cromosomas*, 2023.).*

En el genoma pueden aparecer variaciones, que son cambios en la secuencia del ADN. Un ejemplo gráfico del concepto de variación aparece en la Figura 5. Estas variaciones pueden ser heredadas (germinales) o adquiridas durante la vida (somáticas). Algunas de estas son comunes y benignas, pero otras pueden estar asociadas al desarrollo de una enfermedad. Aunque no todas las variaciones causan enfermedades, todas ellas van a contribuir a la diversidad genética.

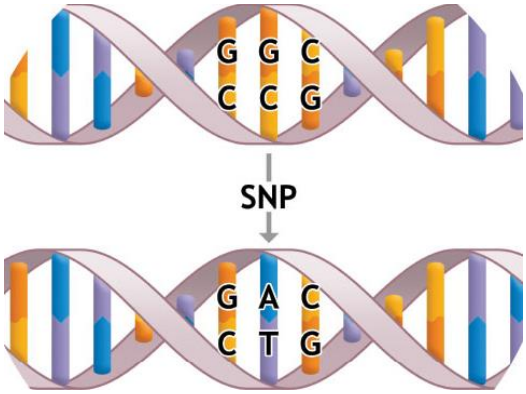


Figura 5. *Variación de Single Nucleotide Polymorphism (SNP). Extraído de (Bhave, 2015).*

Con estos términos generales explicados pasamos a las siguientes partes algo más específicas para poder comprender el desarrollo del pipeline.

4.1.2. Regiones reguladoras

El ADN regulador forma parte del conjunto de ADN no codificante, que constituye más del 95% del genoma humano (García Ordaz, 2011). El ADN regulador (Figura 6) consiste en regiones de ADN que controlan cuánto, cuándo y dónde se lleva a cabo la transcripción de un gen, que es el proceso mediante el cual una célula elabora una copia de ARN desde una pieza de ADN.

En general, cada región reguladora controla la expresión de un único gen. Los genes con patrones de expresión sencillos suelen tener pocos elementos reguladores y localizados en zonas próximas al promotor, estos se encargan de controlar el inicio de la transcripción (Gómez Skarmeta, 2010). Los genes con patrones de expresión complejos, en cambio, tienen muchos elementos dispersos a lo largo del ADN no codificante, tanto en la vecindad del gen como a distancias de hasta cientos de kilobases (Gómez Skarmeta, 2010).

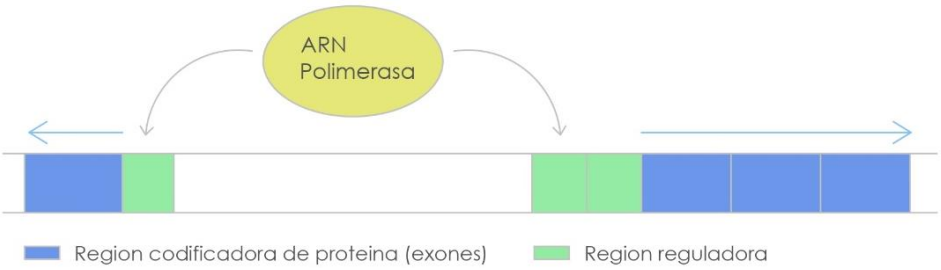


***Figura 6**. Estructura de un gen, regiones reguladoras y codificadoras. Elaboración propia.*

El proceso de transcripción es fundamental para la regulación de la expresión genética. Aunque todas las células tengan el mismo genoma, lo que hace que un tipo celular sea distinto de otro es el conjunto específico de genes que se expresan.

Por otro lado, las regiones reguladoras son de tamaño variable y a estas se van a unir un número de factores de transcripción, que son proteínas que ayudan a activar o disminuir la transcripción de un gen. Cuando estos factores se unan a la región reguladora y esta unión favorezca la transcripción de un gen se les denomina potenciador o *enhancer* (Figura 7) y si es al contrario silenciador o *silencers* (Gómez Skarmeta, 2010). Por lo que las regiones reguladoras lo que hacen es modular la expresión génica.

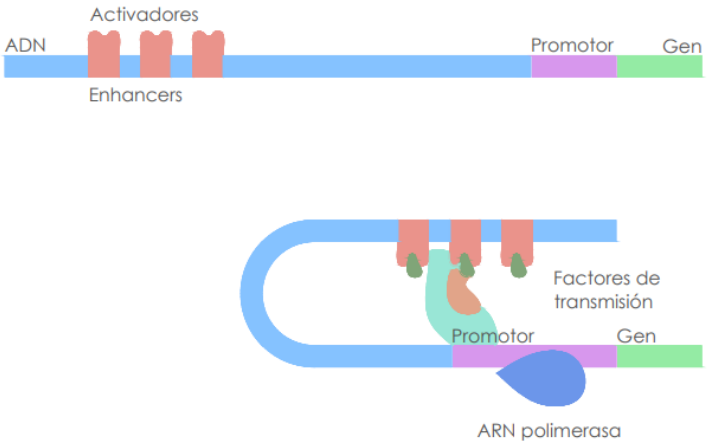


***Figura 7.** Enhancer. Elaboración propia.*

Las secuencias reguladoras pueden contener varios motivos, que son elementos funcionales importantes dentro de estas secuencias. Se le conoce como 'motivo' (*motif* en inglés) a una secuencia patrón de nucleótidos que está distribuida en el genoma y que tiene un significado biológico (Romero Sánchez, 2022).

La identificación de estas regiones reguladoras es muy difícil ya que su código es complejo lo que dificulta predecir donde se localizan y cuál es la información que contienen. Sin embargo, a pesar de esta dificultad como se ha visto son de gran interés debido a su potencial influencia sobre el control de la expresión genética. Lo que puede resultar también en influencia en el desarrollo de enfermedades con la aparición de variaciones que provoquen un funcionamiento incorrecto de las regiones reguladoras.

4.2. Flujo de trabajo general

En la actualidad se usan diferentes técnicas de secuenciación como la secuenciación de exomas completos (WES) o la secuenciación de genomas completos (WGS), estas técnicas se emplean para identificar las variaciones genéticas en el ADN de los individuos. Gracias a estas, la secuenciación del ADN se ha vuelto más económica y sencilla, lo que ha ampliado la posibilidad de hacer diagnósticos genómicos. Estos avances han abierto nuevas líneas de investigación sobre la interpretación del significado biológico de las variaciones genéticas asociadas a determinadas enfermedades. El objetivo es entender qué variaciones son las importantes y en cuáles hay que poner más atención.

Una vez se tiene el ADN del paciente, el primer paso es identificar las variaciones. Para esto, se realiza la secuenciación del ADN que implica usar diferentes programas de alineamiento frente a un genoma de referencia. Este genoma de referencia es aceptado por la comunidad científica y representa a un individuo ideal y estándar, por lo que todo lo diferente a este son variaciones. Uno de estos programas puede ser *Burrows-Wheeler Aligner* (Li & Durbin, 2009) entre otros. Este paso implica detectar las diferencias en la secuencia de ADN entre la muestra analizada y el genoma de referencia. El flujo de trabajo se puede ver en la siguiente Figura 8.

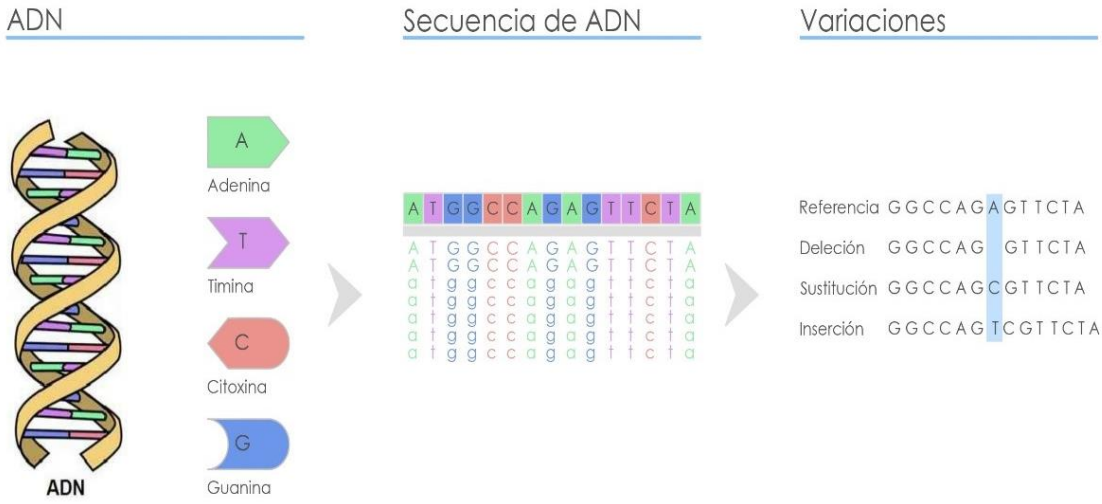


Figura 8. *Proceso de identificación de variaciones en la secuencia de ADN. Elaboración propia.*

La identificación de estas variaciones es solo el primer paso de un proceso más amplio y complejo. Una vez detectadas variaciones, el siguiente paso es determinar su relevancia para una enfermedad. Para esto se realizan etapas de anotación, filtrado y clasificación. Este flujo es importante ya que permite transformar los datos brutos del ADN en información útil para la investigación y la práctica clínica.

4.2.1. Anotación

La anotación de variaciones es el proceso por el cual se le adjudica información biológica a las variaciones, este paso recibe la información de los datos genómicos de un paciente en formato VCF. Este tipo de archivos tienen un formato de texto estándar en bioinformática para almacenar información sobre variaciones de secuencias genómicas. Están formados por una cabecera que describe el contenido del archivo y está indicado con una almohadilla (#) al comienzo de la línea y por 8 columnas obligatorias que están separadas por tabuladores.

En estas columnas obligatorias encontramos los campos de CHROM, que es típicamente el cromosoma que se está registrando. POS, la posición de la variación en la secuencia. ID, un identificador de la variación. REF, la base de referencia en la posición dado para la secuencia de referencia. ALT, es una lista de alelos alternativos para la posición dada. QUAL, calificación asociada con la inferencia de los alelos dados. FILTER, indica el estado de filtrado de la variación. INFO, una lista de campos que van a describir la variación y FORMAT, una lista extensible de campos que sirve para describir las muestras. Un ejemplo de cómo se vería un VCF se puede ver en la siguiente Figura 9.

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With
Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO
FORMAT      Sample1         Sample2         Sample3
2      4370   rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2
GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330   .         T    A       3    q10    NS=5;DP=12;AF=0.017
GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696 rs6055    A    G,T     67   PASS
NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
2/2:35:4
```

***Figura 9**. Ejemplo de VCF.*

Por lo que, en esta etapa se añade información conocida sobre variaciones, desde en que proteína están hasta que frecuencia tienen en la población, etc. Lo que se hace es dar el contexto necesario para determinar la relevancia de una variación. Esta anotación, está reflejada de forma esquemática en la Figura 10.

Las herramientas de anotación en las últimas décadas han experimentado un crecimiento notable. Procesan grandes cantidades de variaciones a través de pipelines de análisis personalizados y las anotan con características completas que se han generado empíricamente. De forma general, las bases de datos de anotación incorporan la información más actual en el campo de la genómica.

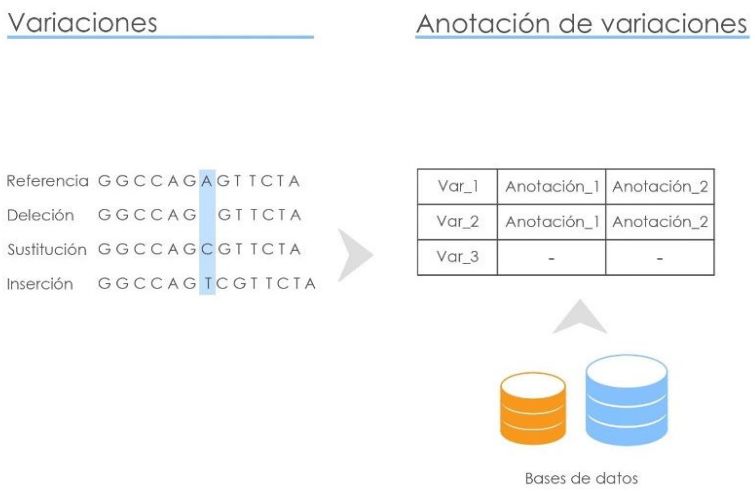


Figura 10. *Proceso de anotación de variaciones. Elaboración propia.*

En los últimos años ha aparecido la anotación basada en datos experimentales y observacionales (Shoemaker et al., 2001) y otras fuentes de información. Uno de los ejemplos más conocidos es ANNOVAR (Wang et al., 2010), publicado en 2010.

Sin embargo, la práctica actual no se limita al uso de ANNOVAR, sino que se han desarrollado más herramientas de anotación que tienen en cuenta nuevos datos e informaciones como, por ejemplo, AnnoGen (Sheng et al., 2020) y Snpeff (Cingolani, 2022).

Hay muchas fuentes de información sobre variaciones que van a ser usadas por estas herramientas de anotación, desde los datos experimentales mencionados conseguidos por diferentes proyectos como ENCODE (de Souza, 2012), GTEx (Carithers et al., 2015) o FANTOM5 (Lizio et al., 2015) entre otros, bases de datos de interpretaciones como ClinVar[1] (Landrum et al., 2018) o bases de datos de proteínas como Uniprot[2] (UniProt Consortium, 2023).

Por tanto, la anotación proporciona información biológica sobre secuencias de ADN. Esta es la base para comprender la función y el impacto de las variaciones en el genoma, es el paso previo a los de filtrado y clasificación.

4.2.2. Filtrado

Existen miles de variaciones en el ADN de un individuo, por lo que reducir el número de variaciones a analizar es fundamental para poder hacer los análisis de manera más eficiente. Para reducir este número elevado se recurre al filtrado. La finalidad de hacer el filtrado es destacar aquellas variaciones que tengan una probabilidad más elevada de poder causar una enfermedad.

En este trabajo nos vamos a centrar en el filtrado de *whole genome secuencing* (WGS) ya que, según lo comentado con los expertos, en el caso de las retinopatías se ha visto que no solo las variaciones que aparecen en el exoma (parte del ADN que codifica para proteína) son las que causan la enfermedad.

[1] https://www.ncbi.nlm.nih.gov/clinvar/
[2] https://www.uniprot.org/

Por lo tanto, es necesario realizar un análisis del genoma completo para una mejor comprensión. En el caso de WGS, se pueden usar varios filtros (Sefid Dashti & Gamieldien, 2017):

- Genes de interés. Este análisis se centra en seleccionar las variaciones que están localizadas en uno o varios genes que se sabe relacionados con la enfermedad para la que se está haciendo el diagnóstico. Los genes de interés se pueden obtener de bases de datos especializadas que proporcionan los genes en lo que se han detectado anteriormente variaciones causantes de la enfermedad de estudio. Adicionalmente, se evalúa la intolerancia a variaciones utilizando scores como el *'missense z'* y puntuaciones pLI de ExAC, que determinan la tolerancia del gen a variaciones *missense* (variaciones que cambian un aminoácido por otro en la secuencia de proteínas) o truncantes (variaciones que generan una proteína acortada o no funcional).

- Herencia. Trabajar con un modelo de herencia o establecer un umbral de frecuencia de variaciones esperadas es una forma útil de filtrar y reducir el número de variaciones que deben analizarse posteriormente con detalle. Este filtro ayuda a identificar variaciones que sean probables de estar relacionadas con la patología en función del patrón de herencia (Sefid Dashti & Gamieldien, 2017).. Además, hay que tener en cuenta los diferentes tipos de herencia. Sin embargo, la capacidad de realizar este paso de filtrado depende de la disponibilidad de genomas/exomas de familiares no afectados.

- Frecuencia (MAF). Se eliminan las variaciones que tengan una frecuencia poblacional superior a un determinado umbral en función del estudio (Sefid Dashti & Gamieldien, 2017). Se considera que por encima de estas frecuencias perjudicarían a demasiadas personas.

- Consecuencia. Las variaciones de truncamiento y *splicing* (variaciones que alteran el proceso de eliminación de intrones y unión de exones en el ARN) son de interés primordial debido a su posible alto impacto celular y sistémico (Sefid Dashti & Gamieldien, 2017). Las variaciones *missense* también se pueden tener en cuenta, estas pueden alterar sitios de *splicing* o de expresión. Por otro lado, las variaciones sinónimas pueden descartarse ya que es menos probable que tengan un efecto relevante.

- Conservación. Las regiones conservadas son aquellas regiones en el genoma que durante la evolución no han sufrido cambios. Estas regiones son muy similares entre diferentes individuos o especies y, a menudo, desempeñan funciones críticas en importantes procesos biológicos (Sefid Dashti & Gamieldien, 2017). Las variaciones en posiciones altamente conservadas tienen más probabilidades de tener un efecto patogénico.

- Regiones específicas, aunque no es lo más común, en ocasiones se puede realizar filtrado por regiones que se sabe que son de interés. De esta manera, si aparece alguna variación en estas regiones se almacena para un futuro análisis más detallado de las mismas.

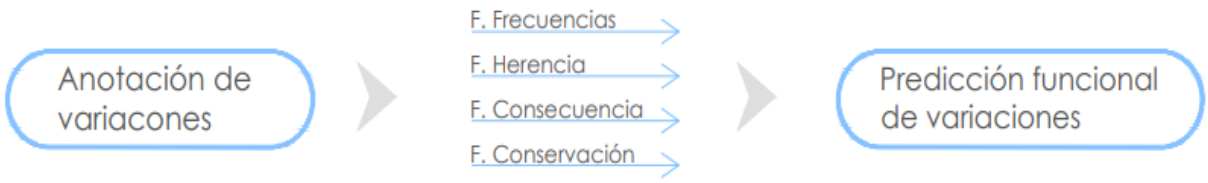


Figura 11. *Proceso de filtrado y predicción funcional de variaciones genéticas. Elaboración propia.*

En la Figura 11 podemos ver un ejemplo de cómo sería un posible proceso de filtrado y predicción ya que no necesariamente hacen falta todos los filtros. Un correcto filtrado aumenta la probabilidad de seleccionar las variaciones con una mayor probabilidad de tener impacto clínico y descartar las que no sean la causa de enfermedad. Por tanto, realizar esta serie de pasos mejora la eficiencia del análisis genómico y aumenta la probabilidad de descubrir variaciones relevantes que podrían pasar desapercibidas.

De esta manera el filtrado es una etapa importante para la identificación de variaciones relevantes de una enfermedad. La combinación de diferentes criterios y bases de datos ayuda a reducir el número de variaciones para un análisis más detallado y preciso.

4.2.3. Clasificación

Tras las etapas de anotación y filtrado, el siguiente paso a llevar a cabo en el análisis de variaciones genómicas es la clasificación. En esta etapa se determinará la relevancia de las variaciones identificadas tras las etapas anteriores. En la actualidad se conoce que una única persona puede contar con millones de variaciones de las cuales 0.6 millones son raras o noveles, aquí es donde empieza la complejidad de interpretar el significado clínico de las variaciones (Amendola et al., 2016). En general, la clasificación va a servir para traducir todos los datos genómicos obtenidos a información que puede ser de utilidad en la investigación biomédica y en la práctica clínica.

La clasificación establece una relación entre las variaciones y la enfermedad mediante una etiqueta que indica cuál es el grado de relevancia de cada una. De forma general, como se ve en la Figura 12, las etiquetas que se usan para la clasificación de una variación son patogénica (*pathogenic*), probablemente patogénica (*likely pathogenic*), benigna (*benign*), probablemente benigna (*likely benign*) y VUS (*variant of uncertain significance*) o variante de significado incierto.

***Figura 12**. Categorías de clasificación de las variaciones de la ACMG/AMP. Elaboración propia.*

Para realizar la clasificación existen diferentes guías, entre las cuales destacan las guías ACMG/AMP (Richards et al., 2015), que fueron desarrolladas con el objetivo principal de servir como recurso de ayuda para los genetistas de laboratorio clínico. Desde su publicación hasta hoy, esta guía ha sido el gran referente en la clasificación.

Tras la publicación de la primera guía, diferentes grupos de investigación y organizaciones han creado sus propias guías a través de la adaptación de las guías ACMG/AMP a contextos clínicos específicos. En estas modificaciones se pueden incluir criterios adicionales o ajustes, para reflejar mejor las características de genes y enfermedades concretas.

Algunos de los grupos que adaptan estas guías son: ENIGMA y ClinGen. Por un lado, ENIGMA que a su vez forma parte de ClinGen, se enfoca principalmente en genes relacionados con cáncer de mama y ovario, como pueden ser BRCA1 y BRCA2. Gracias a las modificaciones introducidas por este grupo en las guías ACMG/AMP, son capaces de reflejar mejor las particularidades de estos genes y las variaciones relacionadas con el riesgo de la aparición de esta patología. Por otro lado, encontramos ClinGen, una organización que ha creado diferentes paneles de expertos y grupos de interpretación de variaciones que se centran en genes y enfermedades variadas.

Estos grupos adaptan los criterios definidos en las guías ACMG/AMP al contexto clínico que estudian, asegurando que las variaciones se interpreten con la mayor precisión posible. Por ejemplo, dentro de ClinGen encontramos criterios para enfermedades como cardiomiopatías, deficiencia del factor de coagulación y en la actualidad están trabajando sobre DHR, que es el objeto de estudio de este trabajo.

Tanto las guías ACMG/AMP originales como las posteriores adaptaciones se basan en evaluar diferentes tipos de evidencia. En estas se pueden encontrar siete grupos de criterios en los que se basan para realizar la clasificación.

En primer lugar, se encuentra el criterio basados en datos poblacionales (*Population Data*), este criterio va a hacer uso de la frecuencia que tiene una variación en diferentes grupos de poblaciones. Para este método se indica que se deben usar diferentes bases de datos como gnomAD (Gudmundsson et al., 2022) y ExAC (Karczewski et al., 2017) para comparar la frecuencia de la variación en la población en general con la frecuencia que se espera en individuos con la patología. En general, las variaciones patogénicas o probablemente patogénicas suelen ser raras en la población, pero comunes en las personas con la enfermedad, por lo que usando umbrales específicos de frecuencia se pueden descartar variaciones comunes y no patogénicas.

Otro criterio es el basado en análisis computacional o predictivo (*Computational and predictive Data*). Para este método se hace uso de algoritmos bioinformáticos que permiten predecir el impacto de las variaciones, entre estos podemos encontrar: CADD (Rentzsch et al., 2019), PolyPhen-2 (Adzhubei et al., 2013a), SIFT (P. C. Ng & Henikoff, 2003). Estos generan puntuaciones que reflejan la probabilidad de que una variación tenga un efecto dañino y según estas se clasifican las variaciones. Por ejemplo, si en CADD se obtiene una puntuación alta la variación tiene mayor probabilidad de ser maligna.

El criterio basado en datos funcionales (Functional Data) se basa en estudios experimentales que evalúan el impacto funcional de una variación. Estos estudios son cruciales para determinar cómo una variación específica afecta el funcionamiento normal de las proteínas y los genes.

Por otro lado, los datos de segregación (Segregation Data) analizan la cosegregación[3] de una variación con la enfermedad dentro de familias afectadas. Si una variación se encuentra consistentemente en los individuos afectados, pero no en los no afectados de una misma familia, es muy probable que sea patogénica.

Los datos de novo[4] (De Novo Data), son especialmente relevantes en enfermedades de alta penetrancia, es decir aquellas enfermedades que van a mostrar signos y síntomas con la presencia de la variación. Si aparece una nueva variación en un niño con una enfermedad grave, y esta variación que no está presente en los padres, la variación tiene una alta probabilidad de ser patogénica. Este método requiere múltiples fuentes y estudios para identificar correctamente una variación *de novo.*

Finalmente, el uso de bases de datos de clasificación de variaciones es crucial. Estas bases de datos contienen información sobre variaciones previamente estudiadas, incluyendo su clasificación y la evidencia asociada.

La clasificación de variaciones genómicas es un proceso complejo que integra múltiples tipos de evidencia para determinar la relevancia clínica de cada variación. Las quías determinan exactamente cómo aplicar los diferentes criterios y como resultado de la aplicación de estos criterios se obtiene la variación de la clasificación.

[3] Transmisión simultánea de dos o más genes del mismo cromosoma por estar situados muy cerca entre sí (*Definición de cosegregación - Diccionario de genética del NCI - NCI*, 2012).

[4] Las variaciones *de novo* son cambios en la secuencia de ADN que se observan por primera vez en un individuo y que no ha aparecido en generaciones anteriores (*Definición de variante de novo - Diccionario de cáncer del NCI - NCI*, 2011), por lo que no es heredada de los padres si no que aparece nueva en un individuo.

4.3. Oráculo genómico de DELFOS

El pipeline generado en este Trabajo Final de Máster está fundamentado en el Oráculo genómico de DELFOS (Leon et al., 2024), una plataforma desarrollada por el grupo PROS del Instituto Valenciano de Investigación en Inteligencia Artificial (VRAIN). Esta plataforma está diseñada para facilitar la gestión y el análisis de datos genómicos con el objetivo de mejorar el diagnóstico y tratamiento de enfermedades, impulsando así la medicina de precisión (Costa Sánchez, 2021).

La estructura de DELFOS permite automatizar la anotación, filtrado, clasificación y generación de informes clínico-genómicos. Su principal ventaja es que permite añadir nuevas fuentes de información a la anotación, elegir las que se quiere usar, y permite al usuario definir sus propias estrategias de filtrado y clasificación. Por tanto, DELFOS ofrece la oportunidad de construir pipelines personalizados.

Debido a las capacidades de personalización ofrecidas por DELFOS en este Trabajo de Fin de Máster se va a usar la base tecnológica de la plataforma, adaptándola para construir un pipeline de análisis de variaciones personalizado al contexto de las distrofias hereditarias de retina (DHR). Se adaptarán las etapas de anotación, filtrado, clasificación y generación de informes a las necesidades del análisis de las DHR.

Para poder desarrollar el pipeline específico para la clasificación de variaciones de las DHR, se realizó una reunión con expertos clínicos e investigadores del Hospital la Fe de Valencia, que nos aportaron su conocimiento para determinar qué aspectos de DELFOS eran más relevantes a la hora de la adaptación a las DHR. Durante esta reunión los genetistas destacaron que en la actualidad no existe una herramienta eficaz para la clasificación de estas patologías, y proporcionaron conceptos clave en los que se debía enfocar el diseño del pipeline. Específicamente, resaltaron la importancia de mejorar el análisis de variaciones, estudiando genes desconocidos, regiones regulatoriaS y áreas no cubiertas del exoma, para esto se buscarán criterios y métricas que permitan esas mejoras.

Además, recomendaron la implementación de predictores nuevos y bases de datos centradas en elementos regulatorios. En base a esta reunión, el trabajo se ha centrado en la búsqueda y análisis de predictores y bases de datos sobre regiones regulatorias y en la adaptación de los apartados de filtrado y clasificación de la herramienta, asegurando que el pipeline se ajuste a las necesidades descritas por estos especialistas. A continuación, se explican los principales componentes de DELFOS.

4.3.1. Bases de datos de conocimiento

DELFOS es una plataforma que integra múltiples fuentes de información genómica que son de interés para la interpretación de las variaciones.La integración de datos genómicos en la plataforma DELFOS se realiza mediante la recopilación y unificación de datos provenientes de diversas fuentes para proporcionar un acceso estandarizado a esta información. Esta estandarización se logra gracias al uso de un modelo conceptual, el Modelo Conceptual del Genoma Humano (Conceptual Schema of Human Genome, CSHG), desarrollado por el grupo PROS, propuesto en la tesis *Diseño y desarrollo de un sistema de información genómica basado en un modelo conceptual holístico del genoma humano* de José Fabián Reyes Román (Román & Fabián, 2018).

En el ámbito de los sistemas de información, un modelo conceptual se define como la descripción del conocimiento sobre el dominio en el que se desarrollará un sistema de información (Olivé, 2007). En el contexto del genoma, el modelado conceptual permite representar de manera clara el conocimiento disponible, facilitando así la comprensión del dominio y la gestión eficiente de los datos genómicos (Costa Sánchez, 2021). El CSHG proporciona una visión holística del genoma humano, organizando los datos en cinco vistas que describen diferentes dimensiones del dominio genómico: vista estructural, vista de transcripción, vista de variación, vista de bibliografía y vista de rutas metabólicas (Costa, García S., et al., 2023).

Estas vistas permiten una organización estandarizada de los datos genómicos, independientemente de su origen, facilitando así su integración en la plataforma DELFOS. En particular, las vistas estructurales, de variación y de bibliografía son de mayor relevancia para el análisis de variaciones genómicas en el contexto de las distrofias hereditarias de retina (DHR).

El uso del CSHG como base conceptual en la plataforma DELFOS asegura que los datos integrados se manejen coherente y uniformemente, facilitando la creación de pipelines personalizados de análisis de variaciones acordes a las necesidades de los expertos clínicos e investigadores. Además, en la actualidad la plataforma incorpora diferentes fuentes de información que mapean al modelo conceptual, lo que permite un uso eficiente y estandarizado de la información.

4.3.2. Anotación

Para realizar la anotación, el modelo hace uso de la herramienta de anotación llamada snpEff, que es ampliamente usada en el campo de la bioinformática para la anotación de variaciones.

SnpEff es un software de código abierto que permite anotar y predecir los efectos de variaciones en secuencias genómicas. Anota las variaciones según sus ubicaciones en el genoma (intrones, sitios de *splicing*…) y predice efectos en la codificación como reemplazo de aminoácidos, ganancia o pérdida de codones de inicio o de parada o cambios en el marco de lectura (Cingolani et al., 2012).

De este destaca su velocidad, ya que puede realizar miles de predicciones por segundo. Su flexibilidad permite añadir genomas y anotaciones personalizadas. Otro aspecto importante es que tiene compatibilidad con múltiples especies y tablas de codones y tiene la capacidad de integrarse con el Genome Analysis Toolkit (GATK) de Broad. Por último permite la anotación de variaciones no codificantes (Cingolani et al., 2012), las cuales como se ha ido viendo a lo largo del trabajo son de gran importancia en el caso de las DHR. Es por eso por lo que se ha considerado este anotador para implementar en el pipeline.

Además de este anotador, el grupo PROS ha desarrollado un módulo de anotación propio. Este permite realizar anotaciones específicas basadas en rangos de posición o anotaciones en base a expresiones más complejas. Por lo que la presencia de estos nuevos anotadores desarrollados junto con snpEff permiten una mayor flexibilidad para adaptar el proceso de anotación a las necesidades específicas del análisis de diferentes patologías.

Haciendo uso del algoritmo de la plataforma de Delfos, para la anotación de variaciones genómicas, se utilizan dos tipos de anotación: primaria y secundaria. Una vez procesadas y mejoradas las bases de datos seleccionadas, se almacenan en archivos bien estructurados para facilitar la anotación.

En cuanto a la primaria, los datos se transforman a formato VCF, este formato cuenta con unos campos específicos que cada variación debe tener definidos, estos campos son CHR, POS, REF, ALT, QUAL, FILTER e INFO. Si los valores QUAL y FILTER no están disponibles, se utiliza '.' como valor por defecto.

En la secundaria, los datos se organizan en archivos delimitados por tabulaciones, utilizando separadores específicos para mantener la integridad y claridad de los datos, permitiendo su anotación secundaria mediante operaciones entre *dataframes*.

El proceso de anotación se automatiza, por lo que una vez proporcionados los archivos genómicos adecuados, es decir, la información genómica del paciente y las bases de datos con las que hacer la anotación, esta se hace de manera automática. De esta manera la herramienta proporciona eficiencia y consistencia.

4.3.3. Filtrado y clasificación

En cuanto al filtrado y clasificación, DELFOS se centra en intentar eliminar las inconsistencias que hay en las guías de clasificación de variaciones actuales. El objetivo es estandarizar y clarificar el proceso de clasificación. Para eso se ha desarrollado una herramienta llamada VariantInsight que permite ejecutar de manera automática los pipelines de filtrado y clasificación basados en guías de interpretación.

Para hacer esto, se basa en un modelo conceptual llamado VarClaMM que representa de manera estandarizada los conceptos claves para la clasificación. Consta de una estructura de tres elementos, cada uno con atributos propios y relaciones entre sí. Estos elementos en los que está estructurado el modelo son: constructos de guías, definen la organización y aplicabilidad de las diferentes guías de clasificación; resultados de evaluación, en esta parte se describe como evaluar las variadas guías; y por último la información contextual, en esta se expone la información necesaria para la clasificación, teniendo en cuenta contexto biológico y clínico (Costa, S., et al., 2023). Todo esto permite generar un marco detallado para la interpretación de los datos genómicos.

De esta manera el modelo va a ofrecer una gran ventaja ya que es capaz de estandarizar la clasificación de variaciones. Se van a definir criterios y métricas de forma precisa, un criterio es una regla o estándar cualitativo que guía en la toma de decisiones y evaluaciones mientras que una métrica es una medida cuantitativa que se usa para evaluar y comparar diferentes aspectos. Así, se mejora la claridad y comprensión de las guías y automatiza el proceso de clasificación.

La primera fase del proceso consiste en el manejo de las bases de datos mediante MySQL. A partir de la información almacenada, se genera un archivo Excel que contiene un conjunto de métricas clave. Estas métricas se han definido en base a las guías de clasificación y filtrado por el grupo de investigación PROS, y son específicas para las distrofias hereditarias de retina.

Para realizar el proceso de filtrado y clasificación, se usa VariantInsight, una herramienta desarrollada en el laboratorio. Esta herramienta toma como entrada un archivo VCF anotado con la información extraída de las bases de datos, junto con una guía de filtrado y una guía de clasificación previamente adaptadas mediante métricas y criterios a las diferentes enfermedades a tratar.

La herramienta VariantInsight se ejecuta mediante un script en Python, que tiene parámetros para personalizar el proceso de filtrado y clasificación. Los parámetros disponibles se pueden ver en la Figura 13.

```
user@anotador:~/dev_repos/variant-insight/src$ python3 main.py --help
usage: main.py [-h] -i INPUT -o OUTPUT [-f FILTER_ID] -g GUIDELINE_ID

Data preprocessing

options:
  -h, --help            show this help message and exit
  -i INPUT, --input INPUT
                        Input file path
  -o OUTPUT, --output OUTPUT
                        Output file path
  -f FILTER_ID, --filter_id FILTER_ID
                        Filter id
  -g GUIDELINE_ID, --guideline_id GUIDELINE_ID
                        Guideline id
user@anotador:~/dev_repos/variant-insight/src$
```

***Figura 13**. Parámetros disponibles en VariantInsight.*

- -i INPUT, --input INPUT: Especifica la ruta del archivo de entrada (VCF anotado).
- -o OUTPUT, --output OUTPUT: Define la ruta del archivo de salida, donde se generará el reporte.
- -f FILTER_ID, --filter_id FILTER_ID: Indica el identificador del filtro que se va a utilizar.
- -g GUIDELINE_ID, --guideline_id GUIDELINE_ID: Especifica el identificador de la guía de clasificación a emplear.

El proceso de evaluación se lleva a cabo mediante un enfoque de comparación booleana. VariantInsight genera un *dataframe* en el que se realiza una evaluación de cada criterio según las métricas definidas. El proceso de clasificación sigue los siguientes pasos:

- Evaluación de Criterios, en la que cada criterio cuenta con una operación específica, por ejemplo, 'IN', '>', '<' y un valor asociado. La herramienta busca en los patrones de columnas especificados y, en función del porcentaje de coincidencia con el valor esperado, determina si un criterio se cumple o no.
- Generación de Resultados, en un *dataframe* que sintetiza si los criterios de clasificación han sido cumplidos.

Este *dataframe* se utiliza para generar un informe en formato PDF que contiene la clasificación final de las variaciones genómicas.

Es importante señalar que en este modelo no están implementados todos los criterios que se encuentran en las guías ACMG/AMP. Solo algunos criterios se han incorporado y estos se muestran en la siguiente Tabla 1.

	Criterio	Descripción
Criterios Patogénicos	PVS1	Se refiere a una variante nula en un gen donde la pérdida de función es un mecanismo conocido de enfermedad. Este criterio se divide en dos categorías: • Very Strong (Muy Fuerte): Aplica a defectos de *splicing*, eliminaciones de exones, mutaciones sin sentido o de desplazamiento de marco, duplicaciones de exones y datos de *splicing* de ARN con evidencia de producción de transcripción alternativa a niveles completos. • Strong (Fuerte): Variaciones sin sentido o de desplazamiento de marco, duplicaciones de exones y deleción de sitios de *splicing*.
	PM1.1	Se refiere a que la variación se localice en un hotspot o en un dominio funcional. Se debe a que las variaciones que aparecen en estas ubicaciones específicas tienen mayor probabilidad de ser patogénicas.
	PM2	La variación está ausente en todas las bases de datos poblacionales disponibles, la ausencia de bases de datos sugiere una posible relación con la enfermedad.
	PM4.1	Es un criterio de nivel *moderate*. La variación no está en una región repetitiva, es una inserción o deleción *in-frame* o es de tipo *stop-loss*. Estos tipos de variaciones son más propensas a causar un impacto funcional significativo.
	PM4.2	En este caso sería nivel *supporting*. La variación no está en una región repetitiva, es una inserción o deleción *in-frame* o es de tipo *stop-loss*. Estos tipos de variaciones son más propensas a causar un impacto funcional significativo
	PP3	Se basa en puntuaciones de predictores, si al menos el 75% de estos la consideran deleterea, entonces esto sugiere alta probabilidad de patogenicidad
Criterios Benignos	BA1	Este criterio se centra en que la frecuencia de la variación sea mayor a 0.008. Esto se debe a que una frecuencia alta puede indicar que es común en la población por lo que no se asocia típicamente con enfermedades raras.
	BS1	En este criterio se tiene en cuenta que la frecuencia alélica de la variación sea mayor a 0.0008. Se debe a una justificación similar al criterio anterior.
	BS2	Si la variación ha sido observada en al menos un individuo adulto sano. Se debe a que si hay presencia en individuos sanos adultos entonces la variación no se asocia con una enfermedad.
	BP4	Se basa en puntuaciones de predictores, si al menos el 75% de estos consideran que la variación es tolerada entonces se puede clasificar como benigna. Esto nos va a indicar una menor probabilidad de que la variación sea patogénica
	BP7	El criterio es que la variación sea sinónima, no esté conservada a lo largo de la evolución y no afecte al *splicing*, ya que este tipo de variaciones son menos propensas a tener efectos negativos

***Tabla 1**. Criterios de clasificación basados en el gen REP65.*

Cada criterio tiene un nivel de evidencia asociado, que puede ser 'Muy Fuerte' (*Very Strong*), 'Fuerte' (*Strong*), 'Moderado' (*Moderate*) o 'Apoyo' (*Supporting*). Dependiendo de la combinación de estos niveles de evidenciase se llega a una clasificación final u otra. Por ejemplo, una variación puede ser clasificada como patogénica si cumple con un criterio 'Muy Fuerte', un criterio 'Fuerte' y dos criterios 'Moderados'. Por otro lado, una variante puede ser clasificada como benigna si cumple con dos criterios 'Fuertes' de evidencia benigna.

Es importante destacar que estas reglas de combinación no son arbitrarias, sino que se basan en la experiencia, en la comprensión de la biología molecular y genética por parte de los genetistas y en la literatura científica. En la siguiente Tabla 2 se observan las reglas de combinación de estos criterios para clasificar las variaciones.

Clasificación	**Combinación de criterios**
Patogénico	1 *Very Strong* y ≥ 1 *Strong*
	1 *Very Strong* y ≥ 2 *Moderate*
	1 *Very Strong* y 1 *Moderate* y 1 *Supporting*
	1 *Very Strong* ≥ 2 *Supporting*
	≥ 2 *Strong*
	1 *Strong* y ≥ 3 *Moderate*
	1 *Strong* y 2 *Moderate* y ≥ 2 *Supporting*
	1 *Strong* y 1 *Moderate* y ≥ 4 *Supporting*
Probablemente patogénico	1 *Very Strong* y 1 *Moderate*
	1 *Strong* y 1 *Moderate*
	1 *Strong* y ≥ 2 *Supporting*
	≥ 3 *Moderate*
	2 *Moderate* y ≥ 2 *Supporting*
	1 *Moderate* y ≥ 4 *Supporting*
	1 *Strong* y 2 *Moderate*
Benigno	≥ 2 *Strong*
Probablemente Benigno	1 *Strong* y 1 *Supporting*
	≥ 2 *Supporting*

Tabla 2. *Reglas para la combinación de criterios extraídas de (Genome Network. (2023)).*

Se seguirá trabajando para que todos los criterios puedan estar definidos y para que se puedan adaptar a las diferentes enfermedades.

En la siguiente Figura 14 se observa un esquema del pipeline enfocado a las DHR.

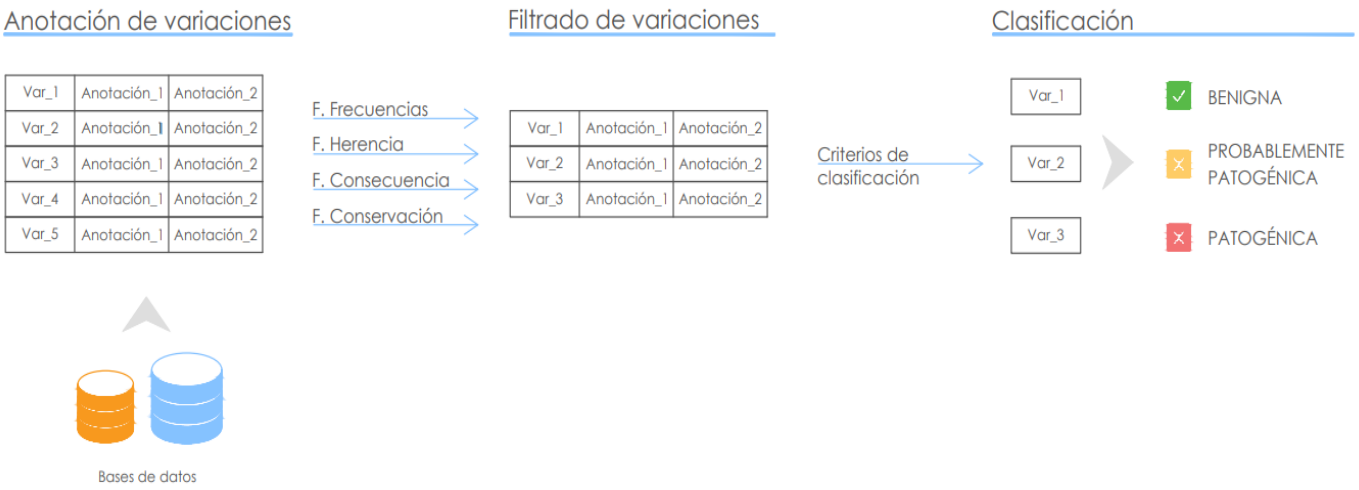


Figura 14. *Esquema general del diseño del pipeline. Elaboración propia.*

4.3.4. Generación de reportes

Una vez realizados los procesos de anotación, filtrado y clasificación se genera un informe en el que aparece toda la información obtenida del análisis de los datos genómicos. En este reporte se van a encontrar los siguientes apartados:

- *Summary of the Results*. Proporciona una visión general de los descubrimientos del análisis (Figura 15). Incluye listado de las variaciones detectadas, gen afectado, alteración y posible consecuencia. Además, se especifica la clasificación de estas variaciones y si alguna de estas ha sido previamente asociada a algún fenotipo de la enfermedad.

Summary of the results

Gene	Transcript	DNA/Protein alteration	Consequence	Zygosity	Classification
CNGB1	ENST00000251102	c.1822G>T, ENST00000251102	stop_gained	A/C	VUS

CNGB1:c.1822G>T, ENST00000251102

Summary

Stop gained located in the CNGB1 gene.
According to RetNet, this gene is associated with the following phenotypes: recessive retinitis pigmentosa. The following description is provided to support this assessment: The following methods were used for determining the gene-disease relationship: homozygosity mapping, candidate gene. The following evidences support the association: consanguineous French family. CNGB1 encodes a complex transcription unit with at least 6 non-overlapping transcripts, one of which is the disease gene in this case.
According to SnpEff, the variant has a HIGH impact on the canonical transcript (ENST00000251102). The variant has been classified as **VUS** by applying the PM2, PM4.2, BP6, BA1@RET, PM1.1, BP7, PM4.1, PM1.2, BS2, BP1, PS4, PP5, BP4, PP3, BS1@RET, PVS1 and BP3 criteria of the ACMG-AMP 2015 interpretation guidelines.
There are 2 publications that mention the variant, as detailed on the bibliography section.

***Figura 15**. Ejemplo del resumen de los resultados del reporte.*

- *Population Frecuency*. Proporciona información de la frecuencia de las variaciones en las diferentes bases de datos poblacionales. Ayuda a determinar si la variación es común o no en la población lo que permite evaluar la relevancia clínica de estas. Cada criterio de clasificación se define y se enumera indicando si la variación lo cumple o no como se puede ver en la Figura 16.

Population Frequency

The variant is present in 2 population databases, with a maximum frequency of 6.570560e-06 and an allele count of 1 in the gnomAD_genomes database.

1000Gp3	exac	gnomAD_exomes	gnomAD_genomes	alfa
.	.	1.368090e-06	6.570560e-06	.

Criteria applied

Name	Description	Result
PM2	Absent from controls	FAIL
BA1@RET	Retinopathy allele frequency is >0.8%	FAIL
BS2	Observed in a healthy adult individual	PASS
BS1@RET	Retinopathy allele frequency is greater than expected	FAIL
PS4	Prevalence in affected individuals is significantly increased compared with controls	FAIL

***Figura 16.** Ejemplo del apartado de frecuencia poblacional del reporte.*

- *Variant Type and Location*. Detalla el tipo de variación y su ubicación precisa en el genoma, indicando si la variación se encuentra en regiones funcionales o estructurales. De nuevo aparecen los criterios aplicados enumerados, definidos y con el resultado de si las variaciones lo cumplen o no (Figura 17).

Variant Type and Location

Stop_gained located in the CNGB1 gene.
The variant is not located in any repetitive region according to Uniprot.
The variant is located in a functional element of the type .. .

Criteria applied

Name	Description	Result
PM4.1	Protein length changes as a result of in frame deletion/insertions in a nonrepeat region	FAIL
PVS1	Null variant in a gene where LOF is a known mechanism of disease	PASS
PM4.2	Protein length changes as a result stop loss variant	FAIL

Figura 17. *Ejemplo del apartado de tipo y localización de variación en el reporte.*

- *Functional Information*. Proporciona un análisis sobre si la variación afecta a dominios funcionales de proteína conocidos basándose en bases de datos como Uniprot o Interpro. Incluyendo detalles de si la variación interfiere o no con regiones críticas para la función de la proteína. Especificándose de nuevo los criterios aplicados (Figura 18).

Functional Information

The variant is not located in any protein domain according to Uniprot and Interpro.

Criteria applied

Name	Description	Result
PM1.1	Located in mutational hotspots	FAIL
PM1.2	Located in functional domain	FAIL

Figura 18*. Ejemplo apartado de información funcional del reporte.*

- *Computational and Predictive Information*. Usa predicciones computacionales para evaluar la probabilidad de que las variaciones tengan un impacto. Se muestra resultados de diferentes herramientas de predicción que evalúan la posible patogenicidad de las variaciones (Figura 19).

Computational and Predictive Information

The variant is predicted as deleterious by 7 predictors, as tolerated by 3 predictors, and as neutral by 3 predictors.

ada_score	rf_score	SIFT	SIFT4G	LRT	MutationAssessor
.	.	.	.	N	.

FATHMM	PROVEAN	MetaSVM	MetaLR	MetaRNN	M-CAP
.	.	.	.	.	.

MutPred	MVP	gMVP	MPC	PrimateAI	DEOGEN2
.	.	.	.	.	.

BayesDel_addAF	BayesDel_noAF	LIST-S2	ESM1b	EVE_Class10	EVE_Class20
D	D	.	.	.	.

EVE_Class25	EVE_Class30	EVE_Class40	EVE_Class50	EVE_Class60	EVE_Class70
.	.	.	.	.	.

EVE_Class75	EVE_Class80	EVE_Class90	Aloft	DANN	Fathmm-MKL
.	.	.	Recessive	D	D

Fathmm-XF	GERP++	PhyloP-vertebrates	PhyloP-mammalian	PhyloP-primate	PhastCons-vertebrates
N	D	T	N	T	T

PhastCons-mammalian	PhastCons-primate
D	D

Criteria applied

Name	Description	Result
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease	FAIL
BP4	Multiple lines of computational evidence suggest no impact on gene or gene product	FAIL
BP7	Synonymous variant with no splicing impact and not conserved	FAIL
BP3	In frame deletion/insertion in a repetitive region	FAIL
PP3	Multiple lines of computational evidence support a deleterious effect	FAIL

Figura 19. *Ejemplo del apartado de información computacional y predictiva del reporte.*

- *Bibliography*. Lista de publicaciones científicas en las que se ha estudiado la variación detallada. Son referencias que apoyan la interpretación de los resultados (Figura 20).

Bibliography

The following publications mention the variant:

- MD Ardell, DL Bedsole, RV Schoborg, SJ Pittler. Genomic organization of the human rod photoreceptor cGMP-gated cation channel beta-subunit gene. Gene 245:311-318 (2000).
- C Bareil, CP Hamel, V Delague, B Arnaud, J Demaille, M Claustres. Segregation of a mutation in CNGB1 encoding the beta-subunit of the rod cGMP-gated channel in a family with autosomal recessive retinitis pigmentosa. Hum. Genet. 108:328-334 (2001).

Figura 20. *Ejemplo del apartado de bibliografía del reporte.*

La plataforma DELFOS se encuentra en fase de validación clínica para asegurar su eficacia y se mejora continuamente. Esta plataforma representa un avance en la gestión y análisis de datos genómicos y permite una mejor comprensión de las variaciones genómicas y su impacto clínico.

5. DISEÑO

En esta sección se va a diseñar un sistema de análisis de variaciones genómicas en el contexto de las distrofias hereditarias de retina para cumplir con la primera parte del segundo objetivo del TFM. Para poder diseñar de forma correcta el pipeline, se ha realizado una reunión con profesionales clínicos del Hospital la Fe de Valencia. Estos expertos indicaron las necesidades actuales en el campo como el conocimiento de información sobre regiones reguladoras y sobre genes específicos de las DHR. Estas necesidades fueron:

- Identificación de genes expresados en DHR, para comprender mejor su rol en la enfermedad. Incluyendo la base de datos de RetNet.
- Estudio regiones regulatorias, ya que pueden influir significativamente en el fenotipo de las DHR.
- Búsqueda predictores y metapredictores que puedan dar información adicional sobre las variaciones genéticas.
- Documentación y análisis de las variaciones, requiriendo documentación exhaustiva de variaciones de DHR, incluyendo estrategias de filtrado y clasificación. Personalización del filtrado y clasificación
- Ajuste de criterios, debe permitir personalizar y ajustar criterios en función de las necesidades específicas del análisis.

Este apartado abarca la recolección de información en literatura científica de las diferentes fases de las que consta el pipeline propuesto, en este caso enfocadas a las DHR. Esta etapa del trabajo se ha centrado en la búsqueda bibliográfica necesaria para adaptar los pasos a las retinopatías.

5.1. Preparación de los datos

Para anotar las variaciones se necesitarán previamente los datos. Para obtenerlos se pueden usar predictores que sirven para descubrir nuevas regiones reguladoras en el ADN, nuevas zonas de splicing y patogenicidad, o de bases de datos. En este apartado se hablará de una revisión de la literatura científica sobre los predictores existentes y los que pueden usarse para las DHR y sobre distintas bases de datos que contienen la información necesaria.

5.1.1. Predictores

Como se ha especificado al comienzo de la sección se van a desarrollar los puntos específicamente sobre las DHR. Para comenzar se hace necesaria una revisión de literatura científica, desde predictores de motivos reguladores hasta la predicción de patogenicidad y de *splicing*. Esto permite conocer la amplia gama de herramientas que se pueden usar. Para poder hacer esta revisión bibliográfica se realizaron búsquedas específicas tanto en PubMed como en Google scholar, estas búsquedas fueron: "*regulatory motif prediction", "pathogenicity predictors", "splicing predictors" AND "genetic disorders"* o *"SpliceAI" "splicing prediction".* El principal objetivo era encontrar predictores que dieran datasets precalculados de los elementos regulatorios, de esta manera no se calcula nada extra sino que se hace uso de información predefinida. No se van a hacer estos cálculos ya que requiere mucha capacidad de cómputo y además los anotadores solo aceptan datos ya generados.

Para comenzar, en cuanto a predictores de regiones reguladoras y de conservación encontramos en primer lugar MEME (Bailey et al., 2006) es una herramienta capaz de descubrir motivos o conjuntos de proteínas o ADN a través del uso del algoritmo de EM (maximización de expectativas), que produce modelos de secuencias probabilístico que representan los motivos encontrados. Por otro lado, BioProspector (Liu et al., 2000) es un conocido predictor que puede identificar motivos de ADN conservados en regiones reguladoras haciendo uso de muestreo de Gibbs y modelos de Márkov. Mencionar también MotifSampler (Thijs et al., 2001), este va a tener modificaciones en el muestreo de Gibbs que van a permitir una estimación del número de copias del motivo en una secuencia además también localiza motivos en el genoma con el uso específico de otra herramienta secundaria denominada MotifLocator (Claeys et al., 2012). Por último, GAME (Wei & Jensen, 2006) usa un algoritmo genético que permite encontrar los motivos óptimos en secuencias de ADN y desarrolla motivos de alta aptitud desde una población aleatoria.

Aunque estos predictores pueden ser relevantes, no se han podido implementar en el trabajo, ya que tienen restricciones de licencias y disponibilidad de datos. Además, estos no daban *datasets* precalculados de los elementos regulatorios que era lo que se buscaba.

A continuación, se exponen predictores de patogenicidad de los cuales en la siguiente sección se especificará cuales se usarán en el modelo.

PhyloP (Pollard et al., 2010) es el primer predictor que se va a tratar, este es capaz de medir la conservación evolutiva en sitios de alineación individuales, cuanto con puntuaciones positivas que indican conservación y puntuaciones negativas que indican cambio evolutivo. Por otro lado, está Grantham (Grantham, 1974) que genera información sobre la variabilidad bioquímica entre aminoácidos haciendo uso de la distancia evolutiva y evalúa el impacto de las variaciones en la funcionalidad de las proteínas. CADD es capaz de integrar variaciones naturales con variaciones simuladas para identificar variaciones causales. PolyPhen-2 va a predecir el impacto de una sustitución de aminoácidos en la estructura y función de una proteína humana con diferentes puntuaciones (Adzhubei et al., 2013b). Por último, se encuentra SIFT que predice si una sustitución de aminoácidos afecta o no a la función de una proteína. Todos los scores de estas herramientas se pueden suponer como predictores ya que pueden predecir un posible impacto funcional de las variaciones genéticas en las proteínas.

Por otro lado, para hacer un correcto análisis de las DHR se pueden a hacer uso de diferentes predictores que nos puedan indicar el posible impacto de las variaciones en los mecanismos de *splicing*. El *splicing* es un mecanismo celular que consiste en la eliminación de los intrones y la unión de los exones, por lo que un impacto posible sobre este mecanismo puede afectar de nuevo a la funcionalidad de las proteínas.

Es importante conocer cómo se regula el *splicing*, ya que se estima que las variaciones genéticas que lo causan cuando no se debe podrían representar hasta el 50% de todas las mutaciones que dan lugar a una disfunción del gen (Cartegni et al., 2002).

En pacientes afectados de DHR no solo se han descrito mutaciones en las secuencias canónicas de *splicing*, sino también en secuencias no canónicas, consideradas como el primer y los 3 últimos nucleótidos de un exón, de hecho, se estima que la prevalencia de variaciones de *splicing* en sitios no canónicos en pacientes con DHR es del 3,5% (Soens et al., 2017). Por lo que se encontraron datos que evidencian la importancia de estos predictores en el marco de las DHR.

Dentro de estos predictores de *splicing* se encuentra en primer lugar, Human Splicing Finder (Desmet et al., 2009) este es una herramienta para predecir los efectos de las variaciones en las señales de *splicing* o para identificar motivos de *splicing* en cualquier secuencia humana. Además de ser un predictor HSF cuenta con una base de datos per-computada que se diseñó para incluir los intrones y exones de todos los genes humanos. Los genes se crearon a partir del conjunto de datos brutos utilizando tanto las coordenadas de los transcritos de Ensembl como las secuencias de la base de datos del navegador del genoma de la UCSC. Actualmente, la base de datos HSF solo contiene genes humanos.

GeneSplicer (Pertea, 2001), se usa en bioinformática para la predicción de genes y la identificación de sitios de *splicing* (*splice sites*) en secuencias de ADN o ARN. Utiliza algoritmos y modelos estadísticos para identificar secuencias de nucleótidos que actúan como sitios de *splicing*, estos sitios son críticos porque determinan dónde se corta y se unen los intrones.

Por último, SpliceAI (de Sainte Agathe et al., 2023) hace uso de redes neuronales profundas para analizar secuencias de ADN y predecir cómo las variaciones genéticas afectan al *splicing*. Este modelo se entrena con grandes conjuntos de datos de secuencias de ADN y sitios de *splicing* conocidos de esta manera la red neuronal aprende características de las secuencias que alteran el *splicing*. Para indicar la probabilidad de que una variación afecte o no al *splicing* genera una serie de puntuaciones. Es muy preciso por usar modelos de aprendizaje profundo, lo que permite una mejor identificación de variaciones. De este predictor se hizo un estudio en profundidad para entender las variables y los pesos de los que hacía uso. Sin embargo, no se ha implementado en el pipeline debido a que necesita costes computacionales muy elevados debido a los modelos de aprendizaje y los tiempos de procesamiento que necesita.

En cuanto a los predictores mencionados es importante destacar que en el pipeline general del grupo de investigación ya se encuentran implementados varios como PhyloP, CADD, PolyPhen-2 y SIFT, por lo que no será necesario volver a implementarlos. Sin embargo, Grantham y SpliceAI aún no están implementados y podrían considerarse para futuros análisis.

5.1.2. Selección de bases de datos

En este apartado se detallará el proceso de selección y anotación de variaciones genómicas usando bases de datos relevantes para enfermedades hereditarias de la retina. Para la recopilación de datos y la investigación del problema, se ha realizado de nuevo una extensa búsqueda en la literatura científica haciendo uso de *PubMed* y *Google Scholar*. Algunas búsquedas exactas fueron: en PubMed *"retinal hereditary diseases" AND "genomic databases"*, *"promoter predictors" AND "databases" y* en Google Scholar *"non-coding variants" "retinal diseases"* o *“gene expression databases" "retinal diseases".*

Con esta búsqueda se encontraron diferentes bases de datos que se pueden ver en la Tabla 3 con la información del proyecto, variaciones, motivos, regiones reguladoras, etc.

Base de datos	Descripción	Referencia
UCSC	Metabase de datos que contiene elementos como TFBS, modificación de histonas, promotores, *enhancers* y *silencers*. Incluye bases de datos diferentes como ORegAnno, NCBI, GeneHancer y VISTA	(Kent et al., 2002)
GTRD	Contiene sitios de unión de fatores de transcripción (TF) identificados mediante experimentos de ChIP-seq.	(Yevshin et al., 2019)
DENdb	Integra información predicha de enhancers en diferentes líneas celulares.	(Ashoor et al., 2015)
CTCFBSDB	Integra información de sitios de unión CTCF, tanto determinados experimentalmente como predichos	(Ziebarth et al., 2012)
EPDnew	Contienen información de promotores reconocidos por la RNA polimerasa	(Dreos et al., 2015)
ORegAnno	Muestra regiones regulatorias seleccionadas por la literatura, sitios de unión de factores de transcripción y polimorfismos regulatorios	(Griffith et al., 2007)
VISTA enhancers	Muestra enhancers potenciales cuya actividad fue validada experimentalmente.	(Visel et al., 2007)
TRANSFAC	Contiene datos experimentales de TFs eucariotas, sus binding sites, secuencias de consenso y genes regulados	(Matys, 2003)
NCBI RefSeq Functional Elements	Incluye regiones reguladoras de genes verificadas experimentalmente, elementos estructurales conocidos, orígenes de replicación de ADN y sitios clínicamente significativos de recombinación de ADN e inestabilidad	(Pruitt, 2004) (Pruitt et al., 2014)
RetNet	Incluye información sobre todos los genes que al presentar determinadas variaciones pueden generar DHR	(Roy et al., 2023)

***Tabla 3**. Recopilación de bases de datos relevantes.*

En cuanto a las bases de datos, en este trabajo se van a seleccionar las dos bases de datos más relevantes de las mostradas en la Tabla 3, una de genes y otra de secuencias reguladoras que es lo que más le interesaba a los expertos, estas se podrán ver en la siguiente sección.

Así, el uso de predictores o bases de datos depende de si se busca realizar nuevos descubrimientos o trabajar con información ya contrastada. Algunas bases de datos dan información sobre regiones regulatorias, mientras que otras detallan elementos como silenciadores, potenciadores o sitios de modificación de histonas, que afectan a la activación o represión de genes, elementos reguladores.

5.2.Anotación

Tras la recopilación de información y preparados los datos necesarios, el siguiente paso del pipeline es la anotación de variaciones para proporcionar contexto sobre su posible impacto biológico. Con esta se puede conseguir priorizar las variaciones que pueden tener un impacto negativo en el fenotipo de las DHR.

Para llevar a cabo este proceso de anotación van a ser necesarias un conjunto de bases de datos especializadas que ofrecen diferentes tipos de información relevante, en la siguiente Tabla 4 se describen las que están ya implementadas en el pipeline original y tras estas se mencionan las dos que se añaden a este, que son específicas para DHR y que se han conseguido gracias a la revisión bibliográfica en la anterior sección.

Bases de datos	Descripción	Referencias
ClinVar	Tiene información de la relación entre variaciones genómicas y enfermedades.	(Landrum et al., 2018)
ClinGen	Tiene anotaciones de genes y variaciones asociadas a la clínica	(Rehm et al., 2015)
GWAS Catalog	Repositorio de estudios de asociación del genoma completo (GWAS), relaciona variaciones con fenotipos o enfermedades	(Sollis et al., 2022)
gnomAD Constraints	Contiene información sobre frecuencia poblacional de las variaciones y su relevancia evolutiva	(Karczewski et al., 2020)
dbNSFP	Incluye predicciones funcionales y datos de conservación	(Liu et al., 2016)
dbscSNV	Contiene anotaciones de variaciones en regiones de splicing, con predicciones de las variaciones en estos puntos	(Liu et al., 2011)
ClinGen Gene	Incluye información sobre la relación de los genes y enfermedades	(Rehm et al., 2015)
NCBI Gene	Tiene información estructural y funcional de genes, con ubicación y funciones de los mismos	(Brown et al., 2015)
UniProt Domains, Mutagenesis, Variants, Repeats	Incluye información de dominios proteicos, variaciones de aminoácidos y secuencias repetitivas.	(UniProt Consortium, 2023)

Tabla 4. *Tabla BBDD incluidas para la anotación en el pipeline general.*

Las nuevas bases de datos que se deben incluir en el pipeline para que sea específico a las DHR son:

- NCBI RefSeq, la cual proporciona información sobre elementos no codificantes como promotores, enhancers y regiones reguladoras. Que como se ha visto son importantes para las DHR. Esta base de datos permite hacer anotaciones que pueden ayudar a identificar variaciones.
- RetNet, es una base de datos en la cual aparecen los genes y variaciones asociadas a las DHR, de esta manera proporciona información específica de los genes.

De esta manera con este diseño se consigue una anotación de variaciones especifica de las DHR.

5.3. Filtrado

Siguiendo con el esquema del pipeline el filtrado es el siguiente paso. En el caso de las DHR como en todas las enfermedades genéticas es fundamental para eliminar aquellas variaciones que no son relevantes. En este apartado se incorporan de manera teórica los filtros generados explicados en la sección 4 pero adaptados a particularidades de las DHR.

Una vez realizada la implementación de las bases de datos se realizan las modificaciones necesarias en la fase de filtrado de las variaciones específicas de las DHR.

Por un lado, la forma de filtrado por herencia no se va a poder implementar debido a que no se tienen disponibles genomas de familiares no afectados o de tejido normal compatible del mismo individuo. En cuanto al filtro de frecuencias se hace uso del estándar ya establecido e implementado en el modelo.

Genes de interés

El principal filtro para obtener las variaciones específicas en el pipeline es el de los genes de interés. Para hacerlo se va a hacer uso de la información detallada en la base de datos de RetNet, es decir se van a utilizar los genes que aparecen en esta base de datos ya que son los que están relacionados con estas enfermedades.

Aunque el número de genes implicados en las DHR es elevado, solo 21 de ellos afectan a 30 o más familias (1%), mientras que los restantes afectan a una o a un número muy limitado de familias (Pérez-Romero et al., 2022). Esto indica que las DHR no están asociadas a variaciones en un único gen, sino que múltiples genes pueden estar involucrados en el desarrollo de la misma patología.

Los siete genes más frecuentemente implicados afectan al 52% de las familias caracterizadas (Figura 21), siendo estos: ABCA4 (22%), USH2A (12%), CRB1 (4%), PRPH2 (4%), RS1 (4%), RPGR (3%) y RHO (3%)(Pérez-Romero et al., 2022).

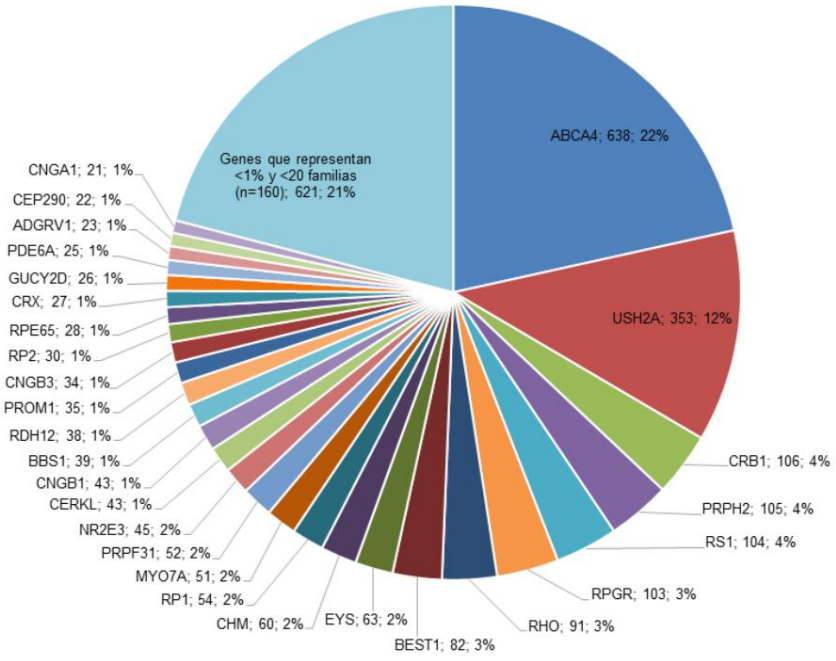


Figura 21. *Porcentajes de los genes que están implicados frecuentemente en las DHR. Extraída de (Pérez-Romero et al., 2022).*

Para esto, en el modelo se añadieron estos genes y aquellas variaciones que no estuvieran en estos eran descartadas. La lista de los genes usados se puede ver en el Anexo en LISTADO DE GENES.

Consecuencia

Como anteriormente se ha dicho las DHR tienen una alta heterogeneidad genética, con 5.064 variantes patogénicas causales observadas. Estas fueron mayoritariamente *missense* (51%), seguidas por las truncantes (*nonsense*, *frameshift*, indels y de *splicing*) (44%) (Pérez-Romero et al., 2022).

Las variaciones de truncamiento y empalme (*splicing*) son de interés primordial en los estudios relacionados con enfermedades debido a su posible alto impacto celular y sistémico, se han detectado en enfermedades de retina (Cremers et al., 1998).

Las variaciones *missense* también pueden alterar sitios de *splicing* o de expresión y como se ha visto son mayoritarias. En el caso de las retinopatías se ha visto que este tipo de variaciones están asociadas a una actividad de ATPasa reducida pero no completamente eliminada, mientras que las mutaciones *nonsense* tienen un mayor impacto en la función de las proteínas (Michaelides et al., 2003).

Por otro lado, las variaciones sinónimas pueden descartarse ya que es menos probable que tengan un efecto relevante.

Por lo que en este tipo de filtro en el caso de las DHR se tienen que seleccionar las variaciones de tipo truncamiento y *splicing*, las variaciones *missense* y las variaciones *nonsense*.

Regiones reguladoras

Por otro lado, la base de datos de NCBI refseq también se va a usar como filtro de variaciones, puede ser útil ya que en específico una de las causas de las distrofias hereditarias de retina son las variaciones en regiones reguladoras. Esta base de datos al contar con información sobre estas regiones se usará como criba.

De esta manera, si aparece alguna variación en estos elementos reguladores que contiene la base de datos NCBI refseq, entonces se seleccionan estas variaciones para un futuro análisis más detallado de las mismas.

Un ejemplo general hipotético de este tipo de filtrado puede ser: se cuenta con un conjunto inicial de 10000 variaciones. Se le aplican dos filtros, el de genes de interés (retiene variaciones en genes específicos causantes de la enfermedad) y el de regiones reguladoras (retiene variaciones dentro de regiones reguladoras importantes cerca de los genes de interés). Con el primer filtro de genes de interés pasan 500 variaciones y con el segundo filtro de regiones reguladoras pasan 200. En total se tendría 700 variaciones que hay que analizar y clasificar.

Se obtienen las variaciones que en principio se interesan reduciendo así el número a analizar, siendo este el principal objetivo de la etapa de filtrado.

5.4. Clasificación

Como se ha expuesto a lo largo del trabajo, la clasificación de las variaciones genómicas es esencial para identificar las mutaciones responsables de una enfermedad y guiar el desarrollo de terapias personalizadas en específico relacionado con las distrofias hereditarias de retina. Para abordar este desafío, se ha adoptado un modelo desarrollado por el grupo PROS, que ofrece estandarizaciones para la clasificación de variaciones de ADN.

Para adaptar el modelo desarrollado en la anterior sección a las distrofias hereditarias de retina, se ha necesitado hacer una serie de modificaciones. Estas modificaciones incluyen criterios y métricas específicos para las DHR, basados en las frecuencias de corte encontrados en ClinGen en el apartado de criterios de clasificación específicos relacionados con el gen RPE65. Este gen es importante ya que la presencia de determinadas variaciones puede causar LCA que es un tipo de distrofia hereditaria de retina.

Se decidió centrar los criterios entorno a este gen ya que actualmente no hay suficiente información sobre otros genes que generan las DHR. Los grupos SVI de ClinGen están desarrollando más criterios específicos para otros genes relacionados, por lo que en un futuro habrá más criterios específicos disponibles y la clasificación se verá mejorada.

En especial, los cambios realizados son en los criterios de frecuencia poblacional y genes de interés. En el criterio de los genes de interés se va a hacer uso de los genes obtenidos en la base de datos de RetNet y en cuanto a los criterios de frecuencia poblacional se va a hacer uso de los umbrales de frecuencia establecidos en la guía de clasificación de variaciones específicas del gen RPE65.

Finalmente, tras todas las implementaciones en el modelo adaptado se consigue una recolección de datos genómicos relevantes de bases de datos especializadas en DHR, la evaluación y clasificación de las variaciones genómicas usando los criterios y métricas adaptados, y la aplicación de herramientas automatizadas para procesar y anotar las variaciones.

Finalmente se puede ver como este modelo presenta numerosos beneficios proporcionando una estandarización que permite adaptarlo fácilmente a cualquier enfermedad en este caso a las distrofias hereditarias de retina. De esta manera se puede facilitar el desarrollo de la medicina personalizada. Con todo esto se mejora la eficiencia y la precisión del proceso que se ha ido exponiendo en los diferentes apartados del TFM.

6. DESARROLLO

En esta sección de desarrollo se tratarán los cambios realizados en el pipeline o modelo original del oráculo genómico de DELFOS para adaptarlo correctamente a las distrofias hereditarias de retina logrando cumplir con el segundo objetivo. Se va a exponer la selección de las bases de datos con las que se va a trabajar y la implementación de las mismas, los filtros que se usan para la parte del filtrado y los cambios que se deben hacer en los criterios de clasificación de variaciones.

6.1.Anotación

Una vez establecidas las bases de la anotación en la sección 4 y en la 5, se tratará como se abordará principalmente en el caso de las DHR.

6.1.1. RetNet

Para comenzar, en cuanto a la implementación de las bases de datos se va a explicar primero el tratamiento de RetNet, en la Figura 22 se puede ver la interfaz de la página web de esta misma. Los datos que interesan en este proyecto son los de *Diseases* y *References*.

En *Diseases* encontramos información de cromosoma, gen, localización del gen dentro del cromosoma, la enfermedad que puede causar, cómo identificarlo y un link de referencias, por otro lado, en *References* encontramos la referencia completa de los estudios donde se encuentra la evidencia.

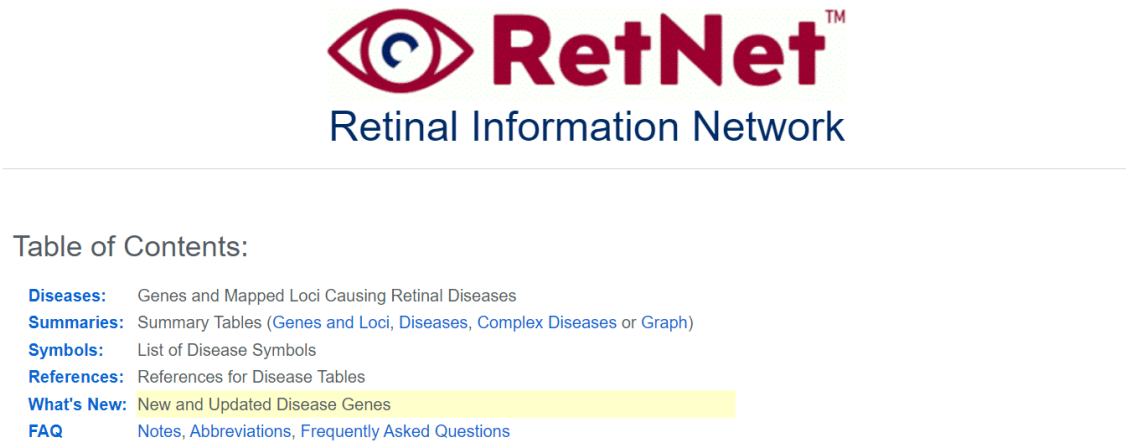


Figura 22. *Página web de la base de datos de RetNet. Extraído de (*RetNet - Retinal Information Network*, 2024).*

Esta base de datos no tiene una opción de descarga directa de los datos, por lo que para poder obtenerlos se usó la herramienta de Google WebScraper.io, esta es una extensión de Google Chrome, que permite una extracción de datos específica de páginas web, como textos, tablas, enlaces y otros elementos. Ofrece una interfaz sencilla sin la necesidad de escribir código y los datos extraídos se pueden exportar en diferentes formatos como CSV, JSON o Excel. Por lo que es una herramienta óptima para la extracción automatizada de datos web.

Una vez conocida esta herramienta se ha visto que es óptima para obtener los datos de la web de RetNet, es por esto por lo que se va a hacer uso de ella generando archivos CSV tanto de los datos de *Diseases* como los de *References*, se van a generar dos archivos CSV diferentes, el motivo es que en la web de la base de datos esta información se encuentra separada como se puede observar en la Figura 22 y se necesitan ambas para poder hacer una correcta implementación.

En la Figura 23 se muestran como están dispuestos los datos de *Diseases* en la web, como se puede ver las referencias no están completas y en nuestro caso se necesita que estén completas.

Chromosome 1				
Symbols; OMIM Numbers	**Location**	**Diseases; Protein**	**How Identified; Comments**	**References**
SAMD11; 616765	1p36.33	Recessive retinitis pigmentosa; sterile alpha motif domain containing 11 protein; [Gene]	homozygosity mapping, whole-exome sequencing.; a single homozygous *SAMD11* p.Arg630* nonsense mutation found in five affected members of two unrelated consanguineous Spanish families with recessive RP; the SAMD11 transcript and protein are abundant in retina; the transcript is present in the early mouse embryo and widely expressed with 45 alternate splice variants; only ocular symptoms were documented in affected individuals; the SAMD11 protein is involved in signal transduction and regulation of transcription and interacts with CRX	Corton 16
NPHP4, SLSN4; 606966, 606996, 607215	1p36.31	recessive Senior-Loken syndrome; recessive nephronophthisis, juvenile; nephronophthisis 4 protein; [Gene] [ClinGen]	linkage mapping, candidate genes; Senior-Loken syndrome involves cystic kidney disease (nephronophthisis) and retinitis pigmentosa or Leber congenital amaurosis; Jobert syndrome is the same with additional cerebellar and cognitive abnormalities; NPHP4 protein interacts with NPHP1 protein	Mollet 02, Otto 02, Schuermann 02

***Figura 23**. Ejemplo de tabla de Diseases. Extraído de (*RetNet - Retinal Information Network*, 2024).*

A continuación, en la Figura 24 se muestran como están dispuestos los datos en *References* y cuál es la referencia completa.

References

[Home | Disease Genes and Loci | Summaries | Symbols | FAQ | Contact Us]

Listed alphabetically by author:

[A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z]

A

- MM Abd El-Aziz, MF El-Ashry, WM Chan, KL Chong, I Barragán, G Antiñolo, CP Pang, SS Bhattacharya. A novel genetic study of Chinese families with autosomal recessive retinitis pigmentosa. 71:281-294 (2006). [PubMed]
- MM Abd El-Aziz, I Barragán, C O'Driscoll, S Borrego, L Abu-Safieh, JI Pieras, MF El-Ashry, E Prigmore, N Carter, G Antiñolo, SS Bhattacharya. Large-scale molecular analysis of a 34 Mb interval on chromosome 6q: major refinement of the RP25 interval. 72:463-477 (2008). [PubMed]
- MM Abd El-Aziz, I Barragán, CA O'Driscoll, L Goodstadt, E Prigmore, S Borrego, M Mena, JI Pieras, MF El-Ashry, LA Safieh, A Shah, ME Cheetham, NP Carter, C Chakarova, CF Ponting, SS Bhattacharya, G Antinolo. *EYS*, encoding an ortholog of *Drosophila* spacemaker, is mutated in autosomal recessive retinitis pigmentosa. 40:1285-1287 (2008a). [PubMed]
- A Abid, M Ismail, SQ Mehdi, S Khaliq. Identification of novel mutations in *SEMA4A* gene associated with retinal degenerative diseases. 43:378-381 (2006). [PubMed]

***Figura 24**. Ejemplo datos de references. Extraído de (*RetNet - Retinal Information Network*, 2024).*

Una vez extraídos los datos, se transforman con el objetivo de obtener las columnas que son relevantes (Tabla 5), en este caso las columnas seleccionadas se pueden ver totalmente desarrolladas en la Tabla 6. Para realizar estas trasformaciones se va a hacer uso de códigos escritos en Python.

Columna
Symbols; OMIM Numbers
Diseases; Protein
How Identified; Comments
References

***Tabla 5**. Columnas seleccionadas para trabajar con su información de la base de datos de RetNet.*

Seleccionadas las columnas relevantes, se generan las mejoras esperadas a cada una, a continuación, se explican estas mejoras y se inserta una tabla ejemplo del resultado del tratamiento de las columnas que se puede ver en la Tabla 6:

1. Separar de OMIM y Genes: Los '*OMIM Numbers*' se deben separar de los '*Symbols*' de los genes, eliminándose de esta forma la información de '*OMIM Numbers'* y seleccionando solo el primer gen de cada lista. Por ejemplo, en la Figura 21 en esta columna se muestra "SAMD11;616765", en este caso se quiere específicamente "SAMD11" en el caso de debajo de este en la figura se obtendría "NPPHP4".
2. Reformatear *Diseases*: La información de la columna se separó en '*Diseases*' y '*Protein'*, se elimina la información de '*Protein'*. Por ejemplo, de nuevo en la Figua 21 vemos en la columa de diseases: "Recessive retinitis pigmentos asterile alpha motif domain containing 11 protein[Gene]" de aquí se obtiene "Recessive retinitis pigmentosa".
3. Completar referencias: Se utilizaron datos del archivo de las referencias bibliográficas para completar la columna '*References*'. Para esto se desarrolla en el código una búsqueda del nombre que aparece en esta columna y se hace un *match* con la referencia completa que aparezca en el archivo CSV de *references*.
4. Unir comentarios: La columna '*How Identified'* y '*Comments*' se combinó para proporcionar una descripción más completa y fluida.

Columna	Ejemplo
Symbols	SAMD11
Diseases	recessive retinitis pigmentosa
How Identified; Comments	The following methods were used for determining the gene-disease relationship: homozygosity mapping, whole-exome sequencing. The following evidences support the association: a single homozygous SAMD11 p.Arg630* nonsense mutation found in five affected members of two unrelated consanguineous Spanish families with recessive RP. The SAMD11 transcript and protein are abundant in retina. The transcript is present in the early mouse embryo and widely expressed with 45 alternate splice variants. Only ocular symptoms were documented in affected individuals. The SAMD11 protein is involved in signal transduction and regulation of transcription and interacts with CRX.
References	A Avila-Fernandez, M Corton, KM Nishiguchi, N Muñoz-Sanz, B Benavides-Mori, F Blanco-Kelly, R Riveiro-Alvarez, B Garcia-Sandoval, C Rivolta, C Ayuso. Ophthalmology 119:2616-2621 (2012).

***Tabla 6**. Resultado de la mejora de los datos de RetNet.*

Los códigos desarrollados para la implementación de esta base de datos se pueden encontrar en el Anexo en el apartado de CÓDIGOS, tanto el de selección de columnas relevantes como el de mejora de los datos.

6.1.2. NCBI refseq

Por otro lado, la base de datos de NCBI RefSeq es más sencilla a la hora de la descarga de datos, esta se puede descargar desde UCSC Genome Browser en la sección de las bases de datos de regiones reguladoras, en la descarga se genera un archivo con las diferentes columnas con las que cuenta (Tabla 7).

Columna	Ejemplo	Descripción
Chrom	chr1	Cromosoma de referencia
ChromStart	167327716	Posición de comienzo en el cromosoma
ChromEnd	167329809	Posición final en el cromosoma
Name	enhancer	Tipo de elemento
Score	0	Su uso es para formato BED
ThickStart	167627716	Posición de comienzo en el cromosoma
ThickEnd	167329809	Posición final en el cromosoma
Reserved	0, 128, 128	Se usa como itemRgb: dependiendo del tipo de elemento se le asigna un color
SoTerm	enhancer	Termino de *Sequence Ontology*
Note	VISTA enhnacer hs1331	Nota descriptiva del elemento
GeneIds	110121063|GeneID: 110121063,5451|GeneID:5451	Gene ID del elemento asociado al gen
PubMedsIds	17135569|PMID: 17135569	ID de publicaciones relacionadas en PubMed
Experiment	EXISTENCE:transgenic organism evidence [ECO:0001131][PMID: 17135569]	Evidencia experimental
Function	Enhancer in: neural tuve[4/7] | hindbrain (rhombencephalon)[4/7] | midbrain (mesencephalon)[5/7] | forebrain[7/7]	Función predicha
_mouseOver	VISTA enhnacer hs1331 | Enhancer in: neural tuve[4/7] | hindbrain (rhombencephalon)[4/7] | midbrain (mesencephalon)[5/7] | forebrain[7/7]	

***Tabla 7**. Estructura de los datos de la BBDD de NCBI refseq.*

Para esta base de datos se realizará una anotación primaria. Tras obtener el archivo original, se descartan las columnas no relevantes de nuevo haciendo uso de códigos desarrollados en Python, obteniendo así las columnas de interés (Tabla 8).

Columna
Chrom
ChromStart
ChromEnd
Name
Note
PubMedsIds
Experiment
Function

***Tabla 8**. Columnas importantes NCBI RefSeq.*

Una vez más tras obtener estas columnas se realizan una serie de mejoras (Tabla 9) que son:

1. Completar referencias: Se extrajeron IDs de PubMed de la columna '*PubMedsIds*' y se completaron con información detallada utilizando una función específica desarrollada denominada 'get_bibliography'. El objetivo de esta función es recuperar la bibliografía de PubMed haciendo uso de los PMIDs que son los identificadoes de PubMed.
2. Unir información: Las columnas '*note*', '*function'* y '*experiment'* se unificaron en una sola columna '*Info'* para mejorar la accesibilidad de los datos.

Columna	Ejemplo	Descripción
Chrom	chr1	Cromosoma de referencia
ChromStart	167327716	Posición de comienzo en el cromosoma
ChromEnd	167329809	Posición final en el cromosoma
Name	enhancer	Tipo de elemento
PubMedsIds	Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. Roh TY, Wei G, Farrell CM, and Zhao K. Genome Res (2007). pmid: 17135569	Publicación relacionada
Info	The description of the element: VISTA enhancer hs1331. The element function is: enhancer in Jurkat T cells and the experimental evidence is: gene assay evidence.	Información general, incluyendo los columnas de *note*, *function* y *experiment*

***Tabla 9**. Ejemplo resultado de la mejora de las columnas de los datos de NCBI RefSeq.*

De nuevo los códigos desarrollados para la mejora e implementación de esta base de datos se pueden encontrar en el anexo en apartado de CÓDIGOS. Tras la anotación de estas bases de datos en los siguientes apartados se van a explicar para qué se han implementado en el pipeline estas bases de datos en específico y cómo se van a usar.

6.2. Filtrado y clasificación

En este apartado, se detallan los métodos y herramientas empleadas para el filtrado y la clasificación de variaciones genómicas en el contexto de las distrofias hereditarias de retina (DHR).

El proceso se basa en el uso de bases de datos procesadas y anotadas previamente con snpEff y un anotador propio del grupo, que como ya se ha dicho permiten incorporar información detallada sobre cada variación, que se integran en el pipeline utilizando MySQL para la gestión de las bases de datos.

Para automatizar las etapas de filtrado y clasificación se usa VariantInsight, la cual se encarga de procesar el archivo de VCF, esta automatización se puede hacer gracias a las métricas y criterios definidos durante la etapa de diseño del pipeline de DHR.

Los criterios específicos modificados son los relacionados con la frecuencia poblacional y los genes relevantes de las enfermedades de distrofias hereditarias de retina.

En primer lugar, se establecen las métricas relacionadas con el criterio de frecuencia poblacional. La primera métrica se llama freq_greater_than_0.008, indica que si una variación tiene una frecuencia mayor que 0.008 es poco probable sea la causa de las DHR, ya que se supone como una variación común en la población. Tras esta se define una nueva métrica llamada freq_less_than_0.0008, en este caso el umbral es menor de 0.0008 lo que ayuda a identificar variaciones raras que podrían ser potenciales patogénicas.

Por último, en cuanto al criterio de genes relevantes se definió una métrica llamada relevant_gene en la que se priorizan las variaciones que estén en genes relevantes de las DHR, se obtienen a partir del listado de los genes de RetNet, lo que permite focalizar el análisis en variaciones dentro de los genes conocidos que están asociados a estas enfermedades. Un resumen de estas métricas se puede ver en la Tabla 10.

Criterio	**Métrica**	**Valor a cumplir**
Frecuencia alélica poblacional	freq_greater_than_0,008	> 0,008
	freq_less_than_0,0008	<0,0008
Genes relevantes	relevant_genes_HRD	Que estén presentes en la lista de genes que hay en el Anexo de LISTADO DE GENES

Tabla 10. *Criterios y métricas modificados para el filtrado y clasificación de variaciones de DHR.*

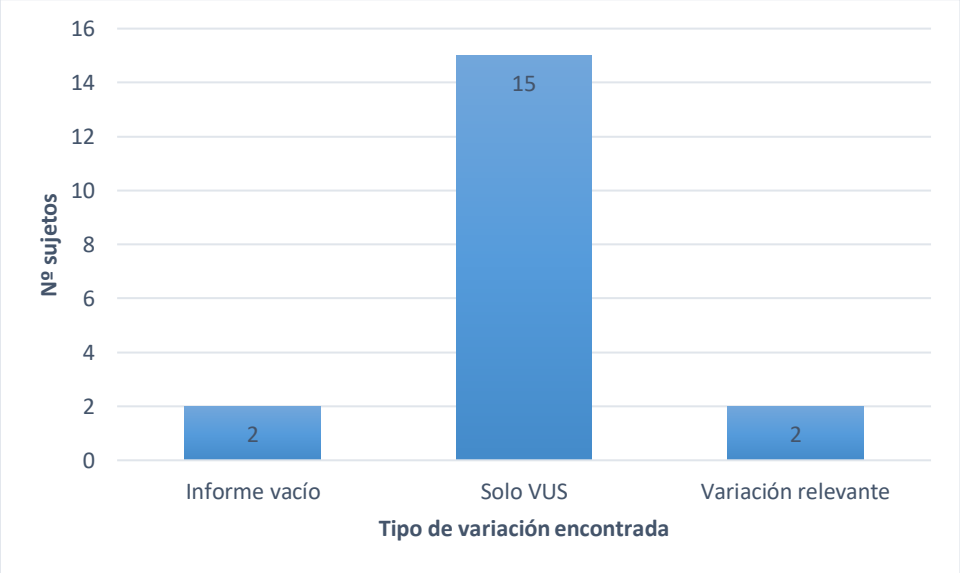
Por lo que la principal ventaja del pipeline es la capacidad de automatizar estas fases en función de los criterios y métricas específicos, generando una clasificación en categorías de patogénica, probablemente patogénica, benigna, probablemente benigna y VUS. Por lo que con la modificación de estas se crea un nuevo pipeline de clasificación específico a una patología en especial diferenciado del estándar que estaba usando el grupo de investigación.

7. CASO DE ESTUDIO Y RESULTADOS

Para abordar el tercer objetivo del presente TFM se realiza un estudio de caso clínico y análisis de resultados. Para poder realizar el caso de estudio se ha hecho uso de archivos VCF o *Variant Call Format.VCF*

El primer paso para poder hacer los casos de estudio es recibir los archivos VCF que contienen variaciones de individuos que padecen alguna distrofia hereditaria de retina, estos archivos VCF fueron proporcionados por el Hospital la Fe de Valencia. Se cargan y procesan en la herramienta del pipeline desarrollado a lo largo del TFM modificado con los criterios y métricas establecidos para las DHR, y se analiza las variaciones de forma automática, este análisis permite determinar si una variación es benigna, de significado incierto o patogénica. La herramienta genera un reporte final detallado para cada variación indicando su clasificación basada en los criterios aplicados. En la Figura 25 se puede ver un ejemplo del resumen de resultados de este reporte.

Para el estudio se han analizado 20 archivos VCF de los que se han obtenido los siguientes resultados. En primer lugar, 2 reportes aparecen vacíos, esto puede indicar que no contenían variaciones relevantes en las regiones de interés o que no cumplían con los criterios establecidos para el análisis, lo que implica que no existe causa genética identificada de las DHR en estos pacientes. Por otro lado, aparecen en 15 de los sujetos variaciones clasificadas como VUS o variación de significado incierto, estas variaciones se han clasificado así ya que no hay suficiente evidencia para determinar si son benignas o patogénicas por lo que requieren una mayor investigación para establecer su relevancia clínica. Es interesante que estas variaciones son la mayoría de tipo *missense*. La clasificación de variaciones como VUS, indica que se necesita más información para determinar su papel en la enfermedad, pero no pueden ser ignoradas. Por último en distintos pacientes, se obtiene 1 variación clasificada como probablemente patogénica (*likely pathogenic*) y 1 patogénica (*pathogenic*) por lo que estas dos variaciones cumplían con las métricas establecidas y son consideradas como variaciones relevantes (Gráfica 1).



Gráfica 1. *Tipos de variaciones encontradas en los sujetos.*

Obtenida la variación clasificada como patogénica se realiza una búsqueda en la literatura científica para identificar si hay estudios previos que hayan asociado esta variación con la enfermedad.

En primer lugar, la herramienta clasifica como patogénica a la variación c2041C>T de tipo *stop gained*[5] del gen ABCA4 (Figura 25). Esta variación ha sido confirmada en la literatura como relevante.

Summary of the results

Gene	Transcript	DNA/Protein alteration	Consequence	Zygosity	Classification
ABCA4	ENST00000370225	c.2041C>T, ENST00000370225	stop_gained	A/G	pathogenic

Figura 25.*Resumen de los resultados del reporte de variaciones en este caso la clasificación de la variación patogénica.*

En el estudio de '*Mutation Spectrum of the ABCA4 Gene in 335 Stargardt Disease Patients From a Multicenter German Cohort—Impact of Selected Deep Intronic Variants and Common SNPs'* (Schulz et al., 2017), se explica que este es un gen muy propenso a desarrollar variaciones que generen un tipo de DHR denominada enfermedad de Stardgardt (STGD1). En los análisis funcionales obtienen que son las variaciones de tipo *stop gained* y *frameshift* las que más asociadas están con este tipo de enfermedades, esto se debe a que se asocian a una acumulación de proteínas malformadas y la pérdida de función de transporte de moléculas en los conos y bastones de la retina contribuyen a la degeneración progresiva. Por lo que esta evidencia científica respalda la clasificación de la variación como patogénica y asociada a la enfermedad de Stargardt.

Por otro lado, se obtiene la clasificación de la variación c.133G>T de tipo *stop gained* del gen CNGB3 como probablemente patogénica (Figura 26). De nuevo esta clasificación también se ve respaldada por literatura científica.

Summary of the results

Gene	Transcript	DNA/Protein alteration	Consequence	Zygosity	Classification
CNGB3	ENST00000320005	c.133G>T, ENST00000320005	stop_gained	A/C	likely pathogenic
CEP290	ENST00000309041	c.3611A>G, ENST00000309041	missense_variant	C/T	VUS

Figura 26. *Resumen de los resultados del reporte de variaciones en este caso la clasificación de la variación probablemente patogénica.*

[5] Una variación de *stop gained* es una variación que va a generar un codón de parada en una zona donde no debería estar, de esta manera se alteran las proteínas generadas.

En el estudio '*Accessory heterozygous mutations in cone photoreceptor CNGA3 exacerbate CNG channel–associated retinopathy'* (Burkard et al., 2018), se detalla cual es el papel del gen CNGB3 en la achromatopsia (ACHM), que es una condición ocular rara, siendo un tipo de DHR. Este gen codifica una subunidad esencial del canal iónico de CNG en los conos de la retina, por lo que variaciones en CNGB3, en especial las de tipo *stop gained* interrumpen la función del canal CNG. En sus estudios funcionales se explica que la mayoría de las variaciones de CNGB3 que generan esos codones prematuros están asociadas con el desarrollo de ACHM, los hallazgos están corroborados por estudios funcionales y modelos animales. Esta clasificación también está respaldada en la literatura, por lo que el resultado obtenido de la herramienta generada es consistente.

Por lo que, la combinación de una búsqueda en literatura científica, la implementación de bases de datos, el estudio de herramientas bioinformáticas y la programación en Python ha permitido desarrollar un pipeline para la clasificación de variaciones genéticas en distrofias hereditarias de retina y el mismo es capaz de aportar resultados en línea con el conocimiento actual de esta enfermedad.

8. DISCUSIÓN

Con el avance que se ha dado estos últimos años en el campo de la genética y las técnicas de secuenciación de nueva generación (NGS), se ha hecho posible la identificación de forma precisa de todas las variaciones de un individuo. Esto permite obtener grandes cantidades de información, lo cual es beneficioso y a la vez un desafío.

Por un lado, se obtiene un elevado beneficio ya que se pueden identificar una gran cantidad de variaciones que podrían asociarse a una determinada enfermedad, esto puede permitir desarrollar un diagnóstico efectivo. Por otro lado, el volumen de información obtenida puede incluir variaciones que no provocan la enfermedad, por lo que se deben desarrollar algoritmos y métodos bioinformáticos complejos para clasificarlas.

En este TFM el objetivo principal ha sido desarrollar un pipeline que permita clasificar variaciones genéticas asociadas a las distrofias hereditarias de retina, ya que actualmente no existe ninguna herramienta específica, lo que se identificó revisando el estado del arte y consultando a profesionales del Hospital de La Fe de Valencia. Para cumplir el objetivo se ha necesario realizar una búsqueda bibliográfica donde se encuentren métricas, criterios, bases de datos y predictores específicos, y la implementación de herramientas para anotar, filtrar y clasificar variaciones. La implementación de estas herramientas se ha realizado haciendo uso de Python y librerías como Pandas, facilitando la manipulación y análisis de los datos.

El pipeline desarrollado se diferencia de otros en algunos aspectos clave como que incorpora bases de datos específicas de DHR y elementos reguladores, lo que permite un mejor enfoque sobre la enfermedad. Además, se han personalizado las etapas de filtrado y clasificación facilitando la adaptación a los nuevos datos y conocimientos que surjan de las DHR y permite que sea una herramienta capaz de adaptarse a cualquier enfermedad según las métricas y criterios se establezcan. Por otro lado, al centrarse en regiones reguladoras entre otras, ofrece una perspectiva más amplia de las posibles causas de DHR, superando limitaciones de otros pipelines que no se centran en estas.

El desarrollo del pipeline se encuentra dentro del marco de la medicina precisión y podría mejorar significativamente el diagnóstico y manejo de DHR. Durante su desarrollo se hizo un caso de estudio que permitió evaluar en primera instancia el pipeline en un entorno controlado. Puede identificar variaciones patogénicas específicas y distinguirlas de las benignas y de las VUS. Además, como se ha visto en la anterior sección aquellas variaciones clasificadas como patogénicas tienen una clasificación correcta y en cuanto a las VUS es necesario un estudio más detallado y continuo.

Algunas limitaciones del trabajo es que la validación solo se ha basado en un caso de estudio y aunque los resultados hayan sido buenos el análisis tiene que ser más exhaustivo con un mayor conjunto de datos para poder estimar de manera precisa la precisión. Además, la implementación de más bases de datos a las que se han implementado es fundamental para tener actualizada la herramienta.

Por otro lado, la colaboración con el grupo PROS del Instituto Valenciano de Investigación en Inteligencia Artificial fue esencial para poder desarrollar este pipeline y por tanto este TFM. A lo largo del proyecto, ha sido necesaria una adaptación a los tiempos de desarrollo del software y a las recomendaciones de los expertos clínicos para una correcta implementación.

9. CONCLUSIÓN

Los objetivos definidos en el TFM se han ido abordando uno a uno. En primer lugar, para cumplir con la investigación del problema se ha hecho uso de herramientas como Google Scholar y PubMed para hacer una revisión de literatura científica y llegar a comprender mejor la genética y las técnicas actuales del análisis de las variaciones de las DHR.

En segundo lugar, para cumplir con el diseño y desarrollo del pipeline se han aplicado los conocimientos adquiridos durante el desarrollo de la investigación del problema, se han implementado las diferentes bases de datos, métricas y criterios relevantes mediante el uso de Python y las diferentes herramientas desarrolladas por el grupo.

Por último, para cumplir con el tercer objetivo de aplicación clínica y futuras direcciones, se realizó una validación preliminar haciendo un caso de estudio. Se ha demostrado que tiene potencial para clasificar variaciones y servir como herramienta de apoyo. Sin embargo, aún queda hacer una validación total colaborando con los expertos clínicos y con más sujetos.

En conclusión, se vio que el correcto análisis de las variaciones genéticas implica una clasificación precisa de estas, lo que puede dar como resultado un mejor diagnóstico a determinadas enfermedades genéticas, con una medicina personalizada desarrollada. En este caso se ha llevado a cabo un pipeline bioinformático para el análisis de las variaciones de las DHR ya que estas enfermedades aún no se habían tratado. Esto es necesario para que las variaciones causantes de estas tengan una correcta clasificación, en la actualidad aún hay una temprana investigación en cuanto a los criterios de clasificación específicos. Este proyecto ha permitido procesar grandes volúmenes de datos de secuenciación de manera sistemática, robusta y eficiente, logrando una correcta anotación, filtrado y clasificación de variaciones genéticas específicas a las DHR.

La implementación de bases de datos específicas, la búsqueda bibliográfica exhaustiva y los criterios de clasificación adaptados a las distrofias hereditarias de retina es lo que ha permitido el correcto desarrollo del proyecto por lo que es a lo que más se le ha dado importancia. Para poder hacerlo se ha trabajado con librerías específicas de Python de tratamiento de datos como Pandas y se han adaptado las guías ACMG/AMP de esta manera se ha conseguido un pipeline coherente.

La metodología de *design science* ha demostrado ser efectiva, permitiendo en primer lugar llegar a entender con profundidad el problema y después generar un diseño y solución a este mismo de esta manera se ha conseguido una clasificación precisa de variaciones genéticas con el objetivo futuro de proporcionar una herramienta para la investigación y el diagnóstico de las DHR.

9.1. Líneas futuras

La principal investigación futura se debe centrar en la validación del pipeline en un entorno clínico real con más pacientes. Por otra parte, otra línea de investigación futura puede ser la mejora constante del mismo, para esto se deben incluir más bases de datos y hacer uso y desarrollar criterios de clasificación más específicos e ir actualizando con nuevos parámetros e información de nuevos estudios de estas enfermedades. Así se da una visión más completa del impacto de las variaciones.

La aplicación de inteligencia artificial y de aprendizaje automático puede ser también una opción para la mejora del pipeline. Los modelos de aprendizaje automático pueden detectar patrones complejos en los datos de entrada del pipeline y clasificar automáticamente las variaciones.

Además de estas líneas también se pueden abrir otras nuevas una vez se llega a comprender de forma correcta la clasificación de las variaciones genéticas como puede ser la farmacogenética y la terapia génica. Se podría desarrollar y probar terapias génicas dirigidas que corrijan las variaciones causantes de la enfermedad. Por otro lado, en el campo de la farmacogenética se puede estudiar cómo es la respuesta de las variaciones ante determinados fármacos por lo que se pueden usar para identificar biomarcadores específicos que predigan la enfermedad

10. OBJETIVOS DE DESARROLLO SOSTENIBLE

Para finalizar la memoria en este apartado se va a ver el grado de relación que tiene el Trabajo de Fin de Máster con los Objetivos de Desarrollo Sostenible de la agenda 2030 (ODS). Para esto se usará la siguiente Tabla 11 con los objetivos y las relaciones, y finalmente se explicará cada una de las relaciones resultantes.

Objetivos de desarrollo sostenible	Alto	Medio	Bajo	No procede
ODS 1. Fin de la pobreza				X
ODS 2. Hambre cero				X
ODS 3. Salud y bienestar	X			
ODS 4. Educación de calidad		X		
ODS 5. Igualdad de género				X
ODS 6. Agua limpia y saneamiento				X
ODS 7. Energía asequible y no contaminante				X
ODS 8. Trabajo decente y crecimiento económico			X	
ODS 9. Industria, innovación e infraestructuras		X		
ODS 10. Reducción de las desigualdades		X		
ODS 11. Ciudades y comunidades sostenibles				X
ODS 12. Producción y consumo responsables				X
ODS 13. Acción por el clima				X
ODS 14. Vida submarina				X
ODS 15. Vida de ecosistemas terrestres				X
ODS 16. Paz, justicia e instituciones sólidas				X
ODS 17. Alianzas para lograr objetivos				X

Tabla 11. *Objetivos de desarrollo sostenible de la agenda de 2030.*

En este trabajo se ha diseñado y desarrollado un pipeline para el análisis de variaciones de distrofia hereditaria de retina y esto puede contribuir a varios Objetivos de Desarrollo Sostenible. Para comenzar está relacionado con 4 objetivos. Primero, con el ODS 3 de salud y bienestar, ya que permite facilitar el diagnóstico haciéndolo más preciso y permitiendo la detección temprana de la patología, en concreto, ODS 4 de educación de calidad, pues puede generar conocimiento nuevo que puede utilizarse en la formación de profesionales de la salud y la investigación genética. ODS 9 industria, innovación e infraestructuras en este caso generar la creación de un pipeline de clasificación de variaciones genéticas implica innovación en herramientas bioinformáticas y, por último, ODS 10 de reducción de desigualdades, al mejorar la comprensión y el tratamiento de una enfermedad genética, contribuyes indirectamente a reducir las desigualdades en salud.

11. BIBLIOGRAFÍA

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013a). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, *76*(1), 7.20.1-7.20.41. https://doi.org/10.1002/0471142905.hg0720s76

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013b). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *Chapter 7*, Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76

Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., Berg, J. S., Biswas, S., Bowling, K. M., Conlin, L. K., Cooper, G. M., Dorschner, M. O., Dulik, M. C., Ghazani, A. A., Ghosh, R., Green, R. C., Hart, R., Horton, C., Johnston, J. J., … Rehm, H. L. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *American Journal of Human Genetics*, *98*(6), 1067-1076. https://doi.org/10.1016/j.ajhg.2016.03.024

Ashoor, H., Kleftogiannis, D., Radovanovic, A., & Bajic, V. B. (2015). DENdb: Database of integrated human enhancers. *Database*, *2015*, bav085. https://doi.org/10.1093/database/bav085

Ayuso, C., & Millan, J. M. (2010). Retinitis pigmentosa and allied conditions today: A paradigm of translational research. *Genome Medicine*, *2*(5), 34. https://doi.org/10.1186/gm155

Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, *34*(suppl_2), W369-W373. https://doi.org/10.1093/nar/gkl198

Bhave, P. (2015, mayo 10). Population Health Research and Disease Management – The Frontier (Part 4). *Clinical Scientist*. https://clinicalscientist.wordpress.com/2015/05/10/population-health-research-and-disease-management-the-frontier-part-4/

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., & Murphy, T. D. (2015). Gene: A gene-centered information resource at NCBI. *Nucleic Acids Research*, *43*(D1), D36-D42. https://doi.org/10.1093/nar/gku1055

Burkard, M., Kohl, S., Krätzig, T., Tanimoto, N., Brennenstuhl, C., Bausch, A. E., Junger, K., Reuter, P., Sothilingam, V., Beck, S. C., Huber, G., Ding, X.-Q., Mayer, A. K., Baumann, B., Weisschuh, N., Zobor, D., Hahn, G.-A., Kellner, U., Venturelli, S., … Ruth, P. (2018). Accessory heterozygous mutations in cone photoreceptor *CNGA3* exacerbate CNG channel–associated retinopathy. *The Journal of Clinical Investigation*, *128*(12), 5663-5675. https://doi.org/10.1172/JCI96098

Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., Compton, C. C., DeLuca, D. S., Peter-Demchok, J., Gelfand, E. T., Guan, P., Korzeniewski, G. E., Lockhart, N. C., Rabiner, C. A., Rao, A. K., Robinson, K. L., Roche, N. V., Sawyer, S. J., Segrè, A. V., … on behalf of the GTEx Consortium. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*, *13*(5), 311-319. https://doi.org/10.1089/bio.2015.0032

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Reviews. Genetics*, *3*(4), 285-298. https://doi.org/10.1038/nrg775

Cingolani, P. (2022). Variant Annotation and Functional Prediction: SnpEff. En C. Ng & S. Piscuoglio (Eds.), *Variant Calling: Methods and Protocols* (pp. 289-314). Springer US. https://doi.org/10.1007/978-1-0716-2293-3_19

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80-92. https://doi.org/10.4161/fly.19695

Claeys, M., Storms, V., Sun, H., Michoel, T., & Marchal, K. (2012). MotifSuite: Workflow for probabilistic motif detection and assessment. *Bioinformatics*, *28*(14), 1931-1932. https://doi.org/10.1093/bioinformatics/bts293

Costa, M., García S., A., León, A., & Pastor, O. (2023). Comprehensive Representation of Variation Interpretation Data via Conceptual Modeling. En T. P. Sales, J. Araújo, J. Borbinha, & G. Guizzardi (Eds.), *Advances in Conceptual Modeling* (pp. 25-34). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47112-4_3

Costa, M., S., A. G., Leon, A., Bernasconi, A., & Pastor, O. (2023). A Reference Meta-model to Understand DNA Variant Interpretation Guidelines. En J. P. A. Almeida, J. Borbinha, G. Guizzardi, S. Link, & J. Zdravkovic (Eds.), *Conceptual Modeling* (pp. 375-393). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47262-6_20

Costa Sánchez, M. (2021). *Diseño y desarrollo de una plataformoa para la gestión de datos genómicos: Oráculo genómico de delfos.* [UPV]. http://hdl.handle.net/10251/173220

Cremers, F. P., van de Pol, D. J., van Driel, M., den Hollander, A. I., van Haren, F. J., Knoers, N. V., Tijmes, N., Bergen, A. A., Rohrschneider, K., Blankenagel, A., Pinckers, A. J., Deutman, A. F., & Hoyng, C. B. (1998). Autosomal recessive retinitis pigmentosa and cone-rod dystrophy caused by splice site mutations in the Stargardt's disease gene ABCR. *Human Molecular Genetics*, *7*(3), 355-362. https://doi.org/10.1093/hmg/7.3.355

Cromosomas: Qué son, función, tipos y características. (s. f.). https://humanidades.com/. Recuperado 4 de julio de 2024, de https://humanidades.com/cromosomas/

de Sainte Agathe, J.-M., Filser, M., Isidor, B., Besnard, T., Gueguen, P., Perrin, A., Van Goethem, C., Verebi, C., Masingue, M., Rendu, J., Cossée, M., Bergougnoux, A., Frobert, L., Buratti, J., Lejeune, É., Le Guern, É., Pasquier, F., Clot, F., Kalatzis, V., … Baux, D. (2023). SpliceAI-visual: A free online tool to improve SpliceAI splicing variant interpretation. *Human Genomics*, *17*(1), 7. https://doi.org/10.1186/s40246-023-00451-1

de Souza, N. (2012). The ENCODE project. *Nature Methods*, *9*(11), 1046-1046. https://doi.org/10.1038/nmeth.2238

Definición de ADN - Diccionario de genética del NCI - NCI (nciglobal,ncienterprise). (2012, julio 20). [nciAppModulePage]. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-genetica/def/adn

Definición de cosegregación—Diccionario de genética del NCI - NCI (nciglobal,ncienterprise). (2012, julio 20). [nciAppModulePage]. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-genetica/def/cosegregacion

Definición de variante de novo—Diccionario de cáncer del NCI - NCI (nciglobal,ncienterprise). (2011, febrero 2). [nciAppModulePage]. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/variante-de-novo

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., & Béroud, C. (2009). Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*, *37*(9), e67-e67. https://doi.org/10.1093/nar/gkp215

Dreos, R., Ambrosini, G., Périer, R. C., & Bucher, P. (2015). The Eukaryotic Promoter Database: Expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Research*, *43*(D1), D92-D96. https://doi.org/10.1093/nar/gku1111

Ferrari, S., Di Iorio, E., Barbaro, V., Ponzin, D., Sorrentino, F. S., & Parmeggiani, F. (2011). Retinitis pigmentosa: Genes and disease mechanisms. *Current Genomics*, *12*(4), 238-249. https://doi.org/10.2174/138920211795860107

Ferreiro, S. (2023, febrero 17). ¿Sabes qué es el ADN y sus funciones y estructura? *ADNTRO*. https://adntro.com/es/blog/aprende-genetica/que-es-adn/

Fotorreceptores. (2017, noviembre 18). American Academy of Ophthalmology. https://www.aao.org/salud-ocular/anatomia/fotorreceptores

García Ordaz, D. M. (2011). *Identificación de secuencias reguladoras mediante agrupamiento*. http://inaoe.repositorioinstitucional.mx/jspui/handle/1009/682

Gómez Skarmeta, J. L. (s. f.). ¿Qué hay en el ADN no codificante? *2010*. https://repisalud.isciii.es/bitstream/handle/20.500.12105/14070/Qu%C3%A9HayADNNoCodificante_2010.pdf?sequence=1&isAllowed=y

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)*, *185*(4154), 862-864. https://doi.org/10.1126/science.185.4154.862

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M. C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S. M., Giardine, B., Hooghe, B., Van Loo, P., Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E., … The Open Regulatory Annotation Consortium. (2007). ORegAnno: An open-access community-driven resource for regulatory annotation. *Nucleic Acids Research*, *36*(Database), D107-D113. https://doi.org/10.1093/nar/gkm967

Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., Genome Aggregation Database Consortium, Rehm, H. L., MacArthur, D. G., & O'Donnell-Luria, A. (2022). Variant interpretation using population databases: Lessons from gnomAD. *Human Mutation*, *43*(8), 1012-1030. https://doi.org/10.1002/humu.24309

Hurtado, C. (2022). Medicina de precisión: Conceptos, aplicaciones y proyecciones. *Revista Médica Clínica Las Condes*, *33*(1), 7-16. https://doi.org/10.1016/j.rmclc.2022.01.002

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434-443. https://doi.org/10.1038/s41586-020-2308-7

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., Birnbaum, D., The Exome Aggregation Consortium, Daly, M. J., & MacArthur, D. G. (2017). The ExAC browser: Displaying reference data information from over 60 000 exomes. *Nucleic Acids Research*, *45*(D1), D840-D845. https://doi.org/10.1093/nar/gkw971

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996-1006. https://doi.org/10.1101/gr.229102

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062-D1067. https://doi.org/10.1093/nar/gkx1153

Leon, A., S, A. G., Roman, J. F. R., Costa, M., & Pastor, O. (2024). The Delfos Platform: A Conceptual Model-Based Solution for the Enhancement of Precision Medicine. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *PP*. https://doi.org/10.1109/TCBB.2024.3377928

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Liu, X., Brutlag, D. L., & Liu, J. S. (2000). Bioprospector: Discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. En *Biocomputing 2001* (pp. 127-138). WORLD SCIENTIFIC. https://doi.org/10.1142/9789814447362_0014

Liu, X., Jian, X., & Boerwinkle, E. (2011). dbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions. *Human Mutation*, *32*(8), 894-899. https://doi.org/10.1002/humu.21517

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, *37*(3), 235-241. https://doi.org/10.1002/humu.22932

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., Mungall, C. J., Arner, E., Baillie, J. K., Bertin, N., Bono, H., de Hoon, M., Diehl, A. D., Dimont, E., Freeman, T. C., … the FANTOM consortium. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology*, *16*(1), 22. https://doi.org/10.1186/s13059-014-0560-6

Matys, V. (2003). TRANSFAC(R): Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, *31*(1), 374-378. https://doi.org/10.1093/nar/gkg108

Michaelides, M., Hunt, D. M., & Moore, A. T. (2003). The genetics of inherited macular dystrophies. *Journal of Medical Genetics*, *40*(9), 641-650. https://doi.org/10.1136/jmg.40.9.641

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812-3814. https://doi.org/10.1093/nar/gkg509

Olivé, A. (2007). *Conceptual Modeling of Information Systems*. Springer. https://doi.org/10.1007/978-3-540-39390-0

Perez-Romero, I., Fernández-Caballero, L., F Iancu, I., Rodilla, C., Martín-Mérida, I., & Ávila-Fernández, A. (2022). Distrofias hereditarias de retinaen España: Tres décadas de estudio epidemiológico, clínico y genético. *2022*. https://analesranm.es/wp-content/uploads/2022/numero_139_03/pdfs/analesranm_13903.pdf#page=60

Pertea, M. (2001). GeneSplicer: A new computational method for splice site prediction. *Nucleic Acids Research*, *29*(5), 1185-1190. https://doi.org/10.1093/nar/29.5.1185

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110-121. https://doi.org/10.1101/gr.097857.109

Pozo Valero, M. del. (2022, octubre 22). *Estudio clínico y molecular de las distrofias hereditarias de retina asociadas a ABCA4 y PROM1*. https://repositorio.uam.es/handle/10486/693810

Pruitt, K. D. (2004). NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(Database issue), D501-D504. https://doi.org/10.1093/nar/gki025

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., … Ostell, J. M. (2014). RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, *42*(D1), D756-D763. https://doi.org/10.1093/nar/gkt1114

¿Qué son las Distrofias Hereditarias de Retina. (2019, junio 14). *GuíaSalud*. https://portal.guiasalud.es/egpc/pacientes_distrofias-que-son/

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L., Plon, S. E., Ramos, E. M., Sherry, S. T., Watson, M. S., & ClinGen. (2015). ClinGen—The Clinical Genome Resource. *The New England Journal of Medicine*, *372*(23), 2235-2242. https://doi.org/10.1056/NEJMsr1406261

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886-D894. https://doi.org/10.1093/nar/gky1016

RetNet—Retinal Information Network. (s. f.). Recuperado 4 de julio de 2024, de https://retnet.org/

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., & ACMG Laboratory Quality Assurance Committee. (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: Official Journal of the American College of Medical Genetics*, *17*(5), 405-424. https://doi.org/10.1038/gim.2015.30

Román, R., & Fabián, J. (2018). *DISEÑO Y DESARROLLO DE UN SISTEMA DE INFORMACIÓN GENÓMICA BASADO EN UN MODELO CONCEPTUAL HOLÍSTICO DEL GENOMA HUMANO* [Tesis doctoral, Universitat Politècnica de València]. https://doi.org/10.4995/Thesis/10251/99565

Romero Sánchez, Y. G. (2022). *Localización de motivos de secuencias de ADN usando un algoritmo genético*. https://hdl.handle.net/20.500.12371/16486

Roy, A., Abdullah, R., Ahmed, F., Mashfi, S., Khan, S. H., & Karim, D. Z. (2023). RetNet: Retinal Disease Detection using Convolutional Neural Network. *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1-6. https://doi.org/10.1109/ECCE57851.2023.10101661

Schulz, H. L., Grassmann, F., Kellner, U., Spital, G., Rüther, K., Jägle, H., Hufendiek, K., Rating, P., Huchzermeyer, C., Baier, M. J., Weber, B. H. F., & Stöhr, H. (2017). Mutation Spectrum

of the ABCA4 Gene in 335 Stargardt Disease Patients From a Multicenter German Cohort—Impact of Selected Deep Intronic Variants and Common SNPs. *Investigative Ophthalmology & Visual Science*, *58*(1), 394-403. https://doi.org/10.1167/iovs.16-19936

Sefid Dashti, M. J., & Gamieldien, J. (2017). A Practical Guide To Filtering and Prioritizing Genetic Variants. *BioTechniques*, *62*(1), 18-30. https://doi.org/10.2144/000114492

Sheng, Q., Yu, H., Oyebamiji, O., Wang, J., Chen, D., Ness, S., Zhao, Y.-Y., & Guo, Y. (2020). AnnoGen: Annotating genome-wide pragmatic features. *Bioinformatics*, *36*(9), 2899-2901. https://doi.org/10.1093/bioinformatics/btaa027

Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., Loerch, P. M., Leonardson, A., Lum, P. Y., Cavet, G., Wu, L. F., Altschuler, S. J., Edwards, S., King, J., Tsang, J. S., Schimmack, G., Schelter, J. M., Koch, J., Ziman, M., … Boguski, M. S. (2001). Experimental annotation of the human genome using microarray technology. *Nature*, *409*(6822), 922-927. https://doi.org/10.1038/35057141

Soens, Z. T., Branch, J., Wu, S., Yuan, Z., Li, Y., Li, H., Wang, K., Xu, M., Rajan, L., Motta, F. L., Simões, R. T., Lopez-Solache, I., Ajlan, R., Birch, D. G., Zhao, P., Porto, F. B., Sallum, J., Koenekoop, R. K., Sui, R., & Chen, R. (2017). Leveraging splice-affecting variant predictors and a minigene validation system to identify Mendelian disease-causing variants amongst exon-captured variants of uncertain significance. *Human mutation*, *38*(11), 1521-1533. https://doi.org/10.1002/humu.23294

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., … Harris, L. W. (2022). The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, *51*(D1), D977-D985. https://doi.org/10.1093/nar/gkac1010

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P., & Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, *17*(12), 1113-1122. https://doi.org/10.1093/bioinformatics/17.12.1113

UniProt Consortium. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523-D531. https://doi.org/10.1093/nar/gkac1052

Visel, A., Minovitsky, S., Dubchak, I., & Pennacchio, L. A. (2007). VISTA Enhancer Browser—A database of tissue-specific human enhancers. *Nucleic Acids Research*, *35*(Database), D88-D92. https://doi.org/10.1093/nar/gkl822

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer.

Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y., & Kolpakov, F. (2019). GTRD: A database on gene transcription regulation—2019 update. *Nucleic Acids Research*, *47*(D1), D100-D105. https://doi.org/10.1093/nar/gky1128

Ziebarth, J. D., Bhattacharya, A., & Cui, Y. (2012). CTCFBSDB 2.0: A database for CTCF-binding sites and genome organization. *Nucleic Acids Research*, *41*(D1), D188-D194. https://doi.org/10.1093/nar/gks1165

PRESUPUESTO

Diseño y Desarrollo de un pipeline para el análisis de variaciones de distrofia hereditaria de retina.

Documento II

Índice del presupuesto

1. INTRODUCCIÓN

El objetivo de este apartado es realizar un estudio de los presupuestos a considerar al realizar este Trabajo Final de Máster para estimar los costes económicos del proyecto. Para realizarlo se han descompuesto los distintos costes en mano de obra, hardware y software para que los cálculos sean más sencillos y que se pueda apreciar los costes de cada uno.

Al final del capítulo se exponen los presupuestos totales tanto de ejecución de material como ejecución por contrata. Para hacer los cálculos se ha usado el programa Arquímedes que sirve para hacer el cálculo de presupuestos de proyectos.

2. PRESUPUESTO DESCOMPUESTO

En esta sección se va a presentar el presupuesto descompuesto total del proyecto. Para esto se van a representar tres tablas resumen de costes de mano de obra (Tabla 12), costes de software (Tabla 13) y costes de hardware (Tabla 14).

2.1. Presupuesto mano de obra

Para llevar a cabo el presente TFM se ha necesitado a un estudiante del Máster de Ingeniería Biomédica que ha sido el encargado de desarrollar y dar forma al proyecto, un tutor catedrático de Universidad, un cotutor doctor contratado y una investigadora predoctoral encargados de dirigir, supervisar y enseñar al estudiante. Se considerarán todas las horas dedicadas para el correcto desarrollo del trabajo.

Para poder comenzar a hacer cálculos se realiza una búsqueda para conocer el sueldo base de los ingenieros, doctores y catedráticos. El sueldo base del Catedrático de Universidad es de 47.573,29 €, el de un doctor es 34454,70 €, el de un investigador predoctoral es 31086,30 € y por último el de un ingeniero *junior* es 23985,08 €, según las tablas retributivas para el 2024 de la Univerdidad Politécnica de Valencia[6]. Considerando que la jornada laboral de ambos es de 8 horas, sin contar fines de semana y vacaciones, se obtienen 224 jornadas laborales al año lo que vienen siendo 1792 h/año. Con el sueldo base y las horas trabajadas se puede deducir el salario por hora de ambos, del Catedrático de Universidad vemos que es 26,55 €/h, del doctor es 19,22€/h, del investigador predoctoral es 17,35 €/h y por último del ingeniero biomédico *junior* vemos que es de 13,4 €/h. Los resultados se muestran en la Tabla 12.

[6] https://www.upv.es/entidades/SRH/retribuciones/787565normalv.html

Tipo	Cantidad (h)	Coste unitario (€/h)	Coste total (€)
Ingeniero biomédico: *título superior*	567	13,40	7 597,8
Tutor: Catedrático de Universidad	25	26,55	663,75
Cotutor: Investigador Predoctoral	80	19,22	1 537,60
Miembro grupo PROS: Doctor Contratado	10	17,35	173,5
Subtotal			**9 972,65**

Tabla 12. *Presupuesto de mano de obra.*

2.2. Presupuesto software

A continuación, se va a estimar el coste del software, se ha tenido en cuenta el coste de las licencias de cada uno de los programas utilizados. Los programas que supondrán costes son los relacionados con Microsoft, tanto Windows 10 como Office, con un uso de ambos de tres años y un coste fijo por cada una. Por otro lado, las diferentes bases de datos descargadas y usadas y el programa de *Visual Studio Code* no suponen costes. El presupuesto de *software* final se puede ver en la Tabla 13.

Tipo	Cantidad (Uds.)	Periodo de uso	Precio (€)	Coste total (€)
Sistema Operativo Microsoft Windows 10	1	6 meses	145,00	24,17
Microsoft Office Hogar y Estudiantes	1	6 meses	149,00	24,83
Bases de datos	1	6 meses	0,00	0,00
Visual Studio Code	1	6 meses	0,00	0,00
Subtotal				**49**

Tabla 13*. Presupuesto de software.*

2.3. Presupuesto hardware

Para terminar, se analizan los costes del *hardware* usado. Debemos tener en cuenta el ordenador usado en el laboratorio junto con la GPU o tarjeta gráfica usada, para calcular sus costes usamos un factor de amortización que se obtiene teniendo en cuenta el periodo de vida útil. El trabajo se ha llevado a cabo durante 6 meses y la vida útil del ordenador y de la GPU generalmente es de 6 años lo que vienen siendo 72 meses. El factor de amortización en este caso es 6/72, el coste total se puede observar en la Tabla 14.

Tipo	Cantidad (Uds.)	Factor amortización	Precio (€)	Coste total (€)
Ordenador	1	6/72	1 000,00	83,33
GPU	1	6/72	780,00	65
Subtotal				**148,33**

Tabla 14. Presupuesto hardware.

3. PRESUPUESTO TOTAL

3.1. Presupuesto de ejecución material

Para calcularlo se realiza la suma de cada uno de los presupuestos parciales (mano de obra, software y hardware), los costes totales los podemos ver en la Tabla 15.

Unidad de obra	Coste total (€)
Mano de obra	9 972,65
Software	49
Hardware	148,33
Total	**10 169,98**

Tabla 15*. Presupuesto de ejecución material.*

3.2. Presupuesto de ejecución por contrata

Finalmente, para hacer el cálculo del presupuesto por contrata hay que tener en cuenta los impuestos establecidos y añadirlos a los costes ya calculados. Los resultados finales se pueden ver en la Tabla 16.

Tipo	Coste total (€)
Presupuesto ejecución material	**10 169,98**
Beneficio industrial (6%)	610,20
Gastos generales (13%)	1 322,10
Total antes de impuestos	**12 102,28**
IVA	2 541.48
Presupuesto ejecución por contrata	**14 643,76**

Tabla 16*. Presupuesto ejecución por contrata.*

Finalmente teniendo en cuenta los impuestos, gastos generales y el beneficio industrial el coste del proyecto asciende a catorce mil seiscientos cuarenta y tres euros con setenta y seis céntimos.

ANEXO

Diseño y Desarrollo de un pipeline para el análisis de variaciones de distrofia hereditaria de retina.

Documento III

Índice Anexo

1. LISTADO DE GENES

En este punto del anexo se incluyen todos los genes que se van a tener en cuenta a la hora de desarrollar el filtro basado en los genes específicos de las DHR.

SAMD11,NPHP4,ESPN,NMNAT1,MFN2,EMC1,PLA2G5,DHDDS,PPT1,ELOVL1,POMGNT1,RPE65,ABCA4,COL11A1,GNAT2,CLCC1,DRAM2,PRPF3,ENSA,SEMA4A,CORD8,ATF6,HMCN1,CFH,CRB1,ADIPOR1,RD3,NEK2,FLVCR1,USH2A,SDCCAG8,OR2W3,NBAS,AGBL5,ZNF513,IFT172,PCARE,EFEMP1,FAM161A,WDPCP,ALMS1,SNRNP200,CNNM4,CNGA3,NPHP1,MERTK,BBS5,CERKL,NEUROD1,TMEM237,KCNJ13,SAG,SPP2,TRNT1,USH2B,SLC4A7,LZTFL1,GNAT1,TREX1,MAPKAPK3,ATXN7,PROS1,ARL6,IMPG2,IQCB1,RHO,NPHP3,RP5,CLRN1,SLC7A14,OPA1,PCYT1A,CEP19,PDE6B,WFS1,HMX1,RAB28,CC2D2A,PROM1,ADGRA3,DTHD1,WDR19,CNGA1,WFS2,MTTP,LRIT3,BBS7,BBS12,MFSD8,PLK4,RP29,LRAT,TLR3,CYP4V2,MCDR3,CWC27,POC5,VCAN,ADGRV1,NR2F1,SLC25A46,CTNNA1,HARS,PDE6A,GRM6,MAK,C2,CFB,TULP1,GUCA1A,GUCA1B,PRPH2,BCAMD,EYS,COL9A1,RIMS1,IMPG1,LCA5,ELOVL4,PRDM13,RTN4IP1,RP63,AHI1,PEX7,RCD1,MDDC,AHR,KLHL7,RP9,BBS9,PEX1,TSPAN12,IMPDH1,OPN1SW,KIAA1549,RP1L1,ADAM9,HGSNAT,RP1,TTPA,CSPP1,OPA6,PEX2,CNGB3,C8orf37,GDF6,RIMS2,VMD1,KCNV2,TOPORS,MIR204,CEP78,INVS,PRPF4,WHRN,TRIM32,TLR4,DYNC2I2,RP8,EXOSC2,INPP5E,PHYH,ACBD5,USH1K,RBP3,ERCC6,RNANC,PCDH15,HK1,CDH23,CDHR1,RGR,KIF11,RBP4,PDE6C,PAX2,PDZD7,ARL3,BBIP1,CORD17,ARMS2,HTRA1,OAT,ZNF408,TUB,TEAD1,USH1C,EVR3,TMEM216,BEST1,ASRGL1,ROM1,BBS1,CABP4,LRP5,CAPN5,MYO7A,TMEM126A,FZD4,DYNC2H1,CEP164,C1QTNF5,MFRP,CACNA2D4,GNB3,PDE6H,COL2A1,CODA1,RDH5,CCT2,BBS10,CEP290,POC1B,MVK,IFT81,C12orf65,ITM2B,RB1,RCBTB1,GRK1,STGD2,ACHM1,RP16,MCDR4,RPGRIP1,NRL,OTX2,RDH11,RDH12,TTLL5,SPATA7,USH1A,TTC8,FBLN5,TRPM1,TUBGCP4,USH1H,SLC24A1,NR2E3,MRST,BBS4,CIB2,RLBP1,GNPTG,IFT140,CLUAP1,ABCC6,RP22,CLN3,ZNF423,RPGRIP1L,BBS2,ARL2BP,CNGB1,OPA8,CDH3,DHX38,ADAMTS18,FHASD,CACD,RCD2,PRPF8,AIPL1,PITPNM3,GUCY2D,CORD4,UNC119,GPR179,MKS1,CA4,RGS9,ARSG,USH1G,PRCD,FSCN2,PDE6G,LAMA1,AFG3L2,OPA4,CORD1,REEP6,RAX2,C3,ARHGEF18,PNPLA6,RGS9BP,MCDR5,CRX,OPA3,PRPF31,IDH3B,PANK2,JAG1,MKKS,KIZ,ABHD12,KIF3B,CEP250,PRPF6,USH1E,C21orf2,OPA5,TIMP3,IFT27,VRD1,MIEF1,ACO2,TUBGCP6,OFD1,RS1,RP6,DMD,OPA2,RPGR,NYX,COD1,RP15,PRD,NDP,AIED,RP2,CACNA1F,PGK1,CHM,TIMM8A,PRPS1,RP24,COD2,RP34,BCM,OPN1LW,OPN1MW,KSS,LHON,MT-TL1,MT-ATP6,MT-TH,MT-TS2,MT-TP

2. CÓDIGOS

A continuación, se explican los códigos desarrollados para poder hacer la implementación de las dos bases de datos específicas para DHR.

2.1. RetNet

En primer lugar, para la transformación de los datos se desarrolla un código guardado con el nombre de transform_retnet_data.py, una vez trasformados los datos se guardan en un archivo CSV. El código es el siguiente:

```
from os import path
from tqdm import tqdm
import sys
from dotenv import load_dotenv


dir_path = path.dirname(path.abspath(__file__))
sys.path.append(path.join(dir_path, "../"))
```

```
from utils import utils


def transform_retnet_data():

    tqdm.pandas(desc="[RetNet Data Transformation]")
    load_dotenv()

    paths_dict = utils.get_transform_paths('retnet')
    for paths in paths_dict.values():
        input_path = paths["input_path"]
        output_path = dict(paths["output_path"])

        retnet_data =utils.read_df_from_csv_latin(input_path["diseases"])

        retnet_data = transform_retnet(retnet_data)
        utils.save_df_to_csv(retnet_data, output_path["diseases"])

def transform_retnet(df):
    df = df[[
        "Symbols; OMIM Numbers", "Diseases; Protein",
        "How Identified; Comments", "References"
    ]]
    df = df.fillna(".")

    return df

if __name__ == '__main__':
    transform_retnet_data()
```

Este código tiene diferentes partes, en la primera se comienza importando las librerías (os, sys, tqdm, dotenv) necesarias para que funcione de forma correcta en script. Tras esto se gestionan las rutas y las variables de entorno, estableciendo las rutas de los archivos de entrada a tratar y las rutas de los archivos de salida una vez los datos estén trasformados.

La función principal de este es la de 'transform_retnet_data', en esta se cargan los datos de los archivos de entrada, se aplica la función de trasformación de los datos llamada 'trasnform_retnet' y se guarda los datos trasformados en un archivo CSV.

La función 'transorm_retnet' se encarga de limpiar y transformar el contenido del *dataframe* de retnet. Se seleccionan solo las columnas relevantes para el análisis y se reemplazan los valores vacíos por un '.' Para evitar errores.

Ahora para hacer la mejora de los datos se desarrolla por otro lado un *script* llamado enhance_retmet_data.py. Es el siguiente código.

```
def enhance_retnet_data():
    tqdm.pandas(desc="[RetNet Data Enhancement]")

    load_dotenv()
```

```
    paths_dict = utils.get_enhance_paths('retnet')
    for paths in paths_dict.values():
        input_path = paths["input_path"]
        output_path = paths["output_path"]["diseases"]

        # Load data from CSV files
        retnet_data =
utils.read_df_from_csv(input_path["diseases"]).fillna('.')
        reference_data =
utils.read_df_from_csv(input_path["references"]).fillna('.')

        retnet_data = enhance(retnet_data, reference_data)

        # Save the result to a new CSV file
        utils.save_df_to_csv(retnet_data, output_path)

        utils.remove_input_dbs("RetNet")

def remove_omim_identifier(df):

    df["Symbols"] = df["Symbols; OMIM Numbers"].apply(split_first,
args=(";",)) # Rename the column
    df.drop(columns=['Symbols; OMIM Numbers'], inplace=True)
    return df

def first_gene(df):

    df["Symbols"] = df["Symbols"].apply(split_first, args=(",",))
    return df
def split_first(x, split_separator):
    return x.split(split_separator)[0]

def remove_protein(df):

    df["Diseases"] = df["Diseases; Protein"].apply(remove_protein_info)
    df.drop(columns=['Diseases; Protein'], inplace=True)
    return df

def remove_protein_info(x):
    elements = [elem.strip() for elem in x.split(';') if not
elem.strip().startswith('protein:')]
    result = '@'.join(elements)
    return result

def remove_gene(df):
    df["Diseases"] = df["Diseases"].str.replace(r'\[Gene\]', '',
regex=True)
    return df
```

```
def find_complete_reference(author_last_name, year_two_digits,
references_df):
    complete_reference = references_df[
        references_df["references"].str.contains(author_last_name,
case=False) &
        references_df["references"].str.contains(year_two_digits,
case=False)
    ]["references"].values

    return complete_reference[0] if len(complete_reference) > 0 else None

def get_complete_references(partial_references, references_df):
    complete_references = []
    for partial_ref in partial_references:
        author_last_names = [name.strip() for name in
partial_ref.split(',')]
        for author_last_name in author_last_names:
            author_last_name, year_two_digits = author_last_name.rsplit('
', 1)
            complete_reference =
find_complete_reference(author_last_name, year_two_digits, references_df)
            complete_references.append(complete_reference or partial_ref)
    return '@ '.join(complete_references)

def obteining_references(df, references_df):

    # Gets references by row
    df["References"] = df["References"].progress_apply(lambda x:
get_complete_references(x.split(';'), references_df))
    return df

def text_transform(x):
    parts = x.split(';') # Separate text by ';'
    how_identified = parts[0] # Extract first element as how_identified

    # Build the new text
    transformed_text = f"The following methods were used for determining
the gene-disease relationship: {how_identified}. The following evidences
support the association: {'. '.join(parts[1:])}"

    return transformed_text

def drop_rows_with_symbol(df, val, column):
    df = df[df[column] != val]
    return df

def enhance(retnet_df, references_df):
```

```
    column_names = {
        'Symbols': 'Gene.official_symbol@RetNet',
        'Diseases': 'Phenotype.preferred_name@RetNet',
        'How Identified': 'EntityPrediction.description@RetNet',
        'References': 'BibliographyReference@RetNet'
    }

    # Apply improvements
    retnet_df = remove_omim_identifier(retnet_df)
    retnet_df = first_gene(retnet_df)
    retnet_df = remove_protein(retnet_df)
    retnet_df = remove_gene(retnet_df)
    retnet_df = obteining_references(retnet_df, references_df)
    retnet_df["How Identified"] = retnet_df["How Identified;
Comments"].apply(text_transform)
    retnet_df.drop(columns=['How Identified; Comments'], inplace=True)
    retnet_df=drop_rows_with_symbol(retnet_df,'(- - -)','Symbols')

    #put the columns in order
    column_order = ['Symbols', 'Diseases', 'How Identified',
'References']
    retnet_data = retnet_data[column_order]

    return rename_columns(retnet_df, column_names, drop=True)

if __name__ == '__main__':
    enhance_retnet_data()
```

De nuevo se explica el código, en primer lugar, se vuelven a cargar las diferentes librerías necesarias para el correcto funcionamiento del mismo. Una vez importadas se gestionan las rutas de los archivos de entrada y salida, en este caso se deben obtener las rutas de entrada de dos archivos y no solo de uno ya que tenemos los datos de los genes de las DHR y las referencias y los carga en dos *dataframes* distintos.

La función principal de mejora de los datos es la llamada ‘enhance’, en esta se aplican una serie de transformaciones sobre las columnas y contenido que están definidas en distintas funciones.

Para empezar, encontramos la función que elimina los identificadores OMIM de la columna *Symbol*, tras esta la que permite seleccionar el primer gen en los casos en los que haya múltiples genes listados, tras esto se elimina la información de las proteínas que están en la columna de *Diseases* y por último se eliminan las referencias específicas de genes en los nombres.

Tras esto se crean una serie de funciones que permiten coger las referencias del segundo archivo y hacer un match con las que hay en el archivo que tiene los datos de las enfermedades. Estas funciones se basan en analizar las iniciales de los autores y los años de las citas. Es la función de ‘get_complete_references’ la que busca coincidencias de los nombres y las fechas en el archivo de referencias y las reemplaza.

Por último, se realiza una función que sirve para transformar el texto, en este caso es en la columna de *how identified* para mejorar la expresión de las asociaciones de gene y enfermedad, es decir se genera un texto más descriptivo. Para terminar, se guardan las transformaciones necesarias en un nuevo archivo CSV

2.2. NCBI refseq

En el caso de esta base de datos al igual que con RetNet también se van a desarrollar dos códigos uno en el que se hace la transformación de los datos o la selección de columnas relevantes y el segundo en el que se hace la mejora de los datos seleccionados. En primer lugar, el primer código se llama 'transform_NCBI_data.py' y es el siguiente:

```
def transform_NCBI_data(NCBI_data, output_path):
    with open(output_path, "w", newline="") as csvfile:
        csvwriter = csv.writer(csvfile)
        csvwriter.writerow(["chrom", "chromStart", "chromEnd", "name",
"pubmedids", "note", "experiment", "function"])

        # chromosome list
        chromosomes = NCBI_data.chroms()


        for chromosome, length in chromosomes.items():

            entries = NCBI_data.entries(chromosome, 0, length)
            for entry in entries:

                # choosen columns
                chrom = chromosome
                start = entry[0]
                end = entry[1]
                name_and_info = entry[2]

                name, *extra_info = name_and_info.split("\t")
                print("Datos en extra_info:", extra_info)


                if len(extra_info) >= 12:
                    note = extra_info [6]
                    pubmedids = extra_info[8]
                    experiment = extra_info[9]
                    function = extra_info[10]
                # write .csv
                csvwriter.writerow([chrom, start, end, name, pubmedids,
note, experiment, function])

def transform_NCBI():
    load_dotenv()
    paths_dict = utils.get_transform_paths('NCBI')
```

```
    for paths in paths_dict.values():

        input_path = paths["input_path"]["maindata"]
        output_path = paths["output_path"]["NCBI"]

        NCBI_data = utils.read_bigwig_file(input_path)
        transform_NCBI_data(NCBI_data, output_path)
        NCBI_data.close()


if __name__ == '__main__':
    transform_NCBI()
```

En primer lugar, se vuelven a cargar las diferentes librerías necesarias, y tras esto se gestionan las rutas de archivos tanto de entrada como de salida. La función principal es la de 'transform_NCBI_data'. Esta función es la encargada de procesar los datos y guardarlos en CSV.

Se recorren todos los cromosomas y para cada uno se procesan las diferentes columnas extrayendo de estas la información relevante que es la posición cromosómica, identificadores de publicaciones (PMIDs), notas y detalles experimentales. Los datos transformados se guardan en un archivo CSV.

Y tras este una vez ya seleccionados los datos relevantes, estos se mejoran con enhance_NCBI_data.py que es el siguiente código desarrollado:

```
def enhance_NCBI_data():
    load_dotenv()
    paths_dict = utils.get_enhance_paths('NCBI')

    for paths in paths_dict.values():
        input_path = paths["input_path"]["NCBI"]
        output_path = paths["output_path"]["NCBI"]

        # load data
        NCBI_data = utils.read_df_from_csv(input_path, separator=',')

        # apply enhancements
        NCBI_data = enhance(NCBI_data)

        # save data
        #utils.save_df_to_csv(NCBI_data,output_path)
        NCBI_data.to_csv(output_path, index=False, sep='\t',
encoding='utf-8-sig')

def fill_missing_values(df):
    return df.fillna(value='.')

def get_bibliography(df):
```

```
    # Transform the pubmedids column to have the appropiate format
    df['pubmedids'] = df['pubmedids'].str.split(',').apply(lambda x:
")|(".join([item.split('|')[0] for item in x]))
    total_pmids = extract_df_pmids(df['pubmedids'])
    df_bibliography = retrieve_pumbed_info(total_pmids)
    df["pubmedids"] = replace_df_pmids(df["pubmedids"], df_bibliography)

    return df


def remove_last_numbers(text):
    # Look for '[#/#]' and remove
    import re
    pattern = r'\[\d/\d\]'
    return re.sub(pattern, '', text)

def remove_numbers(df):
    # Apply remove_last_numbers to 'function' column
    df['function'] = df['function'].apply(remove_last_numbers)
    return df

def clean_experiment(text):
    # Remove first word EXISTENCE:
    cleaned_text = ' '.join(text.split()[1:])
    # Remove from '['
    cleaned_text = cleaned_text.split(' [')[0]
    return cleaned_text

def clean_experiment_column(df):
    # Apply clean_experiment to 'experiment' column
    df['experiment'] = df['experiment'].apply(clean_experiment)
    return df
# Define a function to replace '.' with '(No field information)' after
specific phrases


def replace_no_info(text):
    phrases = ['The description of the element: ', 'The element function
is: ', 'The experimental evidence is: ']
    for phrase in phrases:
        text = text.replace(phrase + '.', phrase + '(No information)')
    return text

def combine_info(df):
    #Remake the text
    df['Info'] = "The description of the element: " + df['note'] + ". The
element function is: " + df['function'] + ". The experimental evidence
is: " + df['experiment'] + "."
```

```
    df.drop(columns=['note', 'function', 'experiment'], inplace=True)

    df['Info'] = df['Info'].apply(replace_no_info)

    return df


def enhance(df):
    column_names = {
        'chrom':'chrom',
        'chromStart':'chromStart',
        'chromEnd':'chromEnd',
        'name': 'DNAEntity.name@NCBI_RefSeq_functional_elements',
        'pubmedids':
'BibliographyReference@NCBI_RefSeq_functional_elements',
        'Info': 'DNAEntity.description@NCBI_RefSeq_functional_elements'
    }

    df = fill_missing_values(df)
    df = get_bibliography(df)
    df = remove_numbers(df)
    df = clean_experiment_column(df)
    df = combine_info(df)

    pd.set_option('display.max_columns', None)
    pd.set_option('display.max_colwidth', None)
    print(df.head(34))

    return rename_columns(df, column_names, drop=True)


if __name__ == '__main__':
    enhance_NCBI_data()
```

Este código también se divide en varias partes que se explican a continuación. En primer lugar, se importan las librerías necesarias y se gestionan las rutas de los archivos de entrada y de salida.

En este caso hay diferentes funciones de transformación en primer lugar encontramos la de rellenar valores faltantes para mantener la consistencia. Después está la de extraer la bibliografía en la que se procesa la columna donde están los identificadores de PubMed y recupera la información de las publicaciones asociadas. Tras esta hay una limpieza de los datos de texto en la que se eliminan números y contenido específico de la columna. Por último, se combina la información de los campos de notas, funciones y evidencia experimental en una solo columna que genera una descripción detallada del elemento.

Para finalizar se organizan y renombran las columnas para que tengan los nombres específicos que se quieren y se guarda el resultado en un archivo CSV.