Universidad Europea

Máster en Bioinformática

ASOCIACIONES DE LA EXPRESIÓN GÉNICA (TRANSCRIPTOMA)-FENOTIPO EN *PMM2-CDG*

Autor: Ricardo Blázquez Martín

Tutor: Vanessa dos Reis Ferreira

Curso 2022-23

Agradecimientos

En el inicio de este proyecto, deseo expresar mi más sincero agradecimiento a todas las personas que me han servido de ayuda. Agradezco especialmente a mi familia y amigos, por su constante ánimo y comprensión. A mis colegas y compañeros de estudio, por las valiosas discusiones y colaboraciones. Sus contribuciones han sido fundamentales para el éxito de este trabajo. Este proyecto no habría sido posible sin su apoyo y confianza.

Índice

Resumen

Los Defectos Congénitos de la Glicosilación (CDG) conforman una creciente familia de trastornos genéticos originados por fallos en los procesos celulares encargados de ensamblar, recortar y añadir glicanos a proteínas y lípidos. En la actualidad, se han identificado alrededor de 160 tipos de CDG y más de 200 fenotipos distintos asociados a ellos, destacando *PMM2-CDG* como el más común. Esta enfermedad se manifiesta con una amplia variedad de síntomas y niveles de gravedad, cuya severidad no depende exclusivamente de las mutaciones en *PMM2*, sugiriendo la influencia de otros genes. Este estudio se centra en explorar las relaciones entre la expresión de genes específicos y las características fenotípicas específicas de *PMM2-CDG*. Para lograrlo, se ha desarrollado una metodología apoyada en herramientas de bioinformática, llevando a cabo análisis de expresión diferencial de genes y análisis de enriquecimiento. Finalmente se ha concluido que existe una correlación entre los pacientes de CDG con variantes en *PMM2* y determinados factores, reforzándose así las ideas previas sobre la relación de esta enfermedad con el sistema inmune y con factores como el TNF-alfa.

Abstract

Congenital Disorders of Glycosylation (CDG) constitute a growing family of genetic disorders arising from failures in cellular processes responsible for assembling, trimming, and adding glycans to proteins and lipids. Currently, approximately 160 types of CDG and over 200 distinct phenotypes associated with them have been identified, with *PMM2-CDG* being the most prevalent. This condition presents a wide range of symptoms and severity levels, where the severity is not solely dependent on *PMM2* mutations, implying the influence of other genes. This study focuses on exploring the relationships between the expression of specific genes and the specific phenotypic characteristics of *PMM2-CDG*. To achieve this, a methodology supported by bioinformatics tools has been developed, conducting differential gene expression analysis and enrichment analysis. Ultimately, it has been concluded that there is a correlation between CDG patients with variants in *PMM2* and specific factors, thereby reinforcing previous notions of the disease's association with the immune system and elements such as TNF-alpha.

Palabras clave

CDG, *PMM2*, glicosilación, herramientas bioinformáticas, secuenciación, RNA-seq.

1. INTRODUCCIÓN Y MARCO TEÓRICO

1.1. Defectos Congénitos de la Glicosilación (CDG)

Los Defectos Congénitos de la Glicosilación (CDG) son una familia creciente de enfermedades genéticas mayoritariamente autosómicas recesivas descritas por primera vez en 1980 por Jaak Jaeken (Jaeken et al., 1980). Se trata de una familia diversa y en rápida expansión causada por defectos en la maquinaria celular responsable de ensamblar, recortar y añadir glicanos a proteínas y lípidos. La mayoría de las CDG consisten en defectos en la N-glicosilación, en la O-glicosilación o en ambas (Verheijen et al., 2020; Francisco et al., 2019; Ferreira et al., 2018). La maquinaria de glicosilación comprende un conjunto de proteínas que procesan, transportan y ensamblan monosacáridos derivados del metabolismo primario en cadenas de glicanos lineales o ramificadas (Varki, 2009).

Actualmente se conocen unos 160 CDG y más de 200 fenotipos diferentes asociados, siendo el *PMM2-CDG* el más prevalente (Ferreira et al., 2018; Péanne et al., 2018). *PMM2-CDG* (MIM: 212065) o fosfomanomutasa 2-CDG, fue el primer defecto de N-glicosilación descrito y es el tipo más común, con aproximadamente 1000 pacientes en todo el mundo (Chang et al., 2018; Jaeken et al., 1980; Vals et al., 2018). Muestra un amplio espectro de signos clínicos y gravedad fenotípica tanto entre diferentes pacientes como dentro del mismo paciente a lo largo de la vida. La afectación neurológica, particularmente hipotonía, ataxia y discapacidad intelectual, es una característica distintiva de esta (y muchas otras) CDG. También suelen verse afectados otros sistemas/órganos, como el tracto gastrointestinal (GI) y el sistema inmunitario (Altassan et al., 2019; Pascoal et al., 2020; Schiff et al., 2017). Dentro de las alteraciones multiorgánicas, podemos encontrar afecciones en los ojos, el hígado, el corazón y el sistema inmunitario (Marques-da-Silva et al., 2017; Pascoal et al., 2020).

Durante la infancia y la edad adulta, esta enfermedad se presenta de manera multisistémica, cursando síntomas como hiporreflexia, convulsiones, episodios de derrame cerebral, neuropatía periférica, distribución anormal de la grasa subcutánea,

rasgos faciales característicos que cambian con la edad, disfunción endocrina, osteopenia y manifestaciones más graves como cardíacas, hepáticas y renales (Lam et al., 1993).

La respuesta inmunitaria depende de innumerables contactos célula-célula y del reconocimiento de antígenos propios y ajenos, y los glicanos están presentes en prácticamente todos los componentes del sistema inmunitario. Por lo tanto, los defectos en la glicosilación pueden tener un impacto diferencial en el estado inmunitario, dando lugar a un amplio espectro de manifestaciones y resultados clínicos (Lyons et al., 2015).

Los CDG son principalmente trastornos multisistémicos complejos que, a pesar de los interesantes avances terapéuticos recientes, siguen careciendo en gran medida de tratamientos eficaces (Iyer et al., 2019; Martínez-Monseny et al., 2019; Yuste-Checa et al., 2017).

Las opciones terapéuticas actuales se limitan principalmente al control de los síntomas, a suplementos de monosacáridos y a fármacos reorientados, pero es cierto que la investigación terapéutica en CDG se ha expandido rápidamente en los últimos años (Pascoal et al., 2020; Brasil et al., 2018). Actualmente hay un total de seis estudios observacionales en curso. Paralelamente, se han producido avances significativos en el desarrollo de modelos de enfermedad y en el descubrimiento de biomarcadores (Chen et al., 2019; Bruneel et al., 2020). Por otro lado, se prevé que la investigación biomédica y clínica impulsada por la terapia CDG se vea obstaculizada por varios retos, muchos de ellos comunes a otras enfermedades raras. Sin embargo, ningún estudio los ha investigado o descrito. Además, apenas se conocen las percepciones, experiencias y expectativas de la comunidad CDG, incluyendo familias, investigadores y profesionales sanitarios, en relación con la investigación y el desarrollo (I+D) de terapias. (Griggs et al., 2009; Augustine et al., 2013).

Los ejercicios de investigación dirigidos por la comunidad y en los que participan múltiples partes interesadas permiten comprender mejor los retos existentes y proponer soluciones. Estos ejercicios presentan varias ventajas, como la mejora de la

calidad de las decisiones, la credibilidad y la aplicación de los avances, así como el fomento de las relaciones entre diferentes sectores y el empoderamiento de la comunidad (Hemati, 2020).

La investigación con métodos mixtos combina la recopilación de datos cuantitativos y cualitativos que son especialmente adecuados para responder a las necesidades de las comunidades. En las enfermedades raras ya se están utilizando métodos mixtos en la investigación terapéutica para evaluar las necesidades sanitarias de los pacientes y diseñar una investigación clínica y traslacional rigurosa centrada en el paciente (Dwyer et al., 2014; Goodspeed et al., 2020; de Freitas et al., 2021).

1.2. Correlación genotipo-fenotipo

Históricamente, a pesar de los estudios realizados que han propuesto algunas relaciones entre el genotipo y el fenotipo dentro de la enfermedad *PMM2-CDG*, es prudente reconocer la existencia de una importante variabilidad fenotípica incluso cuando se trata de pacientes con el mismo genotipo (Lam et al., 1993).

Diversos estudios y análisis retrospectivos, que basan su metodología en la revisión clínica de pacientes que sufren o han padecido esta enfermedad, han sido capaces de destacar ciertas mutaciones genéticas y relacionarlas con los fenotipos más comunes, entre las que se pueden destacar las siguientes (Matthijs et al., 2000):

- Variantes patogénicas en el extremo C-terminal, como *p.His218Leu*, *p.Thr237Met* y *p.Cys241Ser*, se asocian con un fenotipo más leve.
- El genotipo más común, *p.Arg141His*; *p.Phe119Leu*, presenta graves problemas de alimentación, hipotonía, retraso en el desarrollo y disfunción hepática.
- La homocigosidad o la heterocigosidad compuesta para variantes patogénicas con actividad residual casi nula, como *p.Arg141His*, generalmente son incompatibles con la vida.

- También se ha observado un fenotipo grave con alta mortalidad en el genotipo *p.Asp188Gly*;*p.Arg141His*.
- Por otro lado, se han reportado genotipos compuestos heterocigotos, como*p.Thr226Ser*;*p.Ile132Thr* y *p.Arg141His*;*p.Glu139Lys*, que se asocian con un fenotipo más leve sin efusiones pericárdicas, trastornos de coagulación o problemas nutricionales, y algunos individuos pueden caminar de manera independiente.
- La variante patogénica *p.Leu32Arg*, común en Italia, se asocia con un fenotipo más leve, preservando la capacidad de caminar y con un ligero deterioro cognitivo, a pesar de la hipoplasia en la resonancia magnética cerebral.

A pesar de la diferencia entre estudios previos, la mayoría de ellos coinciden en la idea de que la gravedad de la enfermedad no está determinada únicamente por los alelos mutantes de *PMM2*, sino que otros genes probablemente modulan el efecto de las variantes de *PMM2*.Un ejemplo muy representativo de esta conclusión es la variante relativamente común y leve de *ALG6(p.Phe304Ser)*, la cual implica el polimorfismo *c.911T>C* en dicho gen (*ALG6)*, que podría aumentar la gravedad clínica en pacientes con *PMM2-CDG*. Sin embargo, esta información no puede utilizarse actualmente para predecir un resultado clínico informativo válido tanto para los especialistas diagnósticos cómo para los pacientes (Westphal et al., 2002; Bortot et al., 2013).

1.3. Bioinformática en enfermedades raras

En el campo de la investigación de enfermedades genéticas, especialmente aquellas relacionadas con el ADN, se han utilizado ampliamente diversas herramientas y programas de bioinformática. Estas herramientas se han convertido en recursos esenciales para analizar y comprender la complejidad de los trastornos genéticos, incluyendo los Trastornos Congénitos de la Glicosilación (CDG). A través del uso de estas herramientas, los investigadores podemos explorar el genoma, identificar

mutaciones, estudiar la expresión génica y descifrar la base molecular de las enfermedades genéticas.

A lo largo de los años, se han desarrollado diversas herramientas y programas de bioinformática que han sido ampliamente reconocidos y aplicados a nivel internacional en la investigación de enfermedades genéticas. Dentro de las herramientas más utilizadas podemos destacar las siguientes:

- *NCBI* (Centro Nacional de Información de Biotecnología): Es una plataforma integral que proporciona acceso a diversas bases de datos como GenBank, PubMed y dbSNP. Los investigadores podemos utilizar estas bases de datos para buscar secuencias de ADN, estudiar variantes genéticas y acceder a información relevante sobre genes y enfermedades.
- *Ensembl*: Es una base de datos genómica que ofrece una amplia gama de herramientas para el análisis de genes y genomas. Permite la exploración de secuencias de ADN, anotaciones genéticas, variaciones genéticas y la visualización de la estructura genómica en diferentes especies.
- *UCSC Genome Browser*: Es una plataforma interactiva que proporciona acceso a una gran cantidad de datos genómicos y herramientas de visualización. Facilita la exploración y análisis de secuencias genómicas, anotaciones funcionales, variantes genéticas y datos epigenéticos.
- *Variant Effect Predictor* (VEP): Es una herramienta que predice los efectos funcionales de las variantes genéticas. Ayuda a los investigadores a comprender cómo las mutaciones pueden afectar la función de los genes y proteínas, proporcionando información valiosa para el estudio de enfermedades genéticas.
- *GATK* (*Genome Analysis Toolkit*): Es una herramienta bioinformática ampliamente utilizada para el análisis de datos genómicos de alta calidad. Proporciona algoritmos y métodos para el descubrimiento de variantes, SNP y anotaciones funcionales.

Estas herramientas y programas de bioinformática han sido fundamentales en la investigación de enfermedades genéticas, incluyendo el estudio de las CDG. Han permitido a los investigadores analizar datos genómicos, identificar mutaciones, investigar la expresión génica y explorar la base molecular de estas enfermedades. Su uso ha contribuido significativamente a mejorar nuestra comprensión de los mecanismos genéticos subyacentes de las enfermedades genéticas y ha facilitado el desarrollo de posibles enfoques terapéuticos.

En la investigación de enfermedades genéticas como los Trastornos Congénitos de la Glicosilación (CDG), los investigadores enfrentan un desafío al abordar los aspectos inmunológicos de la enfermedad. Aunque existen diversas herramientas y programas disponibles para explorar los aspectos genéticos y de ADN, escasean las herramientas diseñadas específicamente para abordar los componentes inmunológicos de CDG. Esta escasez dificulta la identificación de herramientas adecuadas para explorar los aspectos inmunológicos y obtener una comprensión integral de la enfermedad. Hay que tener en cuenta que el objetivo final siempre es el de dar el diagnóstico más acertado posible para poder determinar el mejor tratamiento o vía de acción para los pacientes, por lo tanto, encontrar herramientas útiles, efectivas y precisas resulta vital.

Para superar este desafío, los investigadores nos vemos en la obligación de explorar las herramientas inmunológicas ya existentes, como la citometría de flujo, que analiza las células inmunológicas e identifica cambios en las poblaciones asociadas a CDG; ELISA, que mide anticuerpos específicos o proteínas inmunológicas, evaluando la respuesta inmunológica en CDG; la inmunohistoquímica, que detecta antígenos en tejidos, proporcionando información sobre las respuestas inmunológicas locales; o la inmunotransferencia, que se emplea para detectar cambios en la expresión de proteínas relacionados con CDG.

Además, la integración de la inteligencia artificial (IA) desempeña un papel fundamental en el abordaje de los aspectos inmunológicos de la CDG. La IA nos permite automatizar la búsqueda y el análisis de la literatura científica, identificando de manera precisa estudios relevantes. Asimismo, proporciona recomendaciones

personalizadas para la selección de herramientas de acuerdo a las necesidades de la investigación. También contribuye a la interpretación de los resultados de las herramientas inmunológicas, permitiendo descubrir patrones y correlaciones entre los aspectos inmunológicos y la CDG.

Un ejemplo claro del uso de la Inteligencia Artificial son los sistemas de apoyo a la toma de decisiones médicas (SAD), que ayudan a seleccionar diagnósticos y tratamientos, acelerar el diagnóstico, corregir diagnósticos erróneos y diagnosticar nuevos pacientes. Los algoritmos de IA mejoran el diagnóstico de enfermedades raras, descubren nuevos mecanismos de enfermedades y utilizan enfoques multiómicos integrados. Sin embargo, el desarrollo y entrenamiento de estos modelos enfrentan limitaciones debido a las pequeñas cohortes de pacientes y las diferencias étnicas. La vasta y desestructurada información clínica en la literatura médica, registros médicos electrónicos y registros de pacientes es un recurso no explotado. Las herramientas de aprendizaje automático se presentan como el futuro del análisis de datos clínicos heterogéneos, pero se necesitan esfuerzos globales para establecer ontologías y procedimientos unificados.

1.4. Abordaje de los retos de las enfermedades raras

Se requiere un mayor progreso para mejorar la investigación y el desarrollo en enfermedades raras. Enfoques centrados en el paciente y tecnologías de IA están mejorando la participación y los resultados de los pacientes, incluyendo la evaluación de resultados reportados por los pacientes, el reclutamiento aumentado a través de las redes sociales y la monitorización de la participación en ensayos clínicos.

La complejidad de los datos biomédicos y los numerosos algoritmos y métodos basados en IA requieren una cuidadosa consideración de las limitaciones técnicas al desarrollar enfoques de IA. Las limitaciones del estudio incluyen la búsqueda solo en una base de datos específica y la posible omisión de datos de otras fuentes. La atención médica, la investigación y la participación de los pacientes están

experimentando un cambio profundo y se necesita una exploración adicional de las posibilidades de la IA. Sin embargo, para un futuro mejor, es necesario abordar cuestiones éticas y legales en la gestión de datos, privacidad y equidad de acceso.

En definitiva, nos encontramos en la era del Big Data, en la que las tecnologías de IA en el ámbito de la biomedicina automatizan tareas y analizan una amplia variedad de datos, abordando eficazmente los desafíos relacionados con las enfermedades raras. La investigación en enfermedades raras, como la CDG, requiere conocimientos y experiencia especializados en el uso de nuevas tecnologías y enfoques de bioinformática. Sin embargo, la falta de personal capacitado y experimentado en este campo está limitando el progreso y la realización de investigaciones de alta calidad en la CDG.

Es crucial abordar estos problemas, dado que las enfermedades raras (ER) presentan desafíos significativos que surgen de su diversidad y la falta de información en comparación con enfermedades más comunes. La implementación de tecnologías de IA permite mejorar el diagnóstico, comprender los mecanismos subyacentes y avanzar en el desarrollo terapéutico en el ámbito de las ER.

1.5. Secuenciación de RNA (RNA-seq)

El empalme (splicing) es el proceso de eliminación de secuencias no codificantes (intrones) de una molécula de ARN seguido de la unión de exones, las regiones codificantes de proteínas de los genes (Rivas et al., 2015). Las variantes en el ADN pueden afectar este proceso, lo que resulta en un empalme incorrecto o alternativo del ARN, como la omisión de secuencias codificantes o la inclusión de secuencias no codificantes en el ARN mensajero (ARNm), lo que puede llevar a una posible pérdida de la función alélica. Los datos de empalme aberrante asociados con una variante de ADN pueden utilizarse como evidencia de patogenicidad, mientras que los datos de empalme normal pueden utilizarse como evidencia de neutralidad (Richards, S. et al., 2015).

La secuenciación de ARN (RNA-seq) ha mostrado un potencial significativo para mejorar el rendimiento diagnóstico y la resolución de las pruebas genéticas de ADN, principalmente debido a los datos funcionales de empalme generados por este análisis. Es importante destacar que el RNA-seq también aborda una limitación técnica que sufren la mayoría de las pruebas genéticas de ADN disponibles clínicamente, las cuales capturan típicamente solo exones y tramos cortos de los intrones adyacentes. Las variantes patogénicas fuera de la secuencia capturada se pasarán por alto con un enfoque basado únicamente en ADN; sin embargo, la adición de RNA-seq proporciona la oportunidad de descubrir empalmes incorrectos causados por eventos intrónicos, lo que lleva a la identificación de variantes patogénicas en la región no codificante de los genes (Scotti et al., 2016).

Los primeros intentos de incorporar el RNA-seq en diagnósticos clínicos han implicado la secuenciación de todo el transcriptoma (WTS) en pacientes con trastornos mendelianos raros que no han recibido un diagnóstico molecular a pesar de someterse a secuenciación de exoma completo (WES) o secuenciación de genoma completo(WGS) (Cummings et al., 2017; Fresard et al., 2019). La adición de esta secuenciación de transcriptoma completo ha demostrado aumentar el rendimiento diagnóstico entre un 7-36%, dependiendo de la enfermedad estudiada. En todos los estudios, se identificaron variantes patogénicas de empalme en regiones normalmente capturadas por los métodos actuales de pruebas de ADN, así como en regiones profundas de intrones, lo que destaca la utilidad del RNA-seq tanto en la identificación como en la interpretación de variantes de empalme causantes de enfermedades (Gonorazky et al., 2019; Lee et al., 2019).

Diversos estudios también han mostrado los beneficios del RNA-seq para los genes de predisposición al cáncer hereditario; sin embargo, este enfoque se ha llevado a cabo tradicionalmente como un seguimiento de las pruebas de ADN que resultaban inconclusas. En un estudio reciente, los resultados de las pruebas genéticas de ARN facilitaron la clasificación del 88% de las variantes de empalme de genes de cáncer como patogénicas o benignas, y se predijo que afectaría a 1 de cada 43 individuos si se realizaba simultáneamente con las pruebas de ADN (Karam et al., 2019). Por lo tanto,

es probable que una proporción sustancial de los pacientes que actualmente reciben pruebas de ADN puedan beneficiarse de la adición de pruebas genéticas de ARN en los protocolos convencionales de diagnóstico de estas enfermedades. Varios estudios también han identificado variantes patogénicas profundas en intrones en una variedad de afecciones de cáncer hereditario, incluyendo el cáncer hereditario de mama y ovario (HBOC), síndrome de Lynch, poliposis adenomatosa familiar, neurofibromatosis y síndrome de Li-Fraumeni. Sin embargo, la prevalencia de las variantes profundas en intrones que predisponen al cáncer no ha sido completamente explorada debido a la limitada escalabilidad de los métodos de pruebas de ARN previos (Landrith et al., 2020).

Para avanzar en el diagnóstico, la secuenciación de ARN (RNA-seq) ha surgido como una herramienta complementaria porque sondea directamente la expresión génica, proporcionando así datos funcionales que respaldan la interpretación clínica de las variantes. Dado que los primeros estudios que utilizaron sistemáticamente RNA-seq para el diagnóstico genético de enfermedades raras se publicaron hace pocos años, el campo es nuevo y carece de flujos de trabajo establecidos (Kremer et al., 2017).

En este contexto, existen protocolos convencionales detallados sobre cómo utilizar datos de RNA-seq para detectar eventos anómalos, como la expresión aberrante, el empalme o splicing alternativo y la expresión monoalélica (MAE) a partir de archivos de formato de alineación de secuencias binarias (BAM) (Li et al., 2009) y archivos de formato de llamada de variantes (VCF) (Danecek et al., 2011). Los protocolos incluyen pasos de control de calidad y gráficos que evalúan los procedimientos de conteo y ajuste del modelo. Para detectar la mezcla o intercambio de muestras, se incluye un paso de validación que compara las variantes llamadas a partir de ARN y ADN. Finalmente, se analizan las muestras individuales utilizando los resultados y objetos derivados del flujo de trabajo.

En un proceso de análisis de datos provenientes de secuenciación, el punto de partida supone la evaluación de la calidad de las reads crudas. Esto implica identificar y, si es necesario, filtrar o eliminar secuencias de baja calidad. La decisión sobre el filtrado

afectará al flujo de trabajo y la cantidad de información disponible. Aunque se debe considerar que no siempre es necesario pre-procesar o filtrar las secuencias, ya que hoy en día se obtienen secuenciaciones de alta calidad. Existen diversas herramientas disponibles para evaluar la calidad de las secuencias en archivos FASTQ. *FASTQC*, una herramienta ampliamente reconocida, procesa archivos FASTQ y genera gráficas detalladas utilizando múltiples métricas para evaluar la calidad de las secuencias, incluyendo estadísticas generales, calidad por base, valores de calidad por secuencia, contenido de bases y contenido de GC por secuencia. *FASTQC* es un programa en Java que se puede ejecutar con una interfaz gráfica o a través de la línea de comandos, generando un informe en formato HTML. En función de los resultados obtenidos se hace un filtrado de secuencias, eliminando las que tengan baja calidad.

El alineamiento de lecturas es un proceso en el que un alineador identifica posibles ubicaciones en un genoma de referencia para cada lectura. En la secuenciación masiva, los fragmentos se superponen, lo que permite secuenciar cada base varias veces y evaluar su confiabilidad. Después de la secuenciación, nos encontramos con millones de secuencias de longitud corta, algunas de las cuales pueden tener múltiples posiciones en el genoma debido a la redundancia genómica, diversidad genética y errores en la preparación. Desde el surgimiento de la secuenciación masiva, se han desarrollado varios alineadores basados en diferentes algoritmos, clasificables según varios criterios, como el tipo y tamaño de las lecturas, el algoritmo subyacente y el origen de la librería. Uno de los alineadores más utilizados es *BWA*, que incluye tres algoritmos: *BWA-backtrack, BWA-SW* y *BWA-MEM*. *BWA-MEM* es preferido en experimentos de Illumina debido a su rapidez y precisión, especialmente para lecturas de 70 pares de bases o más. El alineamiento de secuencias da como resultado un archivo en formato SAM/BAM. El formato SAM es una forma genérica de almacenar alineamientos de secuencias en un genoma de referencia. Es versátil, admite secuencias de diferentes longitudes y es eficiente. Además, consta de dos partes: cabecera (headers) y alineamientos (alignments).

Después del alineamiento, se realiza una segunda fase de control de calidad para evaluar la calidad del proceso. Se revisan las lecturas con calidad de alineamiento

insuficiente y se marcan las lecturas duplicadas. Luego, se recalibran los valores de calidad en el archivo BAM. Generalmente se utilizan herramientas como *Samtools*. Además, es recomendable realizar una inspección visual de los alineamientos utilizando el programa *Integrative Genomics Viewer* (IGV). Las últimas dos fases del protocolo de análisis de secuencias se centran en la identificación de variantes y su posterior filtrado y anotación. En la etapa de identificación de variantes, se busca información relevante, como polimorfismos de un solo nucleótido (SNP), inserciones, deleciones (indels) y cambios en el número de copias de genes, generando una lista de variantes en formato *Variant Call Format* (VCF). Además, se pueden explorar variantes estructurales, como translocaciones e inversiones. En la fase de filtrado y anotación, las variantes identificadas pasan por una validación y filtros de calidad para evitar informar resultados incorrectos. Luego, se utilizan recursos externos, como *dbSNP* y *1000Genomes*, junto con herramientas como *SIFT* y *PolyPhen*, para enriquecer la información, evaluando la frecuencia poblacional y las implicaciones fenotípicas de las alteraciones.

Formato de archivo	Extensión	Descripción
Binary sequence alignment map	.bam	Archivo comprimido binario que contiene las lecturas de secuencia alineadas
FASTA	.fa	Archivo de texto con secuencias de nucleótidos
General transfer format	.gtf	Archivo de texto con anotaciones de genes
Serialized R data	.Rds	Archivo binario comprimido con un único objeto de R
Tab-separated value	.tsv	Archivo de texto con columnas separas por tabuladores
Variant call format	.vcf	Archivo de texto con variantes de secuencias, posición genómica, valor de calidad, etc…

Tabla 1. Diferentes formatos de archivo.

2. HIPÓTESIS Y OBJETIVOS

2.1. Hipótesis

Se postula que existen asociaciones significativas entre la expresión génica y los fenotipos en células de fibroblastos de individuos con *PMM2-CDG* en comparación con individuos sanos (WT), basadas en los genes diferencialmente expresados previamente identificados.

Se espera que la expresión de ciertos genes esté relacionada con características fenotípicas específicas de la enfermedad *PMM2-CDG*, lo que contribuirá a una mejor comprensión de los mecanismos subyacentes de la enfermedad.

2.2. Objetivo primario

El objetivo principal de este trabajo es realizar asociaciones de expresión génica-fenotipo, a partir de los genes expresados diferencialmente entre fibroblastos WT (WildType) y *PMM2-CDG* previamente identificados.

Para ello se deberá seleccionar una plataforma o programa bioinformático adecuado y se desarrollará una metodología de análisis para realizar las asociaciones de todos los genes simultáneamente y presentar los datos obtenidos de manera organizada y fácilmente interpretable, con la intención de evitar el tedioso proceso manual de búsqueda individual de cada uno de los genes.

2.3. Objetivos secundarios

1. **Realizar asociaciones entre los genes diferencialmente expresados y las manifestaciones clínicas de *PMM2-CDG*:**

Para iniciar el proyecto, se deberá seleccionar una plataforma o programa que sea adecuado para realizar las asociaciones de genes con los fenotipos y manifestaciones clínicas que conllevan. Inicialmente se propone la herramienta *Open Target*, que

deberá compararse con las nuevas herramientas encontradas para elegir la más conveniente.

Esta búsqueda de asociaciones entre genes y fenotipos podría realizarse manualmente, pero implicaría mucho tiempo y recursos, por lo tanto, no es idóneo cuando se trata de grandes cantidades de datos. Para ello, se propone desarrollar una metodología que permita automatizar la búsqueda de asociaciones de todos los genes diferencialmente expresados con las manifestaciones clínicas de *PMM2-CDG*, combinándose diferentes programas o software.

A partir de los datos que se hayan obtenido, se deberán presentar los resultados de las asociaciones de manera organizada, con el objetivo de facilitar la interpretación de los datos y proporcionar información relevante sobre cómo los genes que se hayan identificado pueden estar relacionados con las manifestaciones clínicas de *PMM2-CDG.*

2. **Realizar el análisis de datos teniendo en cuenta las manifestaciones clínicas de *PMM2-CDG*:**

Este objetivo se basa en la premisa de la siguiente pregunta científica: ¿Alguna de las asociaciones encontradas previamente explica los fenotipos en *PMM2-CDG*?

Para llevar a cabo este análisis, se usarán los datos que se hayan obtenido en la primera tarea, que se basaba en relacionar los genes con expresión diferencial y las manifestaciones clínicas específicas observadas en *PMM2-CDG*.

La metodología propuesta pretende abordar esta pregunta científica, con el objetivo de analizar estos datos y obtener una mejor comprensión de la relación entre la expresión génica y los fenotipos en esta enfermedad. De esta manera se podrán identificar posibles conexiones entre genes concretos y ciertos fenotipos, lo que nos aportaría información relevante sobre la participación de estos genes en la patogénesis de la enfermedad.

A continuación, se llevará a cabo un análisis de enriquecimiento de funciones para identificar potenciales vías biológicas y procesos celulares que puedan estar implicados

en las manifestaciones clínicas de *PMM2-CDG*. Inicialmente se usarán las librerías "clusterProfiler" y "enrichR", ya que son las librerías más usadas comúnmente para este tipo de análisis y no implican una gran experiencia en programación. No obstante, se valorará el uso de otras librerías más específicas. Se prevé que esta información pueda ayudarnos a comprender cómo la expresión de ciertos genes influye en determinados procesos biológicos y en la presentación clínica de la enfermedad.

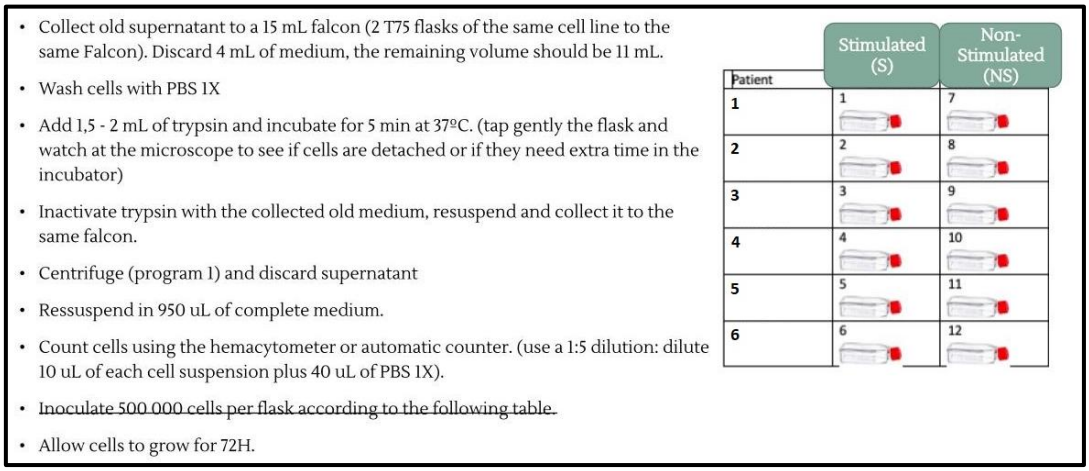
Finalmente, se realizará la validación de resultados para dar veracidad y significado clínico a la información que se haya obtenido hasta el momento. Se buscarán fuentes de datos relevantes, como las "guidelines" de *PMM2-CDG*, los resultados de la encuesta *ImmunoCDGQ* y el cuestionario de priorización de síntomas de CDG, ambas proporcionadas por el tutor, para validar los resultados obtenidos. Comparar los hallazgos del análisis con información adicional fortalecerá la comprensión de las relaciones entre la expresión génica y los problemas clínicos en *PMM2-CDG*.

Se presentarán los hallazgos obtenidos en este análisis de datos que tengan en cuenta las manifestaciones clínicas de *PMM2-CDG*, y se discutirá si alguna de las asociaciones identificadas previamente puede explicar los fenotipos observados en la enfermedad. Los resultados serán interpretados y validados con otras fuentes en el contexto de la metodología aplicada y de las características de la enfermedad para poder dar un significado biológico a los hallazgos.

3. MATERIAL Y MÉTODOS

Los problemas inmunitarios afectan profundamente a la calidad de vida de los pacientes con CDG y son responsables del 20% de la mortalidad infantil. A pesar de que 23 tipos distintos de CDG se caracterizan por presentar alteraciones inmunológicas, el estudio de su inmunopatología ha sido descuidado. Para abordar este problema, las muestras que son objeto de nuestro estudio, tanto los fibroblastos de *PMM2-CDG* (el CDG más común) como los fibroblastos Wild Type (WT), fueron estimulados con la citoquina proinflamatoria TNF-α. Posteriormente, el ARN fue secuenciado utilizando tecnología de *Illumina, Inc*.

Los fibroblastos fueron preparados siguiendo un protocolo concreto basado en un cultivo en PBS con tripsina:



- Collect old supernatant to a 15 mL falcon (2 T75 flasks of the same cell line to the same Falcon). Discard 4 mL of medium, the remaining volume should be 11 mL.
- Wash cells with PBS 1X
- Add 1,5 - 2 mL of trypsin and incubate for 5 min at 37ºC. (tap gently the flask and watch at the microscope to see if cells are detached or if they need extra time in the incubator)
- Inactivate trypsin with the collected old medium, resuspend and collect it to the same falcon.
- Centrifuge (program 1) and discard supernatant
- Ressuspend in 950 uL of complete medium.
- Count cells using the hemacytometer or automatic counter. (use a 1:5 dilution: dilute 10 uL of each cell suspension plus 40 uL of PBS 1X).
- Inoculate 500 000 cells per flask according to the following table.
- Allow cells to grow for 72H.

Figura 1. Protocolo para la preparación de fibroblastos.

Inicialmente se dispone de una cohorte inicial de 12 pacientes, dividida en 2 principales grupos: controles negativos y pacientes de *PMM2-CDG*. A su vez, esta misma cohorte queda segmentada entre las muestras que han sido estimuladas con TNF-alfa y las muestras cuyo crecimiento no ha sido estimulado con esta citoquina. Para el desarrollo de este proyecto es necesario discriminar las muestras que han sido estimuladas con la citoquina proinflamatoria TNF-alfa, puesto que dentro del análisis

que se va a realizar de búsqueda de relaciones entre el genotipo y el fenotipo, es necesario determinar si influye o no la estimulación previa de estos fibroblastos.

Por lo tanto, para este trabajo, la cohorte de pacientes queda reducida a tan solo 6 pacientes, de los cuales, 3 son controles negativos y los otros 3 son afectados por la enfermedad *PMM2-CDG*.

ID	Gen mutado	Mutación	Género	Edad de muestreo
4	PMM2	c.95TA>GC/c.470T>C (p.L32R/p.F157S)	Mujer	7
5	PMM2	c.422G>A/c.647A>T (p.R141H/p.N216I)	Hombre	1
6	PMM2	c.422G>A/c.415G>A (p.R141H/p.E139K)	Mujer	5
1	-	-	Hombre	3
3	-	-	Hombre	10
2	-	-	Mujer	2

Tabla 2. Muestras utilizadas en el análisis.

Las secuencias obtenidas de la secuenciación con tecnología de Illumina han sido analizadas siguiendo la plataforma integrada en *Illumina*, *Bluebee*. El protocolo o "pipeline" empleado comienza con una carpeta comprimida ZIP, que contiene los archivos en formato FASTQ (.fastq). A continuación, se realiza el análisis de la calidad de las secuencias, que genera un informe en formato HTML (página web) que documenta el análisis de calidad. A partir de los resultados del análisis de calidad, los archivos son procesados y recortados para seleccionar las secuencias que cumplen los estándares mínimos de calidad y eliminar las secuencias de peor calidad, que pueden haber sido contaminadas con adaptadores o colas de poli-A. El siguiente paso es el alineamiento, el cual ha sido realizado con la herramienta *STAR* (*Spliced Transcripts*

Alignment to a Reference). Se trata de un software basado en un algoritmo de alineamiento de secuencias de ARN-seq que utiliza la búsqueda secuencial de semillas (Dobin et al., 2013). Como resultado se obtienen archivos en formato BAM (.bam), que son archivos comprimidos binarios que contienen las lecturas de secuencia alineadas. Posteriormente se lleva a cabo la etapa de identificación de variantes, donde se busca información relevante, como polimorfismos de un solo nucleótido (SNP) e indels (inserciones y deleciones) de origen germinal o somático. Esta fase produce una lista de variantes en formato *Variant Call Format* (VCF). Como resultado, en las muestras de pacientes de *PMM2-CDG* se identifican las variantes que aparecen en la tabla 2.

Finalmente, se realiza un Análisis de Expresión Diferencial de Genes (DEG) que es una técnica bioinformática que se utiliza para identificar genes cuya expresión varía significativamente entre diferentes condiciones experimentales, como diferentes grupos de muestras o tratamientos. Este análisis es fundamental en la investigación genética y biológica, ya que permite comprender qué genes están activados o desactivados en respuesta a ciertos estímulos o condiciones.

Los resultados obtenidos a través del análisis de expresión diferencial de genes proporcionan información valiosa sobre cómo la expresión génica se relaciona con los fenotipos observados en un estudio. Los DEG desempeñan un papel esencial en la investigación genética y biológica. Permiten identificar biomarcadores que pueden servir para diagnóstico, pronóstico o seguimiento de enfermedades. Además, son fundamentales para comprender los mecanismos moleculares que subyacen a una amplia variedad de procesos biológicos, como el desarrollo, la respuesta inmunitaria, la regulación celular y la patogénesis de enfermedades. También son cruciales en la identificación de dianas terapéuticas potenciales.

Como resultado del Análisis de Expresión Diferencial de Genes, se obtiene una tabla en formato Excel (.xlsx), donde principalmente se incluyen los diferentes genes que se encuentran diferencialmente expresados. Dicha tabla puede consultarse en los anexos. Concretamente, se realiza una separación entre diferentes grupos de datos:

Por un lado, se muestran todos los genes que tienen una expresión aberrante dentro del grupo control, lo cual puede deberse a múltiples causas, como podría ser cualquier otra enfermedad genética sin relación con nuestra enfermedad de estudio. También se incluyen los genes con expresión aberrante dentro del grupo de pacientes de *PMM2-CDG*. Esta lista de genes, es de especial interés en nuestro análisis puesto que se debe estudiar la posible correlación de estos genes con la enfermedad *PMM2-CDG* y por lo tanto, con los fenotipos más característicos de dicha enfermedad.

Por otro lado, también se incluye información adicional derivada de las 2 principales tablas ya mencionadas. Primeramente, se incluye la lista completa de los 13507 genes analizados en el estudio. Además, se incluyen 2 listas adicionales, donde se muestran los genes con expresión aberrante positiva, es decir, sobre-expresados, en los pacientes de *PMM2* y los genes con expresión aberrante negativa, es decir, sub-expresados, en los pacientes de *PMM2*. También se incluyen otras 2 listas más en las que se incluye la misma información anterior, con los genes sobre-expresados y sub-expresados, pero en este caso en los pacientes control.

Es importante destacar que la información que se representa en estas tablas, no solo muestras los genes diferencialmente expresados, sino también los valores resultantes del análisis DEG, que son utilizados para discriminar estos genes con expresión aberrante de los genes con expresión normal. Los valores que se tienen en cuenta son los siguientes:

- GeneName: Esto se refiere al nombre del gen o genes en estudio. Cada fila de la tabla representa un gen diferente.
- logFC (*Logarithm of FoldChange*): El logFC es una medida de cuánto cambia la expresión de un gen en comparación con una condición de referencia. Es una métrica logarítmica que representa la magnitud del cambio en la expresión génica. Un logFC positivo indica un aumento en la expresión, es decir, sobre-expresión, mientras que un logFC negativo indica una disminución, es decir, sub-expresión.
- logCPM (*Logarithm of Counts Per Million*): El logCPM es una métrica que normaliza la cantidad de lecturas de ARNm (conteo) para un gen en función del

tamaño de la biblioteca y lo presenta en una escala logarítmica. Se utiliza para comparar la abundancia relativa de un gen entre muestras. Cuanto mayor sea el valor de logCPM, mayor será la abundancia relativa del gen en las muestras.

- F (*Valor F*): El valor F es una estadística que se utiliza en pruebas de diferencias entre grupos. En este contexto, se utiliza para evaluar si las diferencias en la expresión de un gen son estadísticamente significativas. Un valor F alto indica que las diferencias en la expresión del gen son significativas.
- PValue (*Valor p*): El valor de p es una medida de la probabilidad de obtener los resultados observados si no hubiera diferencias reales en la expresión del gen. En este caso, un valor de p muy bajo (muy cercano a cero) sugiere que las diferencias en la expresión del gen son altamente significativas desde un punto de vista estadístico. En otras palabras, es poco probable que las diferencias observadas sean el resultado del azar.
- FDR (*Tasa de Descubrimiento Falso*): La FDR es una corrección estadística que controla el riesgo de obtener resultados falsos positivos en un análisis. En este contexto, se utiliza para ajustar los valores de p y controlar el error en la prueba estadística debido a la realización de múltiples pruebas. Un valor de FDR bajo (en este caso, muy cercano a cero) indica que las diferencias en la expresión génica son altamente confiables y no se deben a errores en el análisis.

A partir de todos los datos obtenidos hasta el momento, se ha realizado un análisis de enriquecimiento mediante el software “enrichR”. EnrichR nos permite determinar las vías biológicas, procesos celulares y funciones moleculares que pueden estar relacionados con nuestro conjunto de genes identificados como DEG. Esta interpretación se logra mediante el análisis de enriquecimiento, que compara las listas de genes proporcionadas con una extensa colección de bases de datos de vías biológicas y ontologías. A través de este proceso, *EnrichR* identifica y prioriza las relaciones funcionales más relevantes, lo que nos permite obtener una visión más profunda y completa de sus datos.

```r
# Instalo y cargo los paquetes necesarios
if (!requireNamespace("BiocManager", quietly = TRUE)) {
  install.packages("BiocManager")
}
BiocManager::install("enrichR")
BiocManager::install("enrichRData")
library(enrichR)
library(enrichRData)

# Compruebo si la opción websiteLive está habilitada
websiteLive <- getOption("enrichR.live")

if (websiteLive) {
  # Obtén la lista de bases de datos disponibles si websiteLive está habilitado
  dbs <- listEnrichrDbs()
  if (is.null(dbs)) websiteLive <- FALSE
}

# Lista de genes ejemplo
lista_genes <- c("gen1", "gen2")

# Lista de bases de datos
dbs <- c("GO_Molecular_Function_2015", "GO_Cellular_Component_2015", "GO_Biological_Process_2015", "KEGG_2019_Human")

if (websiteLive) {
  # Realizo el análisis de enriquecimiento
  enriched <- enrichr(gene_list = lista_genes, organism = "Homo sapiens", library = dbs)

  # Grafica de los resultados
  if (!is.null(enriched)) {
    plotEnrich(enriched[[1]], showTerms = 20, numChar = 40, y = "Count", orderBy = "P.value")
  }
}
```

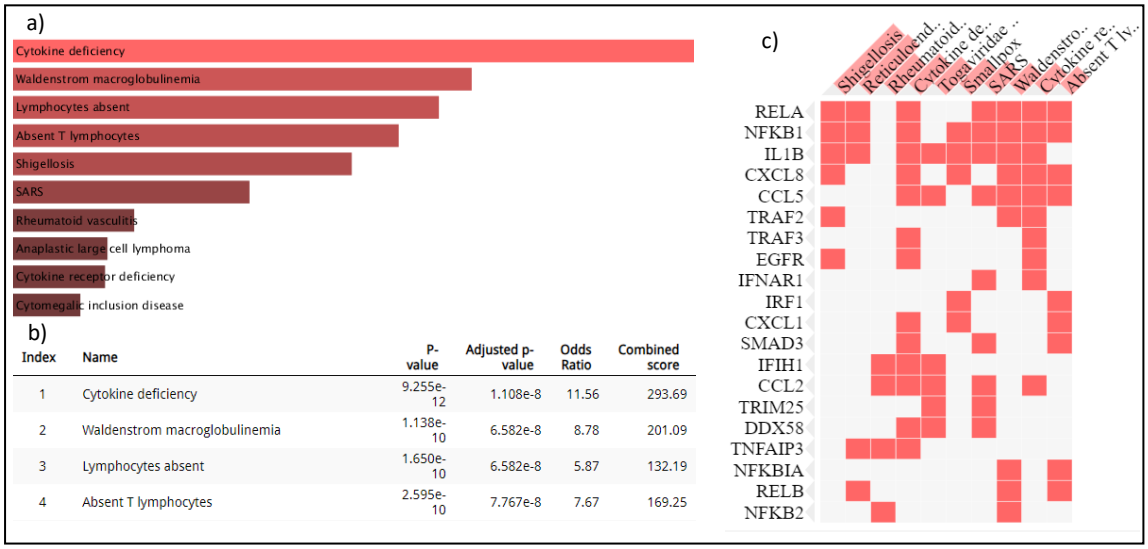
Figura 2. Script de análisis de enriquecimiento con enrichR.

Teniendo en cuenta que este proyecto se centra en encontrar una metodología útil y sencilla para los usuarios, considero importante resaltar que la versión en línea de *EnrichR* puede considerarse como una elección más recomendable. Personalmente, he encontrado que esta versión en línea se ajusta perfectamente a las necesidades de los usuarios que carecen de conocimientos de programación suficientes para los análisis de datos genómicos. Su sitio web (https://maayanlab.cloud/Enrichr/) tiene una interfaz de entrada intuitiva y accesible que proporciona una experiencia de usuario sin complicaciones.

El funcionamiento de la versión en línea destaca por su sencillez. Los usuarios pueden cargar sus listas de genes o, si lo prefieren, emplear listas predefinidas para llevar a cabo el análisis de enriquecimiento. A continuación, se ejecutan los cálculos necesarios y se establecen comparaciones con una amplia gama de bases de datos disponibles. Algunas de las bases de datos más representativas son: *Gene Ontology (GO) Biological Process, KEGG Pathway, Reactome Pathways* o *Disease Ontology* entre otras. Los resultados se presentan de forma comprensible y visual, a menudo acompañados de gráficos y tablas que resaltan las vías, funciones o términos más significativos. Cada informe resultante sirve de guía para dotar de significado biológico a los datos.

En este caso, se introdujo la lista de genes diferencialmente expresados en los pacientes de *PMM2-CDG*, y se seleccionaron varias bases de datos de referencia para comparar los resultados. Concretamente me he centrado en las siguientes bases de información: *RareDiseasesGeneRIF Gene Lists*, que muestra genes asociados con enfermedades raras basándose en búsquedas de PubMed; *KEGG 2021 Human*, que es una base de rutas metabólicas y de señalización de células humanas; *Reactome 2022*, que también se basa en rutas metabólicas; *BioPlanet 2019*, que es un recurso informático que cataloga las diferentes rutas; y *Jensen TISSUES*, que emplea minería de texto para identificar relaciones entre genes y tejidos humanos.

La búsqueda dentro de cada una de estas bases nos muestra diferentes resultados y nos permite visualizar de varias maneras la información. Primero aparece una gráfica de barras que nos ordena los valores según el p-valor; a continuación, se muestra una tabla, que sigue el mismo criterio de ordenación (p-valor), pero que además muestra otros valores como el p-valor ajustado, el odds ratio y el valor score combinado; y finalmente podemos visualizar un *clustergram,* que es un diagrama de agrupamientos que nos muestra patrones, tendencias o relaciones ocultas en los datos.



Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Cytokine deficiency	9.255e-12	1.108e-8	11.56	293.69
2	Waldenstrom macroglobulinemia	1.138e-10	6.582e-8	8.78	201.09
3	Lymphocytes absent	1.650e-10	6.582e-8	5.87	132.19
4	Absent T lymphocytes	2.595e-10	7.767e-8	7.67	169.25

Figura 3. Diferentes visualizaciones de resultados en enrichR: a) Gráfica de barras. b) Tabla de valores. c) Clustergram

4. RESULTADOS

Para la realización de la primera tarea/objetivo, que consiste en seleccionar la base de datos o plataforma adecuada para realizar las asociaciones de genes y fenotipos o manifestaciones clínicas, se propuso la elaboración de una presentación en formato PowerPoint donde se resumiera toda la información necesaria respecto a la búsqueda, evaluación y elección de las herramientas bioinformáticas más adecuadas para este proyecto.

Dicha presentación ha sido grabada en formato .mp4 con el objetivo de ser mostrada en el "6th World Conference for CDG", como miembro de la red *CDG and Allies*. El título de la presentación es: "Overview of Omics and Bioinformatic tools that can be applied for CDG". *[1]

La metodología utilizada para la selección de las herramientas bioinformáticas adecuadas se divide en 4 pasos:

-El primer paso consiste en hacer una búsqueda exhaustiva y detallada en diferentes bases de datos científicas, como *PubMed* y *Scopus*, y en literatura especializada en enfermedades raras, con el objetivo de identificar una serie de herramientas o plataformas que nos resulten potencialmente útiles para el análisis de datos genómicos. Se ha considerado un punto importante que dichas herramientas se puedan aplicar a enfermedades raras, como es el caso de *PMM2-CDG*.

-En segundo lugar, a partir de la búsqueda realizada, se marcaron unos criterios claros de selección para identificar las herramientas más adecuadas para este estudio. Los criterios que se emplearon incluyen la aplicabilidad de las herramientas en casos de enfermedades raras, su disponibilidad y accesibilidad para todo tipo de usuarios, así como las funcionalidades necesarias para poder realizar análisis de este tipo.

-En tercer lugar, cada herramienta identificada en el paso anterior, fue sometida a una evaluación exhaustiva. En concreto, se consultó la documentación técnica y científica proporcionada en las páginas webs oficiales de cada producto. Además, se tuvieron en

*[1] Dicha presentación se encuentra en los anexos.

cuenta las opiniones y recomendaciones de expertos en el campo de la bioinformática y las enfermedades raras.

-Finalmente se realizó la comparación y selección definitiva de las herramientas. Después de la evaluación, se compararon detalladamente las herramientas más prometedoras. Para ello, se evaluaron aspectos clave como la capacidad de análisis, la funcionalidad, la interfaz de usuario o el soporte técnico, y se seleccionaron las herramientas más adecuadas, dando especial importancia a que fueran compatibles con las necesidades específicas de este proyecto. Los resultados de esta búsqueda se muestran en la tabla 3.

HERRAMIENTA	FUNCIÓN	PROPÓSITO	DIFICULTAD	VENTAJAS	DESVENTAJAS
IMGT/HighV-QUEST	Herramienta en línea para el análisis de secuencias de genes de receptores de linfocitos B y T	Estudiar la diversidad y función de los receptores inmunitarios	Baja	Interfaz en línea fácil de usar	Limitado al análisis de secuencias de genes de receptores de linfocitos B y T
V(D)J-Seq Tools	Herramientas bioinformáticas para el análisis de secuencias de genes V(D)J en linfocitos B y T	Estudiar la recombinación de genes V(D)J y la diversidad clonal	Moderada	Foco específico en el análisis de secuencias de genes V(D)J	Requiere ciertos conocimientos de bioinformática y preprocesamiento de datos
MiXCR	Herramienta bioinformática para el análisis de secuencias de linfocitos B y T	Estudiar la recombinación de genes V(D)J, la diversidad clonal y el análisis de expresión	Moderada	Análisis completo de la recombinación de genes V(D)J y la expresión	Requiere instalación y uso de línea de comandos, puede tener una curva de aprendizaje para usuarios no técnicos
Combinación de varios softwares	Una combinación de varias herramientas bioinformáticas para el análisis integral de secuencias genéticas	Estudiar multitud de aspectos de los datos de secuencias genéticas	Moderada	Análisis flexible y personalizable. Abordar muchas necesidades de análisis de secuencias genéticas	Puede requerir experiencia en integrar y usar múltiples herramientas de software

Tabla 3. Herramientas bioinformáticas seleccionadas como adecuadas para el proyecto.

Como resultado he llegado a la conclusión de que optar por una combinación de varios softwares bioinformáticos en lugar de las tres primeras herramientas mostradas en la tabla 3 es preferible debido a su flexibilidad y adaptabilidad. Mientras que las herramientas específicas pueden ser limitadas en cuanto a su capacidad de análisis, la combinación permite seleccionar software especializado para cada aspecto de la investigación genética, lo que se traduce en resultados de mayor calidad y precisión. La única dificultad que puede generar el uso de varios softwares es la necesidad de tener conocimientos mínimos por parte de los profesionales, no obstante, la mayoría de herramientas mencionadas antes también implican ciertos conocimientos. Además, esta estrategia ofrece la ventaja de acceso a una amplia variedad de opciones en la

comunidad científica, lo que permite mantenerse actualizado con las últimas técnicas y herramientas. También facilita el aprendizaje, ya que se pueden elegir herramientas que se adapten al nivel de experiencia del usuario, desde interfaces de usuario amigables hasta software más avanzado. Finalmente, es potencialmente más económico, ya que solo se utilizan las herramientas necesarias, evitando gastos en herramientas costosas que pueden no ser relevantes para el análisis requerido.

En cuanto al análisis de enriquecimiento se consideró que la mejor opción era utilizar el programa enrichR en su versión online, para poder mostrar de manera más sencilla su funcionamiento y explicar los diferentes resultados que nos aporta. Se tomó esta decisión teniendo en cuenta que el objetivo es crear una metodología que pueda ser repetida por usuarios sin experiencia, está especialmente pensadas para personas con conocimientos biológicos pero que tengan limitaciones informáticas. De esta manera se facilita en gran medida la elaboración del análisis gracias a que tiene una interfaz muy sencilla, y se deja en manos del investigador la interpretación posterior de los resultados.

En este caso, como ya he mencionado con anterioridad, para el análisis de enriquecimiento se usaron diferentes bases de datos. A continuación, se muestran los resultados obtenidos en función de cada base de datos:

RareDiseasesGeneRIF Gene Lists

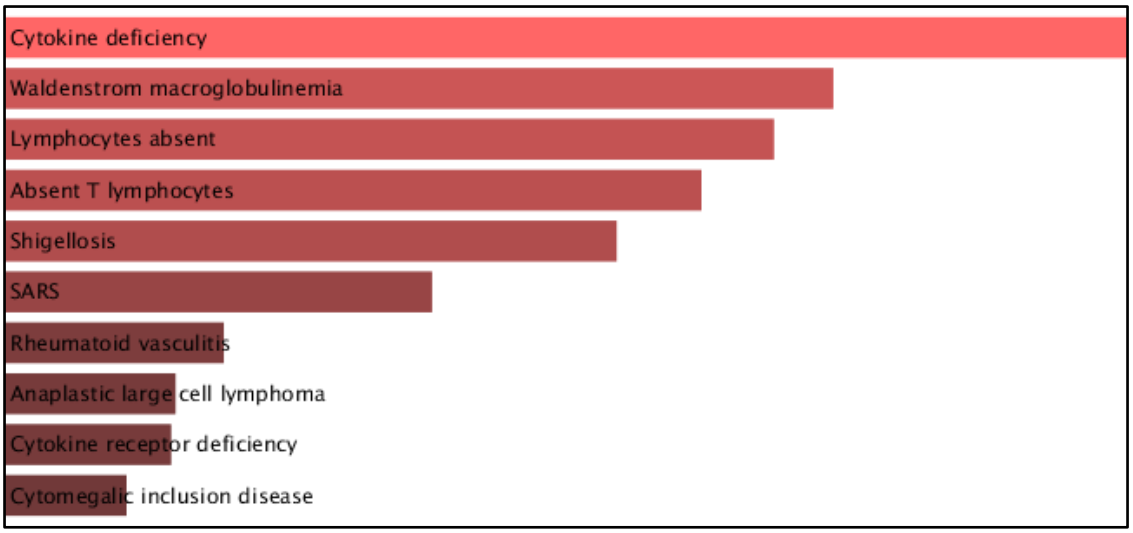


Figura 4. Diagrama de barras de la base *RareDiseasesGeneRIF Gene Lists*

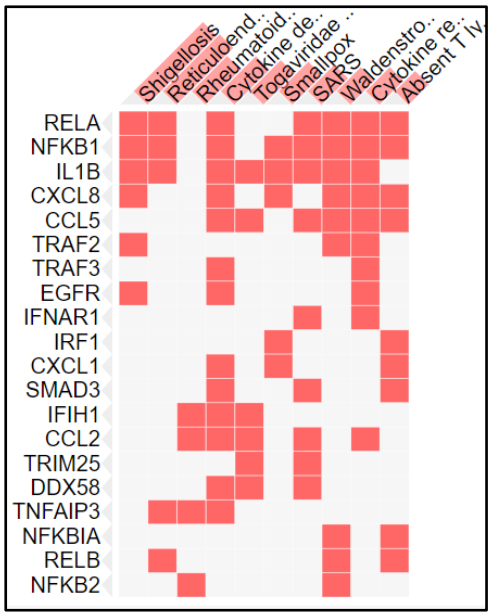


Figura 5. Clustergram de la base *RareDiseasesGeneRIF Gene Lists*

La información que podemos extraer de estos resultados (figura 4) gracias a esta base de datos es que nuestra lista de genes está correlacionada de manera muy significativa con la deficiencia de citoquinas o del receptor de citoquinas. Además, en la figura 5, apreciamos que dentro de los genes que mayor correlación tienen con la deficiencia de citoquinas, están algunos de los genes que aparecen en el archivo de Excel clasificados como DEG. Concretamente se trata de los genes que se encuentran sub-expresados en los casos de *PMM2-CDG* (*TNFAIP3, CCL2, CXCL8*).

BioPlanet 2019

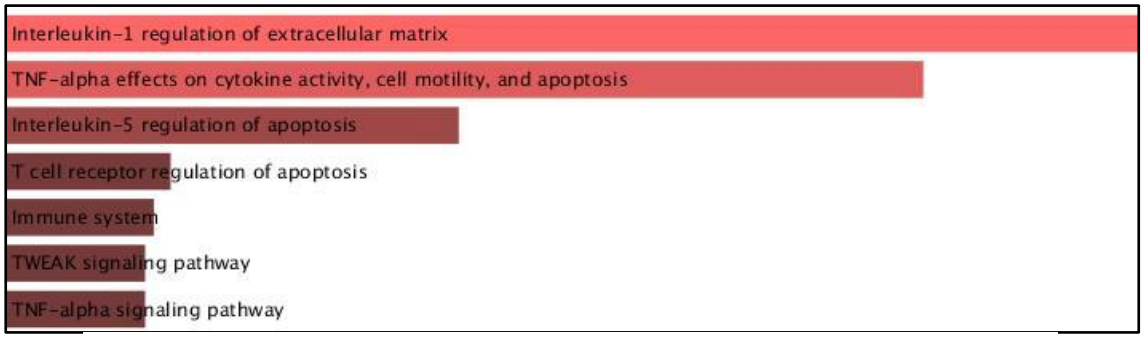


Figura 6. Diagrama de barras de la base *BioPlanet 2019.*

La información que podemos extraer de estos resultados (figura 6) gracias a esta base de datos es que nuestra lista de genes está correlacionada, entre otras cosas, con los efectos de TNF-alfa sobre las citoquinas. Además, en la figura 7, apreciamos que dentro de los genes que mayor correlación tienen con TNF-alfa, están algunos de los genes que aparecen en el archivo de Excel clasificados como DEG. Concretamente se trata de los genes que se encuentran sub-expresados en los casos de *PMM2-CDG* (*NFBIA, NFKB2, RELB, CCL2, CXCL2, TNFAIP3, SOD2, BIRC3, CXCL8*).

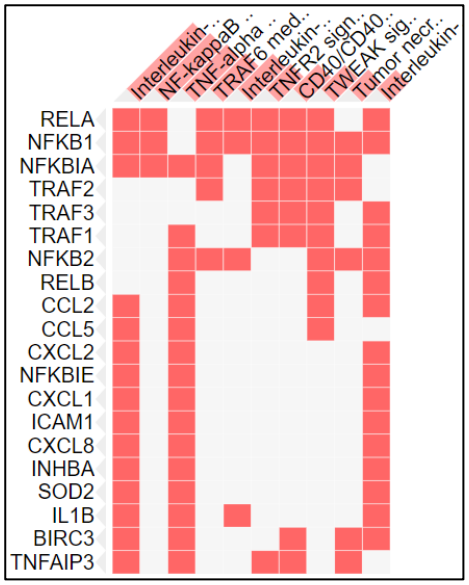


Figura 7. Clustergram de *BioPlanet 2019.*

Reactome 2022

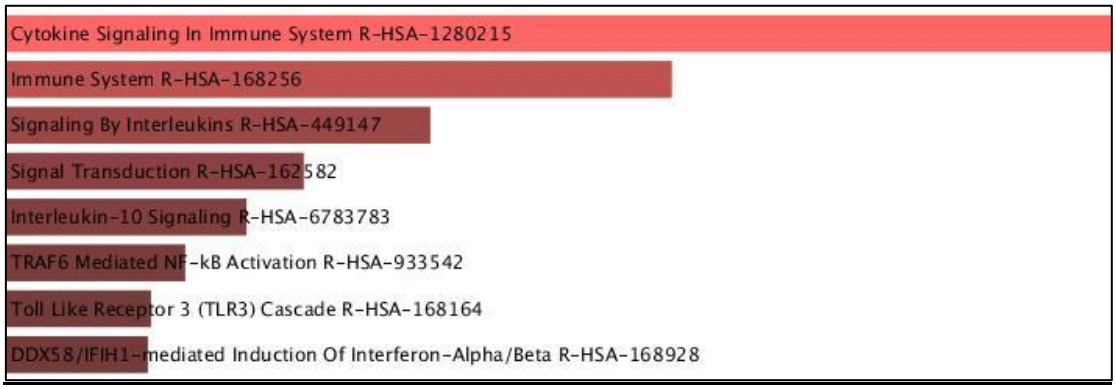


Figura 8. Diagrama de barras de la base *Reactome 2022.*

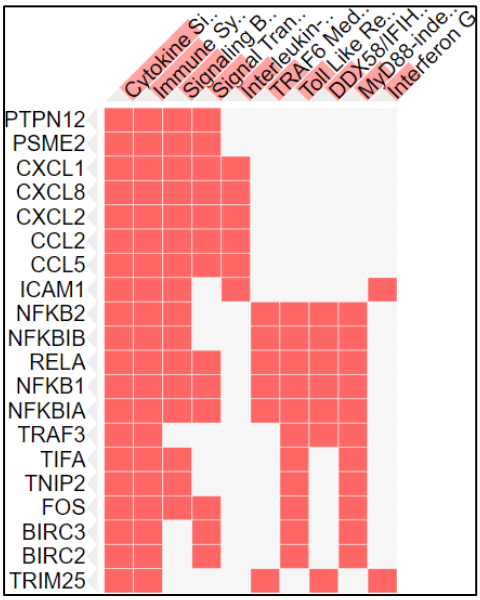


Figura 9. Clustergram de *Reactome 2022.*

La información que podemos extraer de estos resultados (figura 8) gracias a esta base de datos es que nuestra lista de genes está correlacionada de manera muy significativa con la señalización de citoquinas dentro del sistema inmune, lo que es muy representativo de la enfermedad de estudio. Además, en la figura 9, apreciamos que dentro de los genes que mayor correlación tienen con las citoquinas del sistema inmune, están algunos de los genes que aparecen en el archivo de Excel clasificados como DEG. Concretamente se trata de los genes que se encuentran sub-expresados en los casos de *PMM2-CDG* (*BIRC3, CXCL2, CXCL8, CCL2, NFKB2*).

KEGG 2021 Human

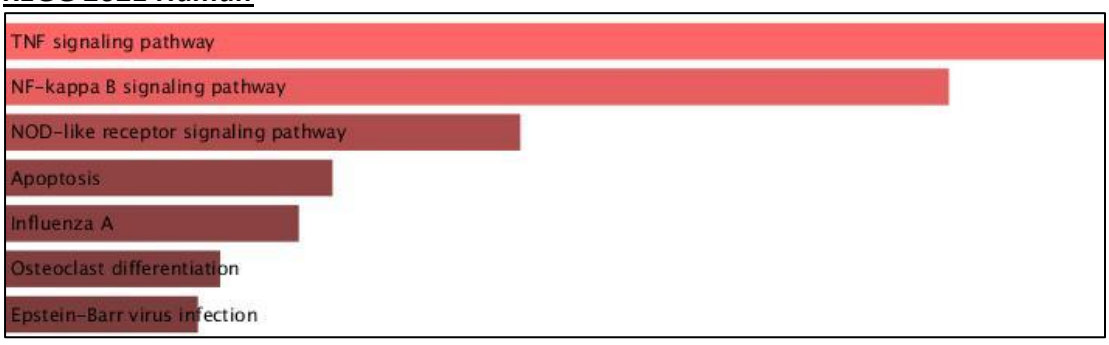


Figura 10. Diagrama de barras de *KEGG 2021 Human.*

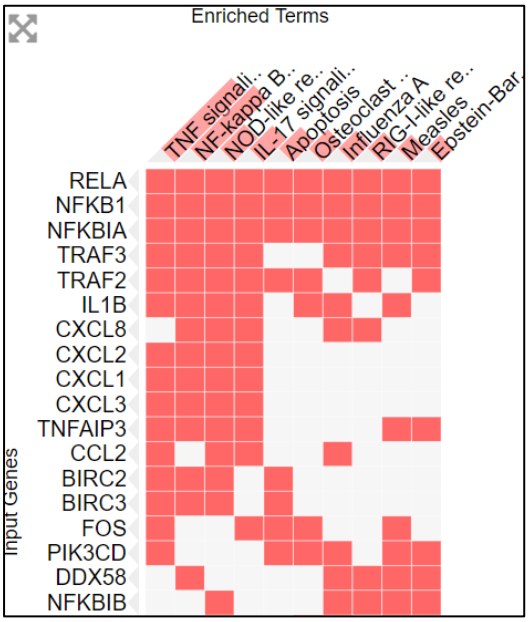


Figura 11. Clustergram de *KEGG 2021Human.*

La información que podemos extraer de estos resultados (figura 10) gracias a esta base de datos es que nuestra lista de genes está correlacionada de manera muy significativa con las rutas de señalización de TNF, algo que es representativo de esta enfermedad. Además, en la figura 11, apreciamos que dentro de los genes que mayor correlación tienen con TNF, están algunos de los genes que aparecen en el archivo de Excel clasificados como DEG. Concretamente se trata de los genes sub-expresados en *PMM2-CDG* (*TNFAIP3, CCL2, CXCL8, CXCL2, BIRC3, NFKBIB*).

Jensen TISSUES

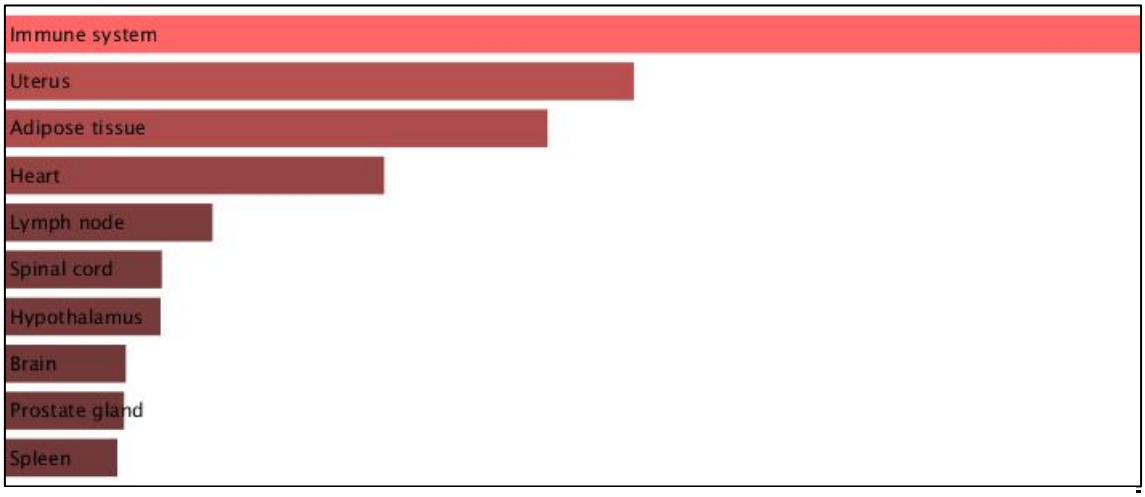


Figura 12. Diagrama de barras de *Jensen TISSUES.*

La información que podemos extraer de estos resultados (figura 12) gracias a esta base de datos es que nuestra lista de genes está correlacionada de manera muy significativa con el sistema inmune. Esto demuestra y refuerza el concepto preestablecido de que la enfermedad *PMM2-CDG* está estrechamente relacionada con las deficiencias inmunológicas. Por lo tanto todos estos genes clasificados como DEG, se correlacionan con *PMM2* en su papel en el sistema inmune.

5. DISCUSIÓN Y CONCLUSIONES

Afrontar los retos de las enfermedades raras implica lidiar con dificultades significativas. La principal razón radica en la necesidad de abordar estos desafíos en el contexto de enfermedades raras (ER). La falta de personal capacitado y experimentado en este campo puede limitar el progreso y la realización de investigaciones de alta calidad en las ER. Además, debido a la naturaleza heterogénea y a la escasez de información disponible en las ER, es fundamental recurrir a enfoques basados en la inteligencia artificial (IA) para un mejor diagnóstico y tratamiento.

Es crucial abordar estos problemas, dado que las ER presentan desafíos significativos que surgen de su diversidad y la falta de información en comparación con enfermedades más comunes. La implementación de tecnologías de IA permite mejorar el diagnóstico, comprender los mecanismos subyacentes y avanzar en el desarrollo terapéutico en el ámbito de las ER.

La investigación en enfermedades raras, como la CDG, requiere conocimientos y experiencia especializados en el uso de nuevas tecnologías y enfoques de bioinformática. Al explorar este tema, estamos abordando un campo relativamente nuevo y poco estudiado, lo que resulta valioso para expandir nuestro conocimiento en esta área y colmar lagunas en investigaciones previas, como las realizadas por Carlota Pascoal, que han servido como base para nuestra investigación.

La necesidad de la comunidad de pacientes con CDG fue el principal motor de este proyecto. Las enfermedades raras plantean numerosos desafíos, como diagnósticos precisos, opciones de tratamiento efectivas y comprensión limitada de las enfermedades en sí. Por tanto, este proyecto se orienta a abordar estas necesidades y mejorar la atención y los resultados para los pacientes. El objetivo es encontrar las mejores herramientas que puedan ayudar a la investigación de la CDG, siempre teniendo en cuenta la información proporcionada por los pacientes sobre su enfermedad, incluyendo síntomas, diagnósticos, emociones, sentimientos y los problemas cotidianos que enfrentan. En este contexto, las redes sociales, los teléfonos

inteligentes y las nuevas tecnologías en general pueden desempeñar un papel fundamental al permitir que los pacientes expresen sus necesidades.

Sin embargo, uno de los grandes problemas para el análisis de secuencias en casos de enfermedades raras es la escasez de muestras, ya que los casos reales son muy pocos. Esto dificulta la función de los softwares y programas informáticos que a menudo requieren grandes cantidades de datos para proporcionar resultados sólidos. Además, existe una carencia de herramientas realmente útiles diseñadas específicamente para analizar secuencias en casos de enfermedades raras. La combinación de la falta de muestras y la limitada disponibilidad de herramientas adecuadas destaca la urgente necesidad de más investigación y desarrollo en esta área, así como la formación de profesionales con los conocimientos necesarios para llevar a cabo estos análisis de manera efectiva.

En cuanto a los resultados concretos de este trabajo, puedo concluir que existe una correlación significativa entre los pacientes de CDG con variantes en *PMM2* y diferentes factores como: la deficiencia de citoquinas, los efectos de TNF-alfa sobre las citoquinas, la señalización de citoquinas del sistema inmune, las rutas de señalización de TNF o con el sistema inmune en general. Esto se ha podido corroborar mediante análisis de expresión diferencial de genes y análisis de enriquecimiento. Con el primer análisis se han determinado los genes que se encuentran sobre-expresados o sub-expresados en los casos de pacientes con *PMM2-CDG* y con el segundo análisis se han establecido las relaciones de estos genes con determinadas rutas metabólicas, vías biológicas o simplemente tejidos humanos. Estas conclusiones fortalecen las ideas preconcebidas sobre esta enfermedad sobre su estrecha relación con el sistema inmune y con determinados elementos como el TNF-alfa.

6. PLAN DE DATOS Y ASPECTOS ÉTICOS Y LEGALES

En este trabajo se abordará el uso de datos FAIR proporcionados por la organización "CDG and Allies". Estos datos han sido facilitados en formato Excel (.xlsx) y contienen información relevante para el estudio de los Defectos Congénitos de la Glicosilación (CDG). Estos datos no son de dominio público y por lo tanto, serán tratados siguiendo las medidas de confidencialidad y privacidad pertinentes.

Se seguirán los principios FAIR para garantizar que estos datos sean almacenados y gestionados de manera adecuada. Para ello, los datos serán fácilmente accesibles y localizables para los responsables e investigadores de este proyecto, y se impulsará la interoperabilidad de los datos, asegurando que se encuentren en un formato estándar y bien documentado. Con esto se pretende facilitar su comprensión y reutilización.

Respecto a la gestión de la privacidad y confidencialidad, estos datos están completamente anonimizados, sin la posibilidad de que sean asociados a personas concretas. Además, el acceso queda restringido a los responsables y a los investigadores designados y se usarán exclusivamente con fines académicos y de investigación, garantizando que no se divulguen o se utilicen con otros fines.

7. Bibliografía

1. Jaeken, J., Vanderschueren-Lodeweyckx, M., Casaer, P. *et al.* Familial psychomotor retardation with markedly fluctuating serum prolactin, FSH and GH levels, partial TBG-deficiency, increased serum arylsulphatase A and increased CSF protein: a new syndrome?: 90. *Pediatr Res* 14, 179 (1980). https://doi.org/10.1203/00006450-198002000-00117

2. Verheijen, J.; Wong, S.Y.; Rowe, J.H.; Raymond, K.; Stoddard, J.; Delmonte, O.M.; Bosticardo, M.; Dobbs, K.; Niemela, J.; Calzoni, E.; et al.,Defining a new immunedeficiencysyndrome: MAN2B2-CDG. J. AllergyClin. Immunol. 2020, 1008–1011. https://doi.org/10.1016/j.jaci.2019.11.016

3. Francisco, R., Marques-da-Silva, D., Brasil, S., Pascoal, C., Dos Reis Ferreira, V., Morava, E., &Jaeken, J. (2019). Thechallenge of CDG diagnosis. *Molecular genetics and metabolism*, *126*(1), 1–5. https://doi.org/10.1016/j.ymgme.2018.11.003v

4. Ferreira, C. R., Altassan, R., Marques-Da-Silva, D., Francisco, R., Jaeken, J., & Morava, E. (2018). Recognizablephenotypes in CDG. *Journal of inheritedmetabolicdisease*, *41*(3), 541–553. https://doi.org/10.1007/s10545-018-0156-5

5. Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., &Etzler, M. E. (Eds.). (2009). *Essentials of Glycobiology*. (2nd ed.). Cold Spring Harbor LaboratoryPress.

6. Péanne, R., de Lonlay, P., Foulquier, F., Kornak, U., Lefeber, D. J., Morava, E., Pérez, B., Seta, N., Thiel, C., Van Schaftingen, E., Matthijs, G., &Jaeken, J. (2018). Congenitaldisorders of glycosylation (CDG): Quo vadis?. *Europeanjournal of medical genetics*, *61*(11), 643–663. https://doi.org/10.1016/j.ejmg.2017.10.012

7. Chang, I.J.; He, M.; Lam, C.T. Congenitaldisorders of glycosylation. Ann. Transl. Med. 2018, 6, 1–13. doi:10.21037/atm.2018.10.45.

8. Vals, M. A., Pajusalu, S., Kals, M., Mägi, R., &Õunap, K. (2018). ThePrevalence of *PMM2-CDG* in Estonia BasedonPopulationCarrierFrequencies and DiagnosedPatients. *JIMD reports*, *39*, 13–17. https://doi.org/10.1007/8904_2017_41

9. Altassan, R., Péanne, R., Jaeken, J., Barone, R., Bidet, M., Borgel, D., Brasil, S., Cassiman, D., Cechova, A., Coman, D., Corral, J., Correia, J., de la Morena-Barrio, M. E., de Lonlay, P., Dos Reis, V., Ferreira, C. R., Fiumara, A., Francisco, R., Freeze, H., Funke, S., … Morava, E. (2019). International clinicalguidelinesforthemanagement of phosphomannomutase 2-congenital disorders of glycosylation: Diagnosis, treatment and follow up. *Journal of inheritedmetabolicdisease*, *42*(1), 5–28. https://doi.org/10.1002/jimd.12024

10. Pascoal, C., Francisco, R., Ferro, T., Dos Reis Ferreira, V., Jaeken, J., &Videira, P. A. (2020). CDG and immune response: Frombedside to bench and back. *Journal of inheritedmetabolicdisease*, *43*(1), 90–124. https://doi.org/10.1002/jimd.12126

11. Schiff, M., Roda, C., Monin, M. L., Arion, A., Barth, M., Bednarek, N., Bidet, M., Bloch, C., Boddaert, N., Borgel, D., Brassier, A., Brice, A., Bruneel, A., Buissonnière, R., Chabrol, B., Chevalier, M. C., Cormier-Daire, V., De Barace, C., De Maistre, E., De Saint-Martin, A., … De Lonlay, P. (2017). Clinical, laboratory and molecular findings and long-termfollow-up data in 96 French patientswith*PMM2-CDG* (phosphomannomutase 2-congenital disorder of glycosylation) and review of theliterature. *Journal of medical genetics*, *54*(12), 843–851. https://doi.org/10.1136/jmedgenet-2017-104903

12. Marques-da-Silva, D., Francisco, R., Webster, D., Dos Reis Ferreira, V., Jaeken, J., &Pulinilkunnil, T. (2017). Cardiaccomplications of congenitaldisorders of glycosylation (CDG): a systematicreview of theliterature. *Journal of inheritedmetabolicdisease*, *40*(5), 657–672. https://doi.org/10.1007/s10545-017-0066-y

13. Lam C, Krasnewich DM. PMM2-CDG. In: GeneReviews®. University of Washington, Seattle, Seattle (WA); 1993. PMID: 20301289.

14. Lyons, J. J., Milner, J. D., &Rosenzweig, S. D. (2015). GlycansInstructingImmunity: TheEmerging Role of AlteredGlycosylation in ClinicalImmunology. *Frontiers in pediatrics*, *3*, 54. https://doi.org/10.3389/fped.2015.00054

15. Iyer, S., Sam, F. S., DiPrimio, N., Preston, G., Verheijen, J., Murthy, K., Parton, Z., Tsang, H., Lao, J., Morava, E., &Perlstein, E. O. (2019). Repurposingthealdosereductaseinhibitor and diabeticneuropathydrugepalrestatforthecongenitaldisorder of glycosylation*PMM2-CDG*. *Diseasemodels&mechanisms*, *12*(11), dmm040584. https://doi.org/10.1242/dmm.040584

16. Martínez-Monseny, A. F., Bolasell, M., Callejón-Póo, L., Cuadras, D., Freniche, V., Itzep, D. C., Gassiot, S., Arango, P., Casas-Alba, D., de la Morena, E., Corral, J., Montero, R., Pérez-Cerdá, C., Pérez, B., Artuch, R., Jaeken, J., Serrano, M., & CDG SpanishConsortium (2019). AZATAX: Acetazolamide safety and efficacy in cerebellarsyndrome in PMM2 congenitaldisorder of glycosylation (*PMM2-CDG*). *Annals of neurology*, *85*(5), 740–751. https://doi.org/10.1002/ana.25457

17. Yuste-Checa, P., Brasil, S., Gámez, A., Underhaug, J., Desviat, L. R., Ugarte, M., Pérez-Cerdá, C., Martinez, A., & Pérez, B. (2017). PharmacologicalChaperoning: A PotentialTreatmentfor*PMM2-CDG*. *Human mutation*, *38*(2), 160–168. https://doi.org/10.1002/humu.23138

18. Brasil, S., Pascoal, C., Francisco, R., Marques-da-Silva, D., Andreotti, G., Videira, P. A., Morava, E., Jaeken, J., & Dos Reis Ferreira, V. (2018). CDG Therapies: FromBench to Bedside. *International journal of molecular sciences*, *19*(5), 1304. https://doi.org/10.3390/ijms19051304

19. Word CDG Organization. CDG Pipeline. 2021. Available online: https://worldcdg.org/drug-development/pipeline (accessedon 25 July 2023).

20. Chen, J., Li, X., Edmondson, A., Meyers, G. D., Izumi, K., Ackermann, A. M., Morava, E., Ficicioglu, C., Bennett, M. J., & He, M. (2019). IncreasedClinical Sensitivity and Specificity of Plasma Protein *N*-GlycanProfilingforDiagnosingCongenitalDisorders of Glycosylationby Use of FlowInjection-ElectrosprayIonization-Quadrupole Time-of-Flight MassSpectrometry. *Clinicalchemistry*, *65*(5), 653–663. https://doi.org/10.1373/clinchem.2018.296780

21. Bruneel, A., Cholet, S., Tran, N. T., Mai, T. D., &Fenaille, F. (2020). CDG biochemicalscreening: Where do we stand?. *Biochimica et biophysica acta. General subjects*, *1864*(10), 129652. https://doi.org/10.1016/j.bbagen.2020.129652

22. Griggs, R. C., Batshaw, M., Dunkle, M., Gopal-Srivastava, R., Kaye, E., Krischer, J., Nguyen, T., Paulus, K., Merkel, P. A., &RareDiseasesClinicalResearch Network (2009). Clinicalresearchforraredisease: opportunities, challenges, and solutions. *Molecular genetics and metabolism*, *96*(1), 20–26. https://doi.org/10.1016/j.ymgme.2008.10.003

23. Augustine, E. F., Adams, H. R., &Mink, J. W. (2013). Clinicaltrials in raredisease: challenges and opportunities. *Journal of childneurology*, *28*(9), 1142–1150. https://doi.org/10.1177/0883073813495959

24. Hemati, M. StakeholderEngagement, 1st ed.; Dahinden, M., Paschke, M., Eds.; Zurich-BaselPlantScience Center: Zürich, Switzerland, 2020.

25. Dwyer, A. A., Quinton, R., Morin, D., &Pitteloud, N. (2014). Identifyingtheunmethealthneeds of patientswithcongenitalhypogonadotropichypogonadismusing a web-basedneedsassessment: implicationsfor online interventions and peer-to-peer support. *Orphanetjournal of rarediseases*, *9*, 83. https://doi.org/10.1186/1750-1172-9-83

26. Goodspeed, K., Bliss, G., &Linnehan, D. (2020). Bringingeveryone to thetable - findingsfromthe 2018 Phelan-McDermidSyndromeFoundation International Conference. *Orphanetjournal of rarediseases*, *15*(1), 152. https://doi.org/10.1186/s13023-020-01389-6

27. de Freitas, C., Amorim, M., Machado, H., Leão Teles, E., Baptista, M. J., Renedo, A., Provoost, V., & Silva, S. (2021). Public and patientinvolvement in health data governance (DATAGov): protocol of a people-centred, mixed-methodsstudyon data use and sharingforrarediseasescare and research. *BMJ open*, *11*(3), e044289. https://doi.org/10.1136/bmjopen-2020-044289

28. Matthijs, G., Schollen, E., Heykants, L., &Grünewald, S. (1999). Phosphomannomutaseeficiency: the molecular basis of theclassicalJaekensyndrome (CDGS typeIa). Molecular genetics and metabolism, 68(2), 220–226. https://doi.org/10.1006/mgme.1999.2914

29. VibekeWestphal, SusanneKjaergaard, ElsSchollen, Kevin Martens, Stephanie Grunewald, Marianne Schwartz, GertMatthijs, Hudson H. Freeze, A frequentmildmutation in ALG6 mayexacerbatetheclinicalseverity of patientswithcongenitaldisorder of glycosylationIa (CDG-Ia) causedbyphosphomannomutasedeficiency, Human Molecular Genetics, Volume 11, Issue 5, 1 March 2002, Pages 599–604, https://doi.org/10.1093/hmg/11.5.599

30. BarbaraBortot, Dora Cosentini, Flavio Faletra, StefaniaBiffi, Eleonora De Martino, Marco Carrozzi, Giovanni MariaSeverini, PMM2-CDG: Phenotype and genotype in fouraffectedfamilymembers, Gene, Volume 531, Issue 2, 2013, Pages 506-509, ISSN 0378-1119, https://doi.org/10.1016/j.gene.2013.07.083.

31. Manuel A. Rivas *et al.*Effect of predictedprotein-truncatinggeneticvariantsonthe human transcriptome.Science348,666-669(2015).DOI:10.1126/science.1261877

32. Richards, S. et al. Standards and guidelinesfortheinterpretation of sequencevariants: a jointconsensusrecommendation of the American College of Medical Genetics and GenomicsandtheAssociationfor Molecular Pathology. Genet. Med. 17, 405–424 (2015). DOI:https://doi.org/10.1038/gim.2015.30

33. Scotti, M., Swanson, M. RNA mis-splicing in disease. NatRevGenet 17, 19–32 (2016). https://doi.org/10.1038/nrg.2015.3

34. Cummings, B. B. et al. Improvinggenetic diagnosis in Mendeliandiseasewithtranscriptomesequencing. Sci. Transl. Med. 9, https://doi.org/10.1126/scitranslmed.aal5209 (2017).

35. Frésard, L., Smail, C., Ferraro, N.M. et al. Identification of rare-disease genes usingbloodtranscriptomesequencing and large control cohorts. NatMed 25, 911–919 (2019). https://doi.org/10.1038/s41591-019-0457-8

36. Gonorazky, H. D. et al. Expandingtheboundaries of RNA sequencing as a diagnostictoolforrareMendeliandisease. Am. J. Hum. Genet. 104, 466–483 (2019).https://doi.org/10.1016/j.ajhg.2019.01.012

37. Lee, H. et al. Diagnosticutility of transcriptomesequencingforrareMendeliandiseases. Genet. Med. https://doi.org/10.1038/s41436-019-0672-1 (2019).

38. Karam, R., Conner, B., LaDuca, H., McGoldrick, K., Krempely, K., Richardson, M. E., Zimmermann, H., Gutierrez, S., Reineke, P., Hoang, L., Allen, K., Yussuf, A., Farber-Katz, S., Rana, H. Q., Culver, S., Lee, J., Nashed, S., Toppmeyer, D., Collins, D., Haynes, G., … Chao, E. (2019). Assessment of DiagnosticOutcomes of RNA GeneticTestingforHereditaryCancer. JAMA network open, 2(10), e1913900. https://doi.org/10.1001/jamanetworkopen.2019.13900

39. Landrith, T., Li, B., Cass, A.A. et al. Splicingprofileby capture RNA-seqidentifiespathogenicgermlinevariants in tumor suppressor genes. npjPrecis. Onc. 4, 4 (2020). https://doi.org/10.1038/s41698-020-0109-y

40. Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T. B., Graf, E., Schwarzmayr, T., Terrile, C., Koňaříková, E., Repp, B., Kastenmüller, G., Adamski, J., Lichtner, P., Leonhardt, C., Funalot, B., Donati, A., Tiranti, V., Lombes, A., … Prokisch, H. (2017). Genetic diagnosis of Mendeliandisordersvia RNA sequencing. Naturecommunications, 8, 15824. https://doi.org/10.1038/ncomms15824

41. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data ProcessingSubgroup (2009). TheSequenceAlignment/Mapformat and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

42. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project AnalysisGroup (2011). Thevariantcallformat and VCFtools. Bioinformatics (Oxford, England), 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

43. Alexander Dobin, Carrie A. Davis, FelixSchlesinger, JorgDrenkow, Chris Zaleski, SonaliJha, PhilippeBatut, Mark Chaisson, Thomas R. Gingeras, STAR: ultrafast universal RNA-seqaligner, *Bioinformatics*, Volume 29, Issue 1, January 2013, Pages 15–21, https://doi.org/10.1093/bioinformatics/bts635

44. En este TFM, se ha empleado inteligencia artificial, específicamente ChatGPT, únicamente como herramienta de búsqueda de información. OpenAI. (2021). ChatGPT. https://openai.com

8. ANEXOS

En el siguiente enlace pueden encontrarse los archivos anexos de este trabajo:

https://drive.google.com/drive/folders/1se7e6d898rz0KPB5b-d2ZJr0nkXm-9ri?usp=share_link