UNIVERSIDAD DE MURCIA

TRABAJO FIN DE MASTER EN BIOINFORMÁTICA

Realizado en el departamento de Bioinformática Clínica y Traslacional del Vall d'Hebron Instituto de Investigación, en colaboración con el departamento de inmunología del Hospital Vall d'Hebron

UNIVERSIDAD DE MURCIA

CARACTERIZACIÓN DE LA VARIABILIDAD GENÉTICA DESCRITA EN LOS PANELES DE ESTUDIO DE LA INMUNODEFICIENCIA COMÚN VARIABLE

AUTORA: LORENA MIRETE GARCÍA DNI: 48746652S

TUTORES VHIR: XAVIER DE LA CRUZ & NATÀLIA PADILLA
TUTOR UMU: JUAN A. BOTÍAS

LÍNEA DE TFM: Caracterización de las variantes genéticas de la secuencia del genoma en fenotipos patológicos específicos mediante el uso de métodos computacionales

MURCIA, SEPTIEMBRE 2021

RESUMEN

A pesar de su elevada prevalencia en continentes como Europa, la Inmunodeficiencia Común Variable (CVID) es un desorden del sistema inmunitario cuyo origen, generalmente poligénico, es desconocido en la mayoría de los casos. Esta situación está cambiando debido a los avances en las tecnologías de secuenciación masiva, que han dado lugar al uso extendido de los paneles NGS (Next-Generation Secuencing) en el entorno clínico. Estos paneles permiten secuenciar un gran número de genes de interés de una forma rápida y menos costosa que la secuenciación del genoma completo (WGS). Así, permiten estudiar en profundidad las enfermedades con un componente genético, caracterizando el número y tipo de variantes genéticas que poseen quienes las padecen. Este conocimiento supone un gran avance en la medicina personalizada de CVID y otras enfermedades poligénicas, pues el uso de estos datos para entrenar modelos de aprendizaje automático podría proporcionar un diagnóstico rápido y concreto, ayudando a los profesionales de la medicina a encontrar la diana para tratar la enfermedad.

En este trabajo se ha caracterizado el perfil genético de una cohorte de 131 pacientes CVID españoles a través del estudio del panel específico de 313 genes, diseñado en el Hospital Vall d'Hebron, y se han comparado con los datos de una cohorte de 785 controles españoles. Se ha desarrollado un patrón de flujo de trabajo generalizado, reproducible y robusto para estudiar futuros datos obtenidos de la secuenciación de paneles para otras enfermedades. Se han caracterizado los genes que presentan mayor número de variantes en CVID y se presenta el posible diagnóstico de un caso individual. Respecto a la diferencia general entre pacientes y controles, se percibe una disgregación pronunciada entre ambas poblaciones, cuya causa no se puede atribuir a la enfermedad. Nuestros resultados sugieren que su origen es probablemente técnico, achacables a los diferentes métodos de secuenciación usados en pacientes y controles. Ello ilustra la complejidad del problema planteado y la necesidad de aumentar la coherencia en las cohortes de pacientes utilizadas para poder realizar una caracterización más profunda.

PALABRAS CLAVE

CVID; panel génico; bioinformática; caracterización de variantes genéticas; enfermedades poligénicas;

ABSTRACT

Despite its high prevalence in continents such as Europe, Common Variable Immunodeficiency (CVID) is a disorder of the immune system whose origin, generally polygenic, is unknown in most cases. This situation is changing due to advances in mass sequencing technologies, which have led to the widespread use of NGS (Next-Generation Sequencing) panels in the clinical setting. These panels allow for a large number of genes of interest to be sequenced quickly and less expensively than Whole Genome Sequencing (WGS). Thus, they allow an in-depth study of diseases with a genetic component, characterizing the number and type of genetic variants possessed by those who suffer from them. This knowledge represents a great advance in the personalized medicine of CVID and other polygenic diseases, since the use of this data to train machine learning models could provide a rapid and concrete diagnosis, helping medical professionals to find the target to treat the illness.

In this work, the genetic profile of a cohort of 131 Spanish CVID patients has been characterized through the study of the specific panel and they have been compared with the data of a cohort of 785 Spanish controls. A robust, reproducible and generalized workflow pattern has been developed to study future data obtained from panel sequencing for other diseases. The genes that present the greatest number of CVID variants have been described, and the possible diagnosis of an individual case is presented. Regarding the general difference between patients and controls, a pronounced disintegration is observed between both populations, the cause of which cannot be attributed to the disease. Our results suggest that its origin is probably technical, attributable to the different sequencing methods used in patients and controls. This illustrates the complexity of the problem posed and the need to increase consistency in the patient cohorts used in order to carry out a more in-depth characterization

KEY WORDS

CVID; gene panel; bioinformatics; characterization of genetic variants; polygenic diseases;

ÍNDICE

1 INTRODUCCION 5
1.1 Inmunodeficiencia común variable 5
1.2 Paneles génicos NGS 6
1.3 Herramientas de predicción de patogenicidad 'in silico'. 6
1.3.1 SIFT "Sorting Intolerant From Tolerant" 6
1.3.2 POLYPHEN -2 "Polymorphism Phenotyping v2" 6
1.3.3 REVEL "Rare Exome Variant Ensemble Learner" 7
1.4 Técnicas de análisis multidimensional 7
1.4.1 PCA "Análisis de componentes principales" 8
1.4.2 TSNE "t-distributed stochastic neighbor embedding" 8
1.4.3 UMAP "Uniform Manifold Approximation and Projection" 8
2 OBJETIVOS 8
2.1 Objetivos de investigación 9
2.1.1 Conocer los estudios preliminares. 9
2.1.2 Caracterizar los genes de diagnóstico monogénico y del panel de estudio 9
2.1.3 Comprobar la existencia de diferencias entre el número y tipo de las variantes en pacientes CVID de causa genética conocida y desconocida. 9
2.1.4 Descartar la Influencia del sexo y la edad en el número y tipo de variantes 9
2.1.5 Ratificar la distinción entre pacientes y controles estudiando los genes del panel 10
2.2 Objetivos técnicos 10
2.2.1 Anotar y seleccionar las variantes de los individuos de la cohorte 10
2.2.2 Representar gráficamente y analizar la correlación entre los grupos de estudio 10
3 RESULTADOS PRELIMINARES 10
4 MATERIALES Y MÉTODOS 11
4.1 Cohortes 11
4.1.1 Cohorte de pacientes 11
4.1.2 Cohorte de controles 11
4.2 Caracterización de los genes de estudio 12
4.2.1 Genes del panel 12
4.2.2 Genes de diagnóstico 12
4.3 FLUJO DE TRABAJO 13
4.3.1 Preprocesado y preparación de los datos 13
4.3.2 Anotación de variantes 14
4.3.3 Procesado y selección de variantes 14
4.3.4 Selección y creación de los grupos de análisis 15
4.3.5 Reducción de la dimensionalidad y representación gráfica 16
5 RESULTADOS 17
5.1 Caracterización inicial de los paneles de pacientes CVID 17
5.2 Estudio comparativo de los paneles de pacientes CVID vs. controles 18
5.3 Análisis de las variantes potencialmente patogénicas 21

6 DISCUSIÓN 21
7 CONCLUSIONES 22
8 BIBLIOGRAFÍA 22

ÍNDICE ABREVIATURAS

ABREVIATURAS	SIGNIFICADO
CID	Inmunodeficiencia Combinada
CVID	Inmunodeficiencia Común Variable
DNA	Ácido Desoxirribonucleico
DSB	Rotura de doble cadena de DNA
ESV	Variantes de secuencia exacta
HVDH	Hospital Vall D'Hebron
Ig	Inmunoglobulina
NGS	Next-Generation Secuencing
OOB	Out of Bounds
PCA	Análisis de Componentes Principales
PCR	Reacción en cadena de la polimerasa
PID	Inmunodeficiencia Primara
SCID	Inmunodeficiencia Combinada Servera
SNP	Polimorfismos de un solo nucleótido
UMU	Universidad de Murcia
VCF	Variant call format
VEB	Virus Epstein-Barr
VEP	Variant ensembling predictor
VHIR	Instituto de investifgación Vall d'Hebron
VODI	Enfermedad venooclusiva con inmunodeficiencia
WGS	Whole Genome Sequencing
XLP	Síndrome proliferativo ligado al cromosoma X

ÍNDICE FIGURAS

FIGURA 1. Diagrama de barras de la prevalencia de CVID en distintas zonas geográficas.. 5
FIGURA 2. Interpretación de los scores de Revel. 7
FIGURA 3. Representación PC1 vs PC2 del análisis de las variantes missense 'Cáncer mama vs control'. 11
FIGURA 4. Ejemplo del significado de los distintos tipos de SNPs. 15
FIGURA 5. Representación PC1 vs PC2 de variables HH en genes importantes 'CVID vs control'. 17
FIGURA 6. Representación de UMAP 'CVID vs control' diferenciados por diagnóstico, edad y sexo. 18
FIGURA 7. Representación PC1 vs PC2 del grupo de mayor volumen de datos 'CVID vs control', con representación de varianza e identificación de genes originales con más aportación tienen sobre PC1. 19
FIGURA 8. Representación PC1 vs PC2 del grupo 'Missense-H-importantes' 'CVID vs control', con representación de varianza e identificación de genes originales con más aportación tienen sobre PC1 y PC2. 19
FIGURA 9 Representación UMAP 'CVID vs control' del grupo 'Missense-H-importates' 19
FIGURA 10 Gráficas de dispersión medianas/medias de variantes 'CVID vs control'.. 20
FIGURA 11. Representación tSNE 'CVID vs control' variantes REVEL patogénicas. 21

REPOSITORIO

- Github: https://github.com/lorenamirete/TFM.git

1 INTRODUCCION

Para la elaboración de este proyecto, ha sido necesaria una documentación bibliográfica previa para comprender en profundidad las peculiaridades de la enfermedad y de las herramientas empleadas para alcanzar los objetivos de la investigación. En este apartado se desarrollan los puntos críticos para entender el contexto del trabajo, que incluyen (i) las características de la Inmunodeficiencia Común Variable, (ii) los Paneles-NGS, herramienta de diagnóstico clínico empleada para obtener los datos de secuenciación, (iii) las Herramientas de Predicción Patológica 'in silico' necesarias, para interpretar estos datos, y (iv) las Técnicas de Análisis Multidimensional que han sido imprescindibles el análisis de dichos datos.

1.1 Inmunodeficiencia común variable

La inmunodeficiencia común variable (CVID) es la más común y clínicamente relevante de las inmunodeficiencias primarias (PID) diagnosticadas en la actualidad. Esta enfermedad se reparte de forma muy desigual según su etiología, siendo Europa el segundo continente con mayor prevalencia (Yazdani, R. et al, 2019), donde representa ~20% de los casos de PID. El 25% de los pacientes europeos de CVID son diagnosticados de forma temprana durante la infancia (Bogaert, D.J.A. et al, 2016; Resnick, E.S. et al, 2012).

El fenotipo de la CVID varía entre pacientes; algunos presentan defectos genéticos que afectan directamente al proceso de maduración de los linfocitos B y a su diferenciación a linfocitos B de memoria (Ameratunga, R. et al, 2018). Esta y otras características menos conocidas dan lugar a una dramática reducción en los niveles de inmunoglobulina G (IgG) en suero, aunque también de las inmunoglobulinas A (IgA) o las inmunoglobulinas M (IgM). La descripción más citada de la enfermedad determina que los pacientes necesitan dos de los siguientes criterios para ser asignados como "probable CVID": Ser mayor de 2 años de edad, poseer un nivel de IgG e IgA menor de 2 desviaciones estándar de la media para la edad, no tener respuesta al tratamiento con isohemaglutinina o respuesta frente a vacunas y/o que no exista ninguna otra causa definida para la hipogammaglobulinemia (Grimbacher, B. 2014). Estos niveles anormalmente bajos de inmunoglobulinas se traducen en una respuesta alterada del proceso inmunológico, desarrollando infecciones crónicas y recurrentes del tracto respiratorio e intestinal, enfermedades autoinmunes, enfermedades granulomatosas y/o complicaciones linfoproliterativas, entre otras (Leonardi, L. et al, 2019).

A pesar de su impacto médico-social, las bases genéticas de CVID son casi desconocidas. En la mayoría de los casos la enfermedad parece surgir de manera esporádica, aunque existe entre un 5 y 25% de pacientes con un *background* de antecedentes familiares. El desencadenante del fenotipo es de origen monogénico solo en el 2-10% de pacientes, lo habitual es que su origen sea más complejo que una enfermedad mendeliana y que los genes modificadores desempeñen un papel crucial en el desarrollo de la enfermedad (Bonilla, F.A. et al, 2016). Por lo tanto, para la comprensión de esta enfermedad, no podemos utilizar las técnicas convencionales de secuenciación dirigida en las que se identifican las variantes genéticas de un gen o un pequeño número de ellos. Necesitamos herramientas más poderosas, que nos permitan investigar decenas de genes candidatos simultáneamente. Los métodos de secuenciación de alto rendimiento, desarrollados en los últimos años, y concretamente los paneles génicos de secuenciación, constituyen una técnica ideal para estudiar la base genética de CVID.

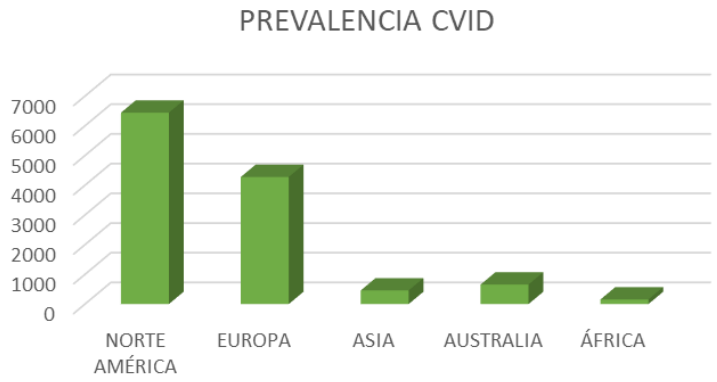


FIGURA 1. DIAGRAMA DE BARRAS DE LA PREVALENCIA DE CVID EN DISTINTAS ZONAS GEOGRÁFICAS. NORTE AMÉRICA (N=6443), EUROPA (N=4279), ASIA (N=459), AUSTRALIA (N=657) Y ÁFRICA (N=156) (YAZDANI, R. ET AL, 2019). LOS DATOS ESTÁN CONDICIONADOS POR EL AVANCE DE LAS TÉCNICAS DE DETECCIÓN EN CADA REGIÓN.

1.2 Paneles génicos NGS

Se denomina *Next-Generation Sequencing* (NGS) a las técnicas de secuenciación masiva del genoma (Gupta, A.K. & Gupta, U.D. 2013). Se desarrolló a raíz del "Proyecto Genoma Humano" en 2003 y es la tecnología de secuenciación que ha revolucionado la genómica. Ha conseguido disminuir el tiempo necesario para secuenciar un genoma humano completo desde casi una década a un solo día. Este rendimiento contrasta con el de técnicas anteriores, Sanger y PCR (Reacción en cadena de la polimerasa) que permiten identificar variantes genéticas de forma más limitada en cuanto a número de genes.

Los paneles génicos, en cuyo análisis hemos centrado este trabajo, son una versión reducida del NGS, una herramienta de diagnóstico clínico basada en el análisis genético simultáneo de un grupo variable de genes asociados a una determinada patología (Morganti, S. et al, 2018). Además de permitirnos detectar mutaciones en muchos genes a la vez, disminuye el tiempo de respuesta y el coste económico por paciente. Son la herramienta de diagnóstico más exitosa para las enfermedades clínicamente heterogéneas, como es el caso de las PIDs (Bisgin, A. et al, 2018), aunque es necesario que los datos sean analizados por centros especializados en detección de variantes relevantes y en variantes raras (con una frecuencia baja en la población) (Bisgin, A. et al, 2021). Si comparamos el panel génico frente al estudio del genoma completo, podemos concluir que ambos presentan ventajas y su uso dependerá de la situación que queramos analizar. En el panel se estudian genes específicos de los que sabemos que ciertas mutaciones pueden causar una u otra enfermedad. Por otro lado, el estudio del exoma (una variante del NGS en la que se secuencia la región codificante del genoma) nos permite detectar mutaciones en cualquier gen, independientemente de si está relacionado o no con la enfermedad. Por eso, este último es interesante cuando se desconoce el gen que causa la enfermedad, cuando los genes conocidos han sido estudiados sin detección de variantes raras y podríamos enfrentarnos a una presentación clínica nueva. Aun así, es una práctica que no suele llevarse a cabo, ya que es una técnica de difícil interpretación (Saudi Mendeliome Group, 2015). En el caso de CVID, los estudios fundamentados en NGS han demostrado ser muy útiles para discernir las bases genéticas de las inmunopatologías poligénicas y desarrollar pruebas genéticas mejoradas (Ameratunga, R. et al. 2018). En este trabajo nos centraremos en el estudio de paneles génicos para esta enfermedad, generosamente proporcionados por el grupo de Inmunología del Hospital Universitario Vall d'Hebron (HVDH), y cuyos resultados de secuenciación se han interpretado con ayuda de herramientas de predicción de patogenidad.

1.3 Herramientas de predicción de patogenicidad 'in silico'

La interpretación de las variantes genéticas obtenidas mediante paneles NGS es crucial para el avance en medicina personalizada. La gran mayoría de variantes descubiertas por NGS son raras, es decir, están presentes en menos del 1% de la población (Abecasis, G.R. et al, 2012). La clasificación de la patogenicidad de las variantes genéticas está basada en múltiples criterios. De ellos, podemos destacar los datos obtenidos de la población, la posición en la que se encuentra la alteración, los casos previamente bien estudiados, y la puntuación mediante predictores 'in silico'. (Richards, S. et al, 2015). En este trabajo nos hemos centrado en esta última aproximación. A continuación, se comentan las tres técnicas de predicción 'in silico' empleadas.

1.3.1 SIFT "Sorting Intolerant From Tolerant"

SIFT es una herramienta que usa la homología en la secuencia para predecir cuando la sustitución de un aminoácido puede afectar a la función de la proteína y, en consecuencia, alterar potencialmente el fenotipo. En otras palabras, su sistema de clasificación se basa en la probabilidad de que dicha sustitución sea evolutivamente viable. Su criterio de clasificación solo distingue entre variantes perjudiciales y tolerables (Ng, P.C. & Henikoff, S. 2001). La clasificación con esta técnica de predicción depende en gran medida de la diversidad de secuencias utilizadas en el alineamiento. Cuando las secuencias utilizadas están estrechamente relacionadas, muchas posiciones parecen conservadas e importantes para que los transcritos sean funcionales. En estos casos, la precisión de la predicción de las sustituciones perjudiciales es alta, pero por contra se predice que muchas sustituciones funcionalmente neutras son perjudiciales.

1.3.2 POLYPHEN -2 "Polymorphism Phenotyping v2"

PolyPhen-2 permite predecir el posible impacto de la sustitución de un aminoácido en la estabilidad y la función de la proteína basándose en un árbol de decisión. Este se fundamenta en características estructurales y comparativas evolutivas según la similitud de la secuencia. Primero, realiza anotaciones

funcionales de polimorfismos de SNPs y asigna los SNPs codificadores a los genes que transcriben, después extrae anotaciones de secuencias de proteínas y atributos estructurales, y en última instancia crea perfiles de conservación. Tras este proceso estima la probabilidad de que la mutación sin sentido sea dañina basándose en una combinación de todas estas propiedades (Adzhubei, I. et al, 2013). Las consideraciones aplicadas para la clasificación tienen el mismo origen que las descritas en SIFT, aunque como diferencia, PolyPhen posee 4 niveles de clasificación. La sensibilidad que presentan ambas es razonablemente alta (cercana al 70%) pero la especificidad, aunque ligeramente más elevada que es SIFT, sigue siendo el punto flojo de PolyPhen que a penas logra alcanzar un 20%. Por este motivo, aunque SIFT y PolyPhen pueden ser útiles para priorizar los cambios que probablemente causen una pérdida de la función de las proteínas, su baja especificidad representa que sus predicciones deben interpretarse con precaución y deben buscarse más pruebas para respaldar o refutar la patogenicidad antes de informar sobre una nueva mutación no sinónima (Flanagan, S.E. et al, 2010).

1.3.3 REVEL "Rare Exome Variant Ensemble Learner"

REVEL es el método más reciente y complejo de los mencionados. Se trata de un metapredictor o predictor de segunda generación que integra puntuaciones de predictores 'in silico' previos como MutPred, FATHMM, VEST, PolyPhen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, y phastCons. REVEL ha sido entrenado con variantes patogénicas y variantes raras neutras de cambio de sentido descubiertas recientemente, excluyendo aquellas que se habían usado para entrenar las herramientas que lo constituyen. Las puntuaciones de REVEL scores son calculadas por el dbNSFP Project (Liu, X. et al, 2020), una base de datos desarrollada para la predicción funcional y la anotación de todas las posibles variantes de un solo nucleótido no sinónimas (nsSNVs) en el genoma humano. En el uso generalizado de REVEL, se consideran patogénicas las variantes con puntaciones por encima de 0.5, las cuales presentan una especificidad cercana al 90% y una sensibilidad del 75% (Ioannidis, N.M. et al, 2016). Esto lo convierte en un predictor más fiable y avanzado que los de primera generación.

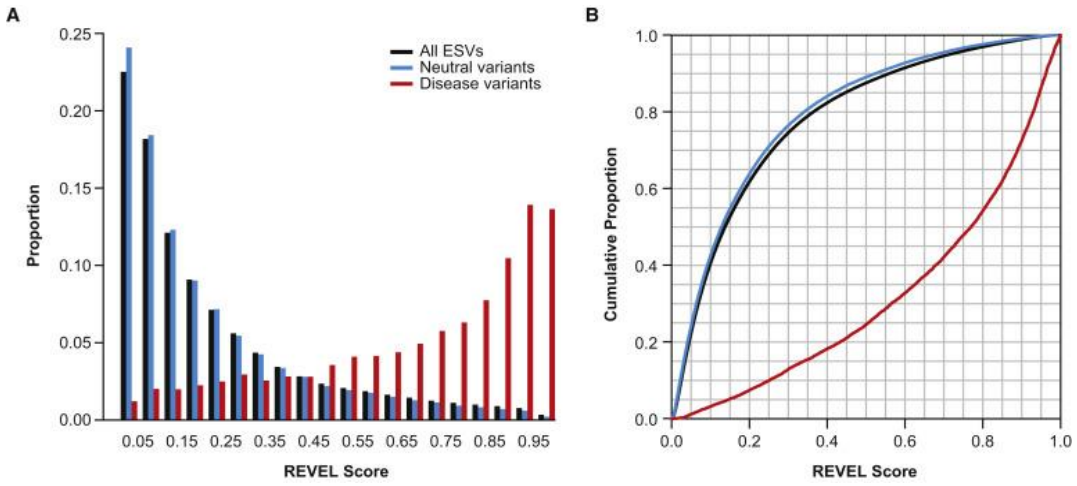


FIGURA 2. INTERPRETACIÓN DE LOS SCORES DE REVEL. (A) DISTRIBUCIÓN DE LAS PUNTUACIONES PARA VARIANTES PATOLÓGICAS (ROJO) Y NEUTRAS (AZUL) DE ENTRENAMIENTO, Y ESV (VARIANTES DE SECUENCIA EXACTA) (NEGRO). LAS PREDICCIONES DE REVEL SON COMPUTADAS SOLO CON LOS OOB (FUERA DE LÍMITES). (B) PERCENTILES DE LA DISTRIBUCIÓN DE LAS PUNTUACIONES PARA LAS VARIANTES DE ENTRENAMIENTO PATOGÉNICAS (ROJO) Y LAS NEUTRAS (AZUL) Y ESVS (NEGRO). (IOANNIDIS, N.M. ET AL, 2016).

1.4 Técnicas de análisis multidimensional

Finalizando esta introducción, se describen brevemente las técnicas de análisis multidimensional que han sido ampliamente utilizadas en esta investigación. Su uso se justifica por el hecho de que, una vez interpretado el significado biológico a través de las herramientas de predicción, para poder estudiar gráficamente todas las variables genéticas a la vez, es necesaria una representación especial para obtener gráficas bi-dimensionales que permitan su interpretación por el ojo humano. Entre las múltiples técnicas para lograr este objetivo, en este trabajo hemos utilizado 3 siguientes:

1.4.1 PCA “Análisis de componentes principales”

El análisis de componentes principales (PCA) no se trata de un análisis gráfico, sino de una técnica de pre-procesado de reducción de la dimensionalidad para eliminar características redundantes y/o ruido (Velliangiri, S. et al, 2019). La reducción se consigue a través de la creación de nuevas combinaciones lineales de las variables que caracterizan los objetos de estudio (Maćkiewicz, A. & Ratajczak, W. 1993). Estas nuevas combinaciones son las denominadas “componentes principales” y tienen que cumplir ciertos criterios matemáticos y estadísticos, maximizando la varianza y preservando las distancias pares grandes.

1.4.2 TSNE “t-distributed stochastic neighbor embedding”

t-SNE es una moderna técnica de visualización y reducción de dimensiones no lineal y no supervisada para datos multidimensionales. Se trata de una variación de la incorporación estocástica de vecinos (Hinton, G. & Roweis, S.T., 2002) mucho más fácil de optimizar y que produce mejores visualizaciones gracias a la reducción de la tendencia a hacinar los puntos en el centro del mapa. Es una técnica muy útil para revelar estructuras a diferente escala, pues mantiene las distancias locales (Van der Maaten, L. & Hinton G., 2008).

Su aplicación requiere definir varios hiperparámetros entre los que destaca “perplexity”. La disminución de su valor está relacionada con el aumento del conocimiento local, mientras que valores altos se asocian con una mejor representación global. El rango recomendado por sus autores está entre 5 y 50, sugiriendo valores próximos al número de individuos por clúster. De lo contrario, las agrupaciones a escalas menores podrían simular una subdivisión. Por eso la aplicación de t-SNE es útil cuando conocemos de antemano los grupos que estamos buscando (Wattenberg, M. et al, 2016).

Respecto a su aplicabilidad en datos genéticos humanos, se ha comprobado que el resultado no solo es similar a las técnicas de reducción de dimensiones, como PCA, cuando se escogen los hiperparámetros de forma adecuada; sino que además t-SNE es más robusto con respecto a la presencia de valores atípicos y puede mostrar patrones continentales y subcontinentales en un solo gráfico (Li, W. et al, 2017). A pesar de esto, su uso no está muy extendido en el estudio de datos genéticos, puesto que presenta ciertas desventajas frente a PCA: la proyección de nuevos datos en el estudio ya realizado es muy complicada, es una técnica no-determinista (cada ejecución de t-SNE puede dar como resultado una solución diferente) y pierde más información global en aras a preservar la información local.

1.4.3 UMAP “Uniform Manifold Approximation and Projection”

UMAP (Aproximación y Proyección Múltiple Uniforme) es una técnica reciente de aprendizaje múltiple no lineal para la reducción de dimensiones. UMAP se construye a partir de un marco teórico basado en la geometría de Riemann y la topología algebraica (Allaoui, M. et al, 2020). El resultado es un algoritmo escalable práctico que se adapta a los datos del mundo real. El algoritmo UMAP es comparable a t-SNE en cuanto a calidad de visualización y presenta una ejecución más rápida que este último. Además, UMAP no tiene restricciones computacionales sobre la incorporación de la dimensión, lo que la hace viable no solo para visualización sino también como una técnica de reducción de dimensiones de propósito general para alimentar modelos de aprendizaje automático (McInnes, L..et al, 2020).

Aunque durante los primeros estudios realizados con esta técnica se avistaba la posibilidad de que UMAP conservara una mayor parte de la estructura global que t-SNE, lo cierto es que esta característica surge a raíz de la forma en la que ambos inicializan sus hiperparámetros. t-SNE usa por defecto unos hiperparámetros random de inicialización, mientras que UMAP tiene implementada una técnica denominada “Laplacian eigenmaps” para inicializar las incorporaciones (Kobak, D. & Liderman, G.C., 2019) de una manera más optimizada y adaptada a los propios datos. Si inicializamos t-SNE con los hiperparámetros adecuados para los datos a tratar, la conservación de la estructura global es similar a la obtenida con UMAP.

2 OBJETIVOS

El **Objetivo global** de este trabajo es confirmar la hipótesis de que se puede caracterizar el perfil genético característico de los pacientes CVID mediante el uso de paneles génicos y por ende se puede diferenciar pacientes de controles a partir de los genes estudiados en el panel. El trabajo asociado a este objetivo se desarrollará mediante una combinación de técnicas bioinformáticas y de análisis multidimensional de datos.

Este objetivo global se descompone en dos bloques de objetivos específicos: objetivos de investigación y objetivos técnicos. Los primeros corresponden a las preguntas científicas principales a las que esperamos responder en este proyecto. Los segundos buscan garantizar la utilidad del trabajo de programación realizado, es decir la transferibilidad de las herramientas bioinformáticas creadas. Las siguientes secciones desarrollan con más detalle estos objetivos.

2.1 Objetivos de investigación

El contexto general de este proyecto, elaborado conjuntamente por los grupos de Bioinformática Clínica y Traslacional y el de Inmunogenética del Hospital Vall d'Hebron de Barcelona, es el de avanzar en el tratamiento y diagnóstico de los pacientes con CVID. Actualmente, las terapias de reemplazo y los tratamientos de apoyo inmunológico son las indicaciones más comunes para estos pacientes. Sin embargo, esta aproximación no es eficaz para todos los individuos, ya que la enfermedad se presenta de manera diferente entre los pacientes debido a la heterogeneidad genética subyacente. Los datos acumulados mediante el uso de tecnologías NGS y pruebas genéticas mejoradas pueden utilizarse para ayudar a determinar la opción de terapia de reemplazo de inmunoglobulina más efectiva durante el proceso de toma de decisiones clínicas (Bisgin, A. et al, 2021).

En este contexto, el objetivo del grupo de Bioinformática Clínica y Traslacional es el de trabajar en el desarrollo de herramientas bioinformáticas para caracterizar a los pacientes CVID caucásicos españoles con desencadenante poligénico y así facilitar y agilizar su diagnóstico. Para validar este trabajo computacional, se utilizará una cohorte de pacientes para los que los datos de secuenciación se han obtenido mediante un panel génico de 313 genes diseñado en el departamento de Inmunología del HVDH.

Este estudio se distingue de las publicaciones previas sobre CVID por el elevado número de individuos que componen la cohorte y por la etnia de los mismos, ya que no hay registros previos de análisis centrados en pacientes caucásicos españoles. Estas dos características son importantes debido a la naturaleza de la enfermedad. En primer lugar, la prevalencia de la CVID varía significativamente entre países y poblaciones, por lo que es esencial descartar la variabilidad por etnicidad. En segundo lugar, la poligenicidad, prevalencia distintiva y heterogeneidad de la enfermedad hace que sea fundamental extender el estudio al mayor número de individuos posible (Bisgin, A. et al, 2021). A continuación, se enumeran y justifican los objetivos parciales asociados al objetivo de investigación.

2.1.1 Conocer los estudios preliminares

El estudio del desarrollo de las investigaciones previas permite entender el contexto biológico del que se parte y reconocer la procedencia de algunos patrones que pueden repetirse en el análisis.

2.1.2 Caracterizar los genes de diagnóstico monogénico y del panel de estudio

Reconocer las características de los genes previamente caracterizados a los que se les atribuye el desarrollo de la CVID nos permite entender el funcionamiento de la enfermedad y ayuda a buscar aquellos con motivos o funciones similares cuyos defectos o suma de ellos pudieran desencadenar el fenotipo.

2.1.3 Comprobar la existencia de diferencias entre el número y tipo de las variantes en pacientes CVID de causa genética conocida y desconocida

Se busca conocer qué genes que no se consideran para el diagnóstico monogénico tienen un número de variantes significativamente superior al resto en los pacientes CVID con diagnóstico genético no concluyente o desconocido, frente a aquellos de causa genética conocida.

2.1.4 Descartar la Influencia del sexo y la edad en el número y tipo de variantes

De la gran mayoría de individuos de la cohorte se conocen las características de edad y sexo. Es importante descartar que ninguna de las dos está enmascarando los datos de nuestro análisis final.

2.1.5 Ratificar la distinción entre pacientes y controles estudiando los genes del panel

La existencia de una tendencia desigual en el número de variantes de algunos genes en pacientes frente a los controles nos permite la caracterización del perfil enfermedad CVID y su distinción ante sanos.

2.2 Objetivos técnicos

Tal como se ha mencionado, el objetivo global del proyecto se abordará mediante el desarrollo y uso de técnicas bioinformáticas, alimentadas de datos procesados mediante programas codificados en Python o R, en nuestro caso. Estos dos objetos, tanto los datos como el código, tienen un valor objetivo para el grupo de acogida, que los puede utilizar para seguir desarrollando el proyecto tras mi marcha del grupo. Por ello, el objetivo técnico de este trabajo es el de construir y ejecutar un flujo de trabajo generalizado, versátil, robusto y reproducible para realizar el análisis del perfil genético en pacientes CVID. Estas herramientas tienen que proporcionar la información necesaria para resolver el problema global, además de permitir su uso posterior en el hospital y su extensión, de forma sencilla, a otros paneles génicos y enfermedades. A continuación, se enumeran y justifican los objetivos parciales asociados a los objetivos técnicos.

2.2.1 Anotar y seleccionar las variantes de los individuos de la cohorte

Será necesario llevar a cabo el preprocesado de los datos de secuencia iniciales para unificar el modelo de entrada de los programas incluidos en el trabajo. Se creará un programa que automatice la anotación de las variantes comparándolo con el genoma de referencia y se llevará a cabo el procesado y la selección de los datos para usar aquellos que aporten valor y descartar el ruido. En última instancia, se crearán los archivos que presentan la información del número de variantes por gen e individuo agrupados según el filtrado que se considere para el estudio. El desarrollo de los programas se hará pensando en la capacidad de reproducir el proceso con otros datos sin necesidad de modificar el código (o alterando lo mínimo viable).

2.2.2 Representar gráficamente y analizar la correlación entre los grupos de estudio

Tras obtener los marcos de datos finales filtrados por diferentes grupos, se llevará a cabo la reducción de dimensionalidad y representación gráfica de los mismos, usando un script en el que resulte sencillo silenciar genes y modificar variables críticas. Este análisis visual se podrá fundamentar con otras técnicas.

3 RESULTADOS PRELIMINARES

Todo proyecto surge de unos resultados anteriores que justifican la necesidad de responder a ciertas preguntas científicas o de desarrollar ciertas aplicaciones biomédicas. En este trabajo, el punto de partida es un estudio preliminar realizado en pacientes de cáncer de mama hereditario dentro del equipo de bioinformática del VHIR. Se pretendía explorar si los paneles de genes se pueden utilizar para discriminar entre pacientes con cáncer de mama hereditario y controles. Para ello, se partió de datos de variantes obtenidos del análisis NGS de un panel de 159 genes en una cohorte de estudio compuesta de 1221 individuos de género femenino, de los cuales 175 eran controles y 1046 pacientes de cáncer de mama. Se anotó cada una de las variantes de estos genes de cada individuo, diferenciando aquellas que suponían un cambio en la secuencia peptídica en neutras y patológicas. A estas últimas se las clasificó según su grado de patogenicidad a partir de los predictores SIFT y PolyPhen-2. Las variantes también se diferenciaron entre heterocigotas y homocigotas, realizando un análisis que incluía ambas y otro solo con homocigotas, por ser más críticas para el fenotipo. Para valorar la viabilidad de realizar un algoritmo de clasificación de paneles (pacientes vs. control), el primer paso es analizar si realmente existe correlación entre las variantes y la enfermedad. Para el análisis visual fue necesario reducir la dimensionalidad usando las técnicas PCA y t-SNE. De forma alternativa, se utilizaron técnicas de *random forest*.

Con el fin de descartar que la representación de la información estuviera enmascarada por un exceso de variantes no informativas, se refinaron los genes de estudio a grupos más reducidos. El grupo *GS1* estaba compuesto por 10 genes relacionados de forma directa con el diagnóstico del cáncer de mama (Schroeder, C. et al, 2015) y el grupo *GS2* por 29 genes relacionados con la reparación de la rotura de doble hélice de DNA (DSB) por recombinación homóloga (GO:0000724 Double-strand break repair via homologous recombination, 2009).

Por otra parte, el dominio de BRCA1 y BRCA2 en las componentes principales impedía evaluar la influencia del resto de genes. Además, no se intuye correlación alguna entre la enfermedad y el número de variantes de ambos genes. Por este motivo se silenciaron y se repitió el análisis sin su presencia.

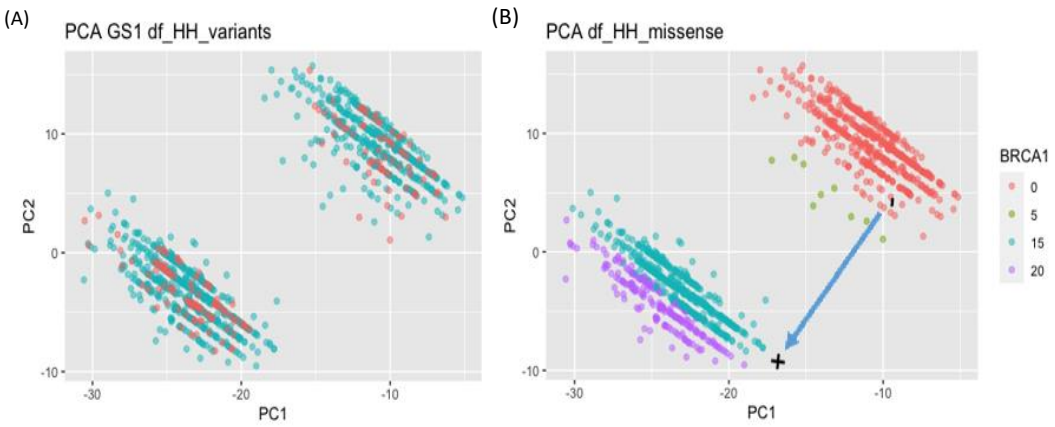


FIGURA 3. (A) REPRESENTACIÓN DE PC1 VS PC2 DEL ANÁLISIS DE LAS VARIANTES "MISSENSE" DE GENES DEL GRUPO GS1. CONTROLES (ROJO) Y PACIENTES (AZUL). (B) REPRESENTACIÓN DE PC1 VS PC2 DEL ANÁLISIS DE LAS VARIANTES "MISSENSE" DEL GRUPO GS1. DIFERENCIADOS POR NÚMERO DE VARIANTES BRCA1: 0-4 (ROJO), 5-9 (VERDE), 10-15 (AZUL) Y 16-20 (LILA)

A pesar de este filtrado y depurado de los datos, en una primera visualización no se observó discriminación entre controles y pacientes, tanto visualmente con representación de PCA y t-SNE como con la salida de *random forest*, descartando la hipótesis de una posible correlación.

Estos conocimientos previos constituyen el punto de partida del presente proyecto, tanto científico como técnico. En concreto, se ha podido adaptar alguno de los programas utilizados (o crear programas análogos), generalizándolo y permitiendo su aplicabilidad tanto al caso CVID como a otros estudios posteriores, añadiendo nuevos modos de visualización y predicción e implementando nuevas librerías más versátiles.

4 MATERIALES Y MÉTODOS

4.1 Cohortes

Una de las peculiaridades de este proyecto ha sido las cohortes de estudio, debido a sus características de etnicidad y el gran número de individuos incluidos en ellas. Aunque el rango de edad es más estrecho en la cohorte de los controles, en el estudio previo a los pacientes se determinó que el número de variantes no está interferido por edad ni sexo. A continuación, se desarrollan los rasgos de las cohortes.

4.1.1 Cohorte de pacientes

Los datos de pacientes del estudio CVID fueron facilitados por el departamento de inmunología del HVDH. Se secuenciaron las variantes de una cohorte de 131 pacientes residentes en Cataluña, con una edad comprendida entre 11-80 años, de los cuales 60 eran mujeres y 71 varones. El resultado de la causa genética desencadenante proporcionado por el departamento de inmunología, basándose en el estudio de dichas variantes, hasta el momento se divide en: 12 con mutación monogénica, 18 con deficiencia en proteína TACI "Activador transmembrana y modulador de calcio", 2 dudosos y 99 no concluyentes.

4.1.2 Cohorte de controles

Los datos de controles fueron amablemente cedidos por el instituto de investigación Germans Trias i Pujol (IGTP), liderado por el Dr Rafael de Cid. Los genomas se secuenciaron en el proyecto *"GCAT/Genomas por la vida*", un estudio de genómica en salud que investiga cuáles son los componentes genéticos y las

derivaciones, enfermedades y tratamiento (Obón-Santacana, M. et al, 2018). De dicho estudio, se seleccionaron las variantes de una cohorte de 785 individuos catalanes, con una edad comprendida entre los 40-66 años en el momento de la secuenciación, de los cuales 397 eran mujeres y 388 varones.

4.2 Caracterización de los genes de estudio

4.2.1 Genes del panel

El panel de estudio, compuesto por 313 genes, fue diseñado por el departamento de inmunología del HVDH basándose en sus años de experiencia en detección de enfermedades inmunes. Los genes poseen una función con impacto en el sistema inmune y pueden estar relacionados directa o indirectamente con el fenotipo CVID. El departamento también facilitó una lista de 91 genes del panel cuya función estaba más relacionada con la enfermedad. En un estudio superficial, se observa la predominancia de funciones de regulación celular, especialmente de linfocitos, y factores de transcripción. Las proteínas codificadas presentan amplio rango de tamaños con un alto porcentaje de dominios de regulación y transmembrana.

Además de la lista de 91 genes importantes, dentro del panel tenemos otra distinción de genes en los que están documentadas variantes específicas que desarrollan la enfermedad por sí solas o con ayuda de factores aún desconocidos. También se ha facilitado amablemente la información de los pacientes de la cohorte que las poseen. Sus características se detallan en el siguiente apartado.

4.2.2 Genes de diagnóstico

Haciendo uso de la bibliografía y las ontologías *amiGO* (http://amigo.geneontology.org/) y la ontología propia del IRB *Dsysmap* (https://dsysmap.irbbarcelona.org/) se ha realizado una caracterización de los genes que han sido calificados como desencadenante genético principal por el departamento de inmunología del HVDH. Debido a la gran influencia en desarrollo de la enfermedad de estos genes, el diagnóstico de quienes los portan es clasificado como monogénico, aunque también sea necesaria la actuación de otros factores. En este apartado se enumeran y describen aquellos genes utilizados habitualmente en el diagnóstico molecular del CVID y que han desempeñado un papel importante en los análisis de este trabajo.

- **BTK** *Tirosina quinasa de Bruton*. Este gen actúa en la maduración de linfocitos B. Mutaciones en BTK afectan a la fosforilación evitando la cascada de señalización, lo que se traduce en la no formación de linfocitos B maduros dando lugar a un cuadro de hipogammaglobulinemia (Rip, J. et al. 2018).

- **CTLA4** *Antígeno-4 asociado al Linfocito T Citotóxico*. Es un receptor proteico de la membrana celular de los linfocitos T, cuya función se ve inhibida por la estimulación de este receptor. Mutaciones en este gen provocan el desarrollo del *Síndrome Autoinmune Linfoproliferativo5*, un desorden de la apoptosis que causa acumulación de linfocitos autorreactivos, caracterizado por linfadenopatía no maligna con anemia hemolítica autoinmune y hepatoesplenomegalia, trombocitopenia y neutropenia (Kniffin, C.L. 2014).

- **DKC1** *Pseudouridina Sintasa 1 Dyskerina*. Gen que codifica la subunidad del complejo ribonucleoproteico H/ACA que participa en el mantenimiento de los telómeros. Los defectos en el gen pueden generar una multiplicación celular anormal. Algunas desencadenan una enfermedad hereditaria rara, la disqueratosis congénita (Knight, S. et al, 2001). Las personas con esta afección tienen un riesgo más alto de presentar insuficiencia de la médula ósea, síndrome mielodisplásico, fibrosis pulmonar y ciertos tipos de cáncer.

- **IKBKG** *Modulador Esencial NF-kappa-B***.** Gen ligado al cromosoma X que transcribe la subunidad reguladora del complejo IKK que fosforila inhibidores de NF-kappa-B, provocando la disociación del complejo y la degradación del inhibidor. Ciertas mutaciones se asocian a la *Inmunodeficiencia 33* (Converse, P.J. & Kniffin, C.L. 2007) que solo afecta a varones y se caracteriza por infecciones graves de inicio temprano a consecuencia de una disgammaglobulinemia y disminución de células B de memoria.

- **IKZF1** *IKAROS Familia de Dedos de Zinc 1*. La proteína de unión a DNA Ikaros es un factor de transcripción crítico para la diferenciación mieloide y linfocítica (Yoshida, N et al, 2017). La mayoría de los portadores de variantes alélicas caracterizadas por haploinsuficiencia presetan fenotipo CVID.

- **LRBA** codifica la proteína del mismo nombre (del inglés **"***Lipopolysaccharide-responsive and beige-like anchor protein***")** involucrada en la transducción de señales de acoplamiento y la deposición en la membrana de moléculas efectoras inmunes. Su mutación homocigota provoca deficiencia de LRBA y causa CVID8, caracterizada por el inicio temprano de infecciones respiratorias recurrentes con trastornos autoinmunes (Lopez-Herrera, G. et al, 2012).

- **NFKB1** *Factor nuclear NF-kappa-B (Subunidad1).* Codifica el precursor de una subunidad del factor de transcripción NF-kB. Variantes heterocigóticas "missense" en NFKB1 se han catalogado recientemente como la causa monogénica conocida más común de CVID entre los europeos, que da como resultado un defecto progresivo en la formación de células B (Tuijnenburg, P. et al, 2018).

- **PIK3R1** *Fosfatidilinositol 3-quinasa regulador subunidad alfa*. La quinasa PIK3 es necesaria para el aumento de captación de glucosa estimulado por insulina y la síntesis de glucógeno (National Center for Biotechnology Information, 2021). Modificaciones en su activación o su inactivación tienen consecuencias en el sistema inmune. Un ejemplo es el síndrome de PI3K-delta activado, un desorden caracterizado por infecciones respiratorias recurrentes, linfopenia, aumento de células B de transición circulantes, aumento de IgM y disminución de IgG en suero y alteración en la respuesta a vacunas (Preite, S. et al, 2019).

- **TNFRSF13B** *Miembro de la Superfamilia de Receptores de Factor de Necrosis Tumoral 13B*. La proteína que codifica es TACI (ligando de interacción de ciclofilina, activador transmembrana y calcio modular) y está directamente relacionado con la activación de linfocitos B y la CVID-2. Las mutaciones en TACI representan las variaciones de secuencia de DNA más comunes en individuos afectados por CVID, encontrándose en más del 10% de los sujetos (Park, M.A. et al, 2008). La mayoría de los parientes de un CVID portadores de mutaciones heterocigóticas en TACI no son hipogammaglobulinémicos. Aun así, presentan defectos detectables de células B in vitro, por lo que no se termina de comprender el papel de TACI en el desarrollo de la enfermedad (Martinez-Gallo, M. et al, 2013).

4.3 FLUJO DE TRABAJO

A continuación, se exponen los pasos seguidos para desarrollar el proyecto y las herramientas que se han usado en el proceso. Se explica el flujo de trabajo general y las peculiaridades de los datos propios.

4.3.1 Preprocesado y preparación de los datos

Los tránscritos secuenciados han sido analizados y facilitados para su caracterización bioinformática en el formato estándar *Variant Call Format* (VCF) (Danecek, P. et al, 2011). Los VCFs son ficheros de texto en los que se almacenan polimorfismos de un solo nucleótido (SNPs), indels y variaciones estructurales de la secuencia de genes y su información. Este formato se ha desarrollado a raíz de grandes proyectos de secuenciación del DNA como el Proyecto 1000 Genomas (The 1000 Genomes Project Consortium, 2010).

En el caso concreto de nuestro estudio, los genes de los pacientes y de los controles no han sido secuenciados siguiendo la misma técnica. Mientras que la secuenciación de los pacientes se ha realizado usando el panel mencionado de estudio de CVID del HVDH, los controles provienen de un estudio de WGS realizados para el proyecto GCAT y amablemente cedidos por el IGTP. Este hecho será importante a la hora de interpretar los resultados obtenidos, ya que la precisión de los cebadores usados puede ser inferior en el estudio WGS. Sumado al posible error durante el solapamiento de las secuencias para obtener su posición en el genoma, podría ser un artefacto decisivo en el análisis final. Los datos también difieren en el genoma de referencia empleado, siendo ChGR38 en el caso de los pacientes y ChGR37 en el de los controles. Para los controles trabajamos únicamente con las variantes correspondientes a los mismos genes que analizamos en el panel. Los datos de todos los controles se nos proporcionaron en el mismo archivo, por lo que fue necesario someterlo a un proceso de segregación por individuo. Los pasos ejecutados para preparar los datos de cara a su análisis posterior fueron:

- Cortar los VCF por IDs para poder estudiarlos de formar individual. Para ello es necesario obtener una lista con los IDs que queremos extraer a partir del encabezado del archivo original. A continuación, se creó un script de shell que busca en el archivo original los IDs de interés y separa toda la información del individuo en diferentes archivos, gracias al paquete *VCFtools.*

- En el caso de los controles, como los archivos provienen de WGS, el siguiente paso a realizar es un filtrado para seleccionar únicamente los genes de interés. Para ello seguiremos los pasos descritos a continuación. **Primero**, extraer el listado de las posiciones de genes de panel a través de un Bioconductor en R. Utilizando *biomaRt* y el dataset de *Ensembl* podemos obtener el listado con las posiciones de inicio y final de cada gen en el genoma de referencia GRCh37, y lo exportaremos a un archivo csv. **Segundo**, se comprimen los archivos VCFs, previamente segregados por individuo, a formato ".gzip" y para cada uno de ellos se crea un archivo en formato ".tbi" con los índices, mediante los comandos de la extensión tabix, agrupados en un shell script. **Y tercero**, crearemos los archivos finales en formato VCF mediante las consultas de tabix en python; estos archivos estarán formados únicamente por los genes de estudio del panel.

  Repositorio github: "separar_VCF.sh", "posicionvariantes37.R", "crear_tabix.sh", "filtroVCFs.py"
  Librerías: "VCFtools" y "Tabix" en Bash; "gzip", "os", "Tabix" y "pandas" en Python; "biomaRt" en R

4.3.2 Anotación de variantes

La anotación de variantes es un paso crítico en la interpretación de los datos biológicos obtenidos por NGS. La información de posición y de el/los nucleotido/s transmutados que se obtiene al referenciar los datos de secuenciación con el genoma de referencia es limitada desde un punto de vista funcional. Necesitamos comprender qué genes son portadores de variantes y cómo cada una de estas afecta a su función. Este proceso se conoce como anotación de las variantes y puede realizarse de diferentes maneras. En nuestro caso se ha utilizado la herramienta de código abierto de ensembl VEP: *Variant Effect Predictor* (Mc Laren, W., et al, 2016). VEP es de uso gratuito y reproducible, permite el análisis, anotación y priorización de variantes genómicas en regiones codificantes y no codificantes, proporcionando acceso a una amplia colección de anotaciones genómicas. La ejecución sería la siguiente:

- Anotar variantes con VEP. **Primero,** se ejecutó el análisis desde la página web de *Ensembl* (https://www.ensembl.org/Tools/VEP) de tres individuos con diagnóstico genético conocido. **Segundo**, tras corroborar la coherencia entre el resultado de la anotación con la herramienta y los datos de diagnósticos facilitados, se escribió un script perl/bash para automatizar el proceso en el resto de individuos. Además de la información obtenida por defecto, se añadieron los siguientes *pluggins*: "sift" y "polyphen" para conocer la puntuación y clasificación de las variantes según ambos predictores de patogenicidad; "symbol" para obtener la ID del gen en el que se encuentra la mutación; y "canonical", "biotype" y "regulatory", que aportan información adicional útil para filtrar los transcritos más adecuados para incluir en nuestro estudio. **Y tercero**, la alineación con los genomas de referencia se realizó de forma local con los archivos obtenidos de la página web oficial de *Ensembl*. Fue necesario crear una versión del script para la anotación de los controles, ya que VEP usa por defecto como referencia el genoma GRCh38 y la secuenciación de estos individuos se realizó con GRCh37. El output de este proceso da lugar a un archivo por individuo en formato txt.

  Repositorio github: "scriptvep.sh", "scriptvep37.sh,"
  Paquetes: "VEP" en Perl.

4.3.3 Procesado y selección de variantes

Una vez realizada la anotación de las variantes, se llevará a cabo la adición de información relevante que se pierde en el cambio de formato, como la frecuencia alélica, o que se ha obtenido por vías diferentes a VEP, como las puntuaciones de REVEL. También se filtrará la información para eliminar el ruido y poder trabajar con los tránscritos de interés, puesto que la selección de buenos datos es otro paso crucial para realizar un análisis coherente y de calidad. Para este propósito se ejecutarán los siguientes subprocesos, todos automatizados a través de Python:

- Añadir la descripción del genotipo. El hecho de que una variante sea bialélica generalmente se traduce en un efecto más pronunciado en el fenotipo; por esta razón es de suma importancia recuperar esta información no presente en el archivo de salida de VEP. Para recuperarla se usó el VCF original de cada paciente. La presencia alélica de la variante puede estar expresada de cuatro formas: 1/1, para aquellas variantes homocigóticas; 1/0, para las variantes que se encuentran en un solo alelo (heterocigóticas); 1/2, para aquellos locus que presentan dos variantes diferentes en la misma posición y ninguna de ellas coincide con el genoma de referencia. Por último, en las ocasiones en las que tenemos a todos los individuos en un mismo archivo, aquellos que son homocigotos para el genoma de referencia se

representan con 0/0. Para añadir esta información a nuestro archivo anotado, **primero** se creó una columna en el archivo VCF denominada *'#Uploaded_variation'* con las coordenadas cromosómicas y el cambio de nucleótido, creando así una identidad única para cada variante. Esta columna, presente en ambos archivos, se usa de guía para añadir el código de cigosidad en una columna nueva del archivo de variantes anotadas. **Segundo**, se sustituyen los códigos por su significado biológico (H para homocigoto y HH para heterocigoto, incluyendo los completamente heterocigotos "1/2") y **tercero**, se exporta el *dataframe* a un nuevo archivo en formato "*.csv".*

- Filtrar los trascritos de interés y eliminar los homocigotos para el genoma de referencia. Al analizar superficialmente los archivos obtenidos de la anotación de los pacientes con VEP, se observó que en un solo individuo encontrábamos alrededor de 13000 variantes en aproximadamente 500 genes diferentes. Este número distaba de los 313 genes del panel de secuenciación, sugiriendo un posible problema de procesado. Parte de la diferencia se puede explicar por la existencia del fenómeno "overlapping" genético (Makalowska, I. et al, 2005), es decir, la forma en la que los genes se superponen en las mismas regiones del genoma, lo cual supondría que una variante en esa región afectaría a varios genes. Aun y así, el número de variantes observado era demasiado elevado, teniendo en cuenta que supera a todas las variantes encontradas en los 785 individuos control (11791). Otro posible origen del problema es que la anotación de VEP para una mutación se representa tantas veces como funciones biológicas se vean afectadas. Para eliminar esta fuente de redundancia, se introdujo un filtro para seleccionar únicamente transcritos a través de su selección en la columna "Feature_type", en el apartado de biotipos se escogieron los que codifican proteínas y se seleccionaron los que cumplían el criterio "Canonical". Con este cribado nos quedamos con el 15% de las variantes iniciales en pacientes.

- Añadir las puntuaciones de REVEL. A diferencia de SIFT y PolyPhen, REVEL es un predictor actual que se encuentra en desarrollo constante y en el momento de nuestro análisis no existía la posibilidad de añadir el *plugging* en la anotación de VEP por no corresponder con la versión actual. Fue necesario incluir las puntuaciones a las variantes a través de un texto plano, obtenido de la web oficial de dbNSFP, con todas las variantes identificadas y su posición en los genomas de referencia hg19 y GRCh38. Su adaptación y adición a los archivos supuso un lapso de tiempo importante y un alto gasto computacional.

Repositorio Github: "añadir_frecuencia.py", "añadir_frecuencia_controles.py", "elegir_transcritos.py", "añadir_revel.py"
Librerías: "pandas" y "os" en Python

4.3.4 Selección y creación de los grupos de análisis

Para seleccionar los grupos de análisis se siguieron los criterios que se detallan a continuación:

- Consecuencia de la variante**.** Las variantes *missense* corresponden a un cambio en la secuencia de la proteína, pero no todas tienen la misma repercusión biológica. Las variantes neutras no tienen un efecto detectable sobre la función proteica. Por contra, las patogénicas tienen un impacto molecular que puede desencadenar un fenotipo clínico. Para distinguir las variantes neutras de las patogénicas, usamos los 3 métodos de predicción i*n silico* descritos previamente. Esta segregación da lugar a 5 grupos de análisis:

  1. Variantes de todo tipo
  2. Variantes "Missense". Es fácil segregarlas del resto, ya que VEP nos brinda por defecto la información de la consecuencia de la variante.
  3. Variantes patológicas según predicción SIFT. Se escogerán aquellas clasificadas como "deleterius".
  4. Variantes patológicas según predicción PolyPhen. Se escogerán aquellas clasificadas como "probably_damaging " o "possibly_damaging".
  5. Variantes patológicas según predicción REVEL. Se escogerán aquellas con puntuación mayor de 0.5

	Tipo	DNA	mRNA	Aminoácido
VARIANTE	REFERENCIA	**TTC**	**AAG**	Lys
	SINÓNIMA	**TT**T	**AA**A	Lys
	"NONSENSE"	**A**TC	**U**AG	-
	"MISSENSE" CONSERVATIVA	**T**C**C**	**A**G**C**	Arg
	"MISSENSE" NO CONSERVATIVA	**T**G**C**	**A**C**G**	Thr

FIGURA 4. EJEMPLO DEL SIGNIFICADO DE LOS DISTINTOS TIPOS DE SNPS. NUCLEÓTIDO MUTADO EN ROJO. AMINOÁCIDO CODIFICADO BÁSICO (AZUL), NEUTRO (AMARILLO) Y AUSENCIA DEL MISMO POR CODÓN DE STOP (ROJO). LAS VARIANTES "NONSENSE" O LAS "MISSENSE" NO CONSERVATIVAS SON AQUELLAS QUE TIENEN MAYOR PROBABILIDAD DE RESULTAR PATOGÉNICAS.

- Variantes homocigotas (H) y heterocigotas (HH). Como se ha comentado previamente, habitualmente las variantes homocigotas presentan un fenotipo más crítico y desencadenan respuesta en aquellas mutaciones de expresión recesiva. Esto hace que resulte interesante el estudio de todos los tipos de presentación de la mutación y también solo las variantes homocigotas por otro.

- Número de variantes reales en el gen (Único) y la suma de las funciones biológicas en las que participa y se ven afectadas (Todo). La causa raíz de esta división es comprobar si el aumento de consecuencias de una variante puede tener un significado biológico explícito, un ensayo de si es posible medir la magnitud biológica de la mutación partiendo del número de funciones afectadas.

- Genes del panel y genes importantes, La creación de un grupo de análisis únicamente con las variantes del listado de genes importantes facilitados por el departamento de inmunología nos permite acotar la búsqueda de la influencia de genes y eliminar el posible ruido de aquellos que no parecen repercutir, igual que se hizo en el estudio preliminar de cáncer de mama.

Después de toda esta selección, nos encontramos con 40 marcos de datos con sus características específicas y salida en formato ".csv". En ellos se sustituyen los valores ausentes de número de variantes por "0" y eliminamos columnas y filas de varianza 0, o lo que es igual, se eliminan las filas de pacientes que no presentan ninguna variante para los genes de análisis y las columnas de genes donde el número de variantes no difiere entre ninguno de los individuos. Se añade también otra información como el diagnóstico genético en aquellos casos en los que es conocido, el sexo y la edad.

Repositorio Github: "CREAR_DF.py"
Librerías: "pandas" y "os" en Python

4.3.5 Reducción de la dimensionalidad y representación gráfica

Una vez seleccionados, creados y depurados los grupos que queremos contrastar, vamos realizar su análisis a través de las técnicas de reducción de dimensionalidad y de representación gráfica antes citadas. Para ello se crearán dos programas en R con una estructura similar a diferencia de que en el segundo tendremos la visualización de pacientes segregada por rangos de variantes que presentan en un gen específico escogido por el usuario. Estos programas se dividen en cuatro bloques:

- Carga y elección de los datos. Tras cargar el marco de datos deseado, tenemos la opción de escoger una alícuota del mismo, por ejemplo, con un diagnóstico genético específico. También podemos silenciar algunos dentro del marco para poder ver la influencia de otros con más claridad. Por último, escogeremos como datos de entrenamiento únicamente los valores numéricos de las variantes.

- Análisis y representación de PCA. Los datos se centran y se estudian escalados y también con su valor original, puesto que todos están expresados en la mismas unidades (número de variantes). Los *loadings* nos aportan información de la participación de las variables originales en las componentes principales (PCs). Con *ggplot2* representamos las dos PCs a analizar, generalmente PC1 y PC2, y dependiendo de la distribución varianza, también PC3 vs PC4. Finalmente, la herramienta *Biplot* sitúa en el mapa de distribución los vectores de la aportación de cada gen y ayuda a comprender el significado de la gráfica.

- Análisis y representación tSNE. Se aplicó a través de la librería "Rtsne", en la que el hiperparámetro *perplexity* tiene un valor de 30 por defecto. En la creación del código se decidió que esta fuera una variable introducida por el usuario, para adaptarla a las distintas necesidades de cada *dataframe*. Gracias a esto y al conocimiento previo de los dos clústeres (CVID y control) conseguimos imágenes bastante informativas sin invertir un esfuerzo significativo en encontrar los parámetros de inicialización correctos. Su representación se realiza usando *ggplot*.

- Análisis y representación UMAP. El análisis por UMAP se lleva a cabo a través de la librería para R del mismo nombre y se mantienen los hiperparámetros usados por defecto. Su representación se realiza mediante *ggplot*.

Repositorio Github: "PINTARGENESINMUNO.R"
Librerías: "ggplot2", "Rtsne" y "umap" en R

5 RESULTADOS

En este apartado se detallan los resultados obtenidos utilizando las herramientas detalladas previamente. En los primeros análisis se estudiaron solo los pacientes CVID, pues eran los únicos datos disponibles. Cuando se comenzaron los análisis gráficos de los datos, se descartó continuar el estudio con el grupo que contaba todas las funciones afectadas por la variante, por ser el número de tránscritos del gen un gran artefacto y no demostrarse la coherencia biológica. Los resultados descritos a continuación corresponden al grupo de estudio que denominamos 'Único', en el que se trabajó únicamente con un tránscrito por gen.

5.1 Caracterización inicial de los paneles de pacientes CVID

Los análisis de PCA identificaron varios genes con un número de variantes superior al resto (Figura 5AB):

- **SPINK5**. Codifica el inhibidor de la serina proteasa Kazal-tipo 5, conocido como LEKTI (150kDa), No está documentado que sea directamente responsable del fenotipo CVID, pero sí está involucrado en otras enfermedades inmunes como dermatitis atópica y asma (Dežman, K. et al, 2017).
- **DOCK8**. Codifica la proteína del mismo nombre, que está implicada en el mantenimiento de la estructura celular, migración, adhesión y funcionamiento del sistema inmune (190kDa). Su deficiencia da lugar a una inmunodeficiencia combinada rara caracterizada por alergia y eosinofilia con elevados niveles de inmunoglobulina E, entre otros (Ruiz-García, R. et al, 2014).
- **SP110.** Es un antígeno nuclear (110kDa) estimulado por interferón presente en muchas células humanas Su deficiencia causa enfermedad venooclusiva con inmunodeficiencia (VODI), una PID grave donde el número de células T y B suele ser normal. Fabian Baldin demostró su papel como regulador inmunitario intrínseco clave de las células T en su tesis doctoral "*The role of Sp110 in human T cell apoptosis and immunopathology*" (2018).

Los tres genes codifican proteínas de gran tamaño y, exceptuando DOCK8, no parecen tener un historial de correlación con el fenotipo patológico CVID, por lo que no se descarta que esta superioridad en número de variantes sea consecuencia única del tamaño del gen y no tenga un valor explicativo de la patología.

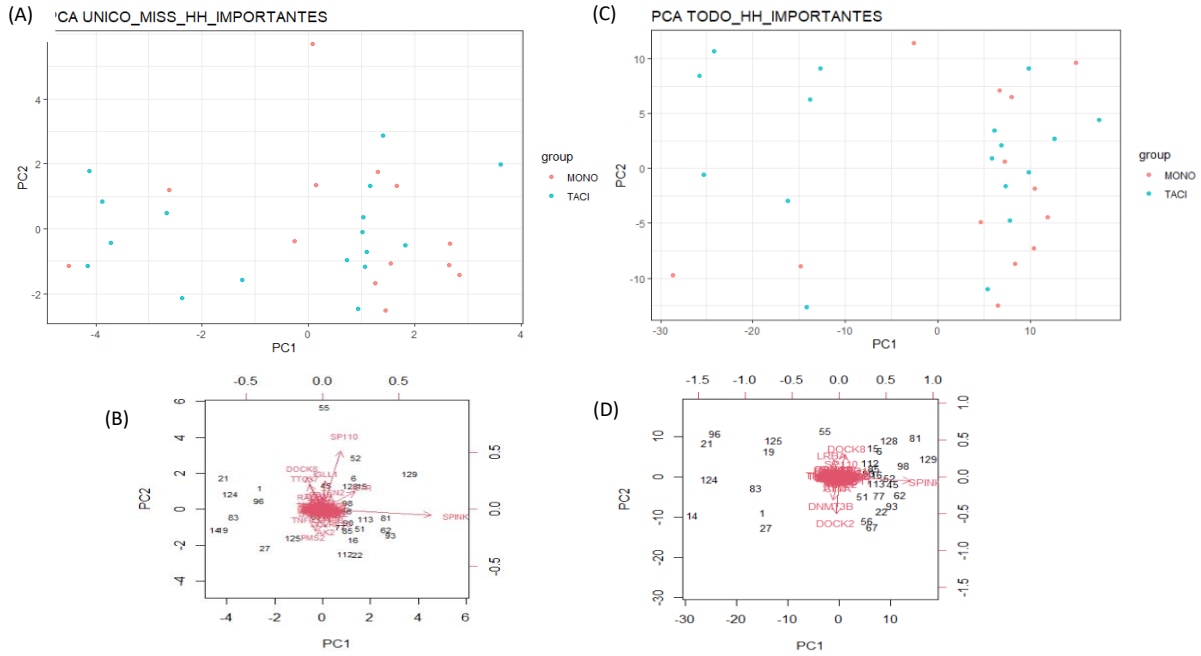


FIGURA 5 (A) ANÁLISIS PC1 VS PC2 DE LAS VARIABLES 'MISSENSE' HOMOCIGOTAS Y HETEROCIGOTAS (HH)EN GENES IMPORTANTES. LOS PACIENTES CON DIAGNÓSTICO NO CONCLUYENTE (NC) ESTÁN SILENCIADOS. SE MUESTRAN LOS PACIENTES CON CAUSA GENÉTICA MONOGÉNICA (ROJO) Y CON DEFECTOS EN TACI (AZUL). (B) REPRESENTACIÓN POR BIPLOT DE LOS VECTORES DE LOS GENES DE ESTUDIO RESPECTO A SU CONTRIBUCIÓN A LAS PCS. (C) ANÁLISIS PC1 VS PC2 DE LAS VARIABLES HH EN GENES IMPORTANTES. LOS PACIENTES CON DIAGNÓSTICO NC ESTÁN SILENCIADOS. SE MUESTRAN LOS PACIENTES CON CAUSA GENÉTICA MONOGÉNICA (ROJO) Y CON DEFECTOS EN TACI (AZUL). (D) REPRESENTACIÓN POR BIPLOT DE LOS VECTORES DE LOS GENES DE ESTUDIO RESPECTO A SU CONTRIBUCIÓN A LAS PCS.

A continuación, se analizó el resto de marcos de datos de variante única y se comprobó si existía algún sesgo en el número de variantes atribuible al sexo o a la edad (Figura 6). También se analizó la correlación entre estas dos últimas características y el diagnóstico genético. A pesar de que se esperaba una tendencia ligeramente diferente entre sexos debido a los genes ligados al cromosoma X, no se perciben grupos delimitados condicionados por este motivo.

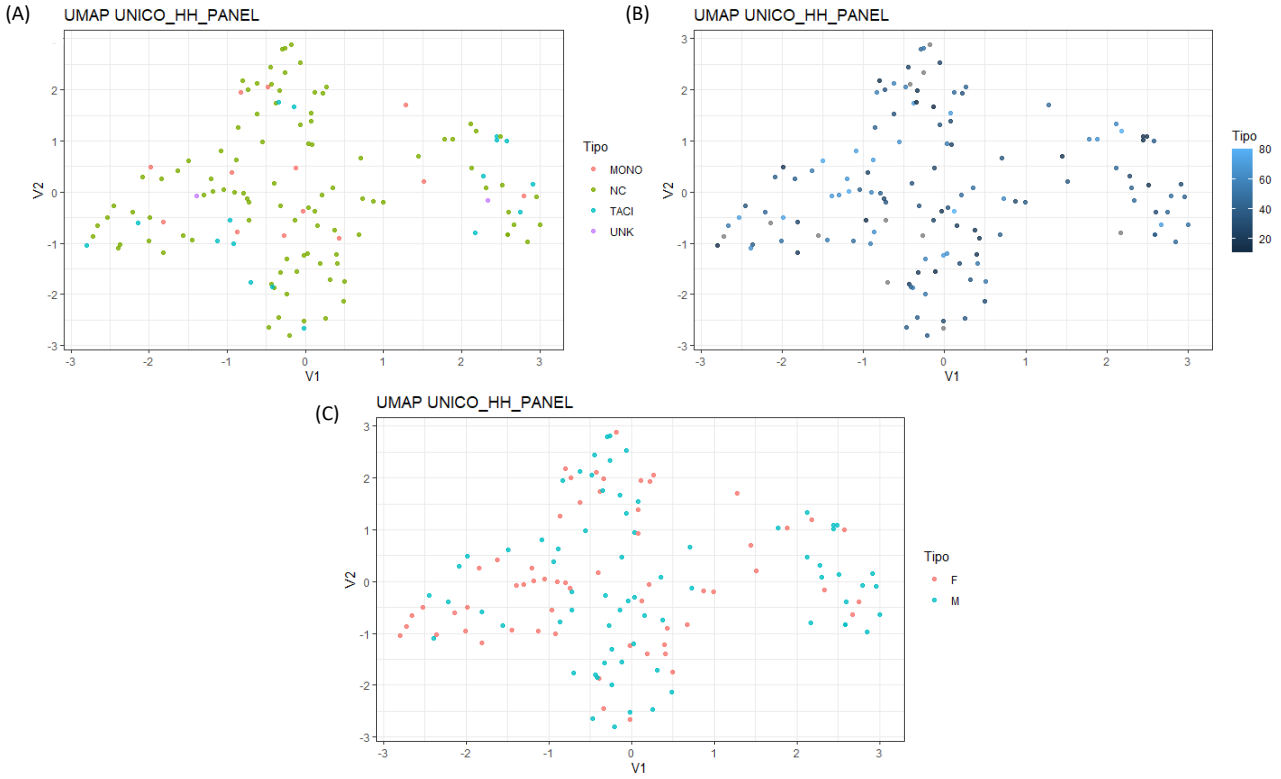


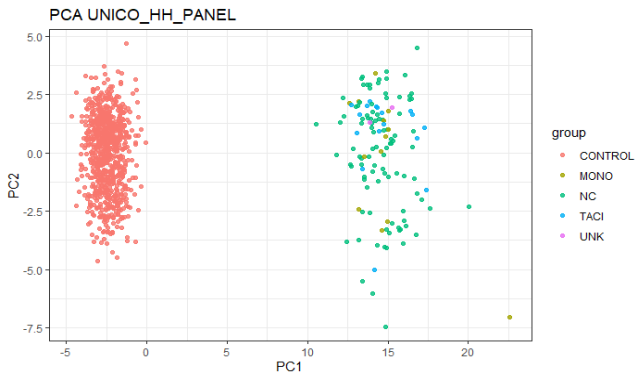
FIGURA 6 REPRESENTACIÓN DE UMAP DE LOS PACIENTES CVID ANALIZADOS POR SU NÚMERO DE VARIANTES EN TODOS LOS GENES DEL PANEL. (A) AGRUPADOS POR DIAGNÓSTICO GENÉTICO MONOGÉNICO (ROJO), NO CONCLUYENTE (VERDE), DEFECTO EN TACI (AZUL) Y DESCONOCIDO (VIOLETA). (B) REPRESENTADOS POR EDAD EN GRADIENTE DE COLOR (OSCURECE AL DISMINUIR LA EDAD). (C) AGRUPADOS POR SEXO, MUJERES (ROJO) Y HOMBRES (AZUL).

5.2 Estudio comparativo de los paneles de pacientes CVID vs. controles

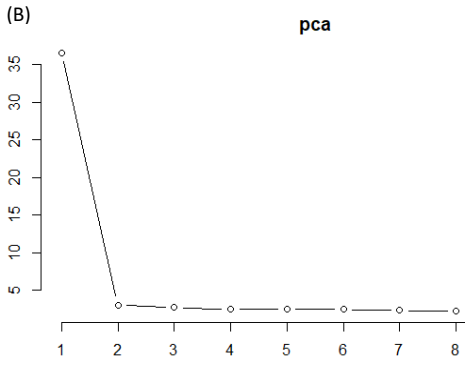
Una vez en posesión de los datos de controles, se llevó a cabo el análisis conjunto de las dos cohortes. En el análisis gráfico de PC1 vs PC2 de mayor número de datos, que incluye todo tipo variantes de genes del panel, se observa un gran alejamiento entre grupos de controles y los pacientes CVID (Figura 7A) que resulta artificial. En efecto, casi la totalidad de varianza se acumula en PC1 y la aportación de las componentes originales está repartida de forma casi equitativa entre un número relativamente alto de genes (Figura 7BC). Esta separación era más pronunciada al aumentar el volumen de datos en el análisis, ya fuera por aumento de número de genes de estudio o por la disminución del filtrado en las diferentes agrupaciones descritas en el apartado *"Procesado y selección de variantes".* Esto sugiere una tendencia general atribuible a efectos técnicos, ya que los paneles de controles y pacientes CVID tienen características técnicas diferentes. Para corroborar o refutar esta posibilidad, se representó la mediana del número de variantes totales para cada gen en controles vs pacientes CVID. La gráfica resultante (Figura 10A) muestra que el número de variantes es generalmente superior en pacientes, apoyando así la presencia de un componente técnico en los resultados.

Independientemente, en la representación gráfica PC1 vs PC2 de variantes de genes importantes-H-missense (Figura 8) se observa una distribución de los controles y CVID bastante coherente y en línea a lo que esperábamos encontrar al plantear el objetivo inicial. Esta tendencia parece menos afectada por el efecto técnico descrito, ya que, al realizar la gráfica de dispersión de la media del número de variantes de este grupo, se continúa observando una ligera tendencia como la descrita previamente, pero no aparece tan acentuada ni tan artificial (Figura 10B). Utilizamos otras técnicas como UMAP, para corroborar este resultado (Figura 9) pero la tendencia positiva no se reproducía, por lo que no podemos aceptar los datos de la Figura 8 como un apoyo a nuestra teoría inicial.

(A)



(B)

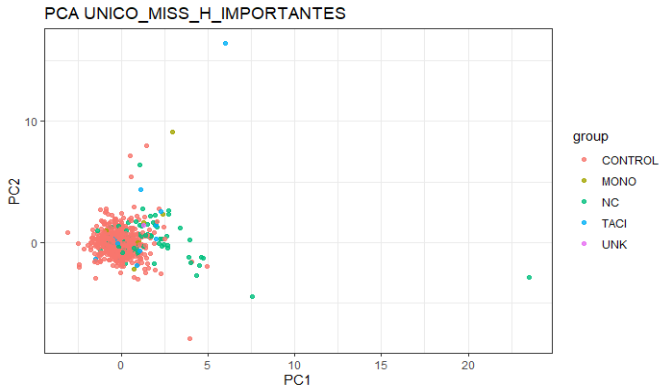


(C)

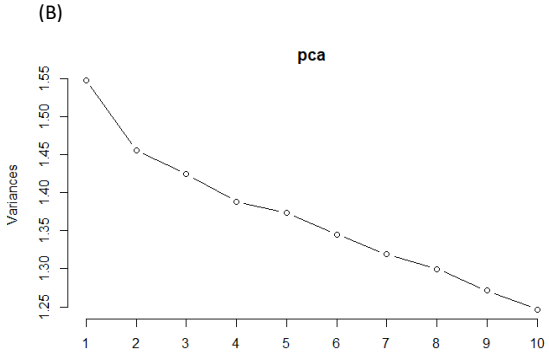
	STN1	MAP3K14	C4B	DGKE	AK2	RASGRP2	NLRP12	DKC1	BLNK	C4A
PC1	0.156829	0.147212	0.145442	0.140597	0.135014	0.134537	0.132416	0.131407	0.131214	0.129961

FIGURA 7 (A) REPRESENTACIÓN GRÁFICA PC1 VS PC2 DEL GRUPO TODAS LAS VARIANTES-HH-PANEL. (B) GRÁFICA DE VARIANZA EN LAS 8 PRIMERAS COMPONENTES PRINCIPALES. (C) LOS 10 GENES QUE MÁS APORTACIÓN TIENEN SOBRE PC1.

(A)



(B)



(C)

	CD3E	BTK	STAT3	TCN2	RNF168
PC1	0.5019683	0.4490092	0.3524424	0.1857234	0.1837673
	TNFRSF13C	**RASGRP1**	**XIAP**	**CIITA**	**IL7R**
PC2	0.4413160	0.4094651	0.2499697	0.2417884	0.2268353

FIGURA 8 (A) REPRESENTACIÓN GRÁFICA PC1 VS PC2 DEL GRUPO MISSENSE-H-IMPORTANTES. (B) GRÁFICA DE VARIANZA EN LAS 10 PRIMERAS COMPONENTES PRINCIPALES. (C) LOS 5 GENES QUE MÁS APORTACIÓN TIENEN SOBRE PC1 Y PC2.

FIGURA 9 REPRESENTACIÓN UMAP DEL GRUPO MISSENSE-H-IMPORTATES. NO SE APRECIA AGRUPACIÓN POR DIAGNÓSTICO.

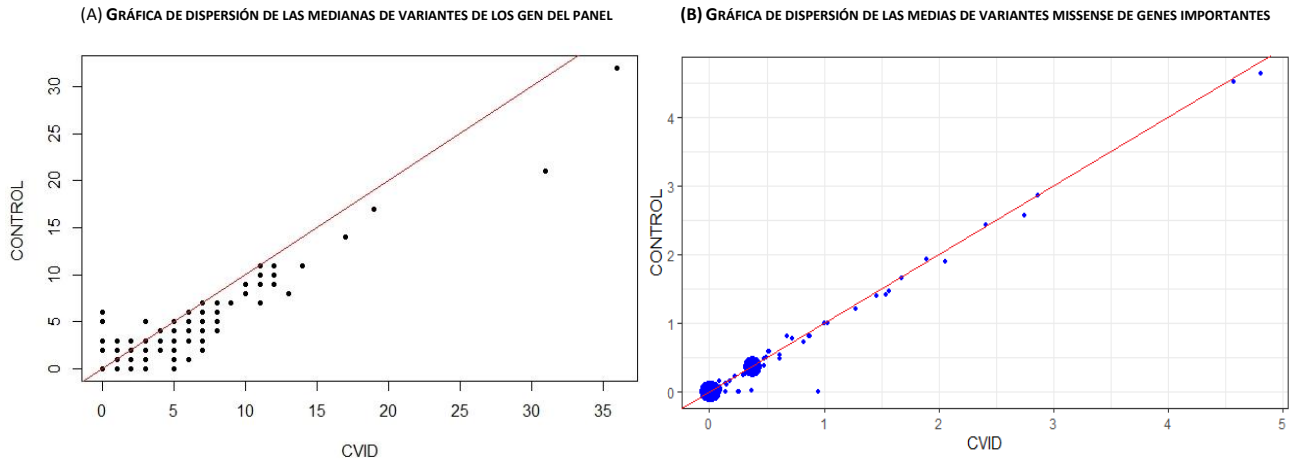


FIGURA 10 GRÁFICAS DE DISPERSIÓN (A) LAS MEDIANAS DE TODO TIPO DE VARIANTES EN LOS GENES DEL PANEL COMPLETO EN CONTROLES VS CVID. CADA PUNTO PUEDE REPRESENTAR MÁS DE UN GEN (NEGRO) , PUESTO QUE SE SOLAPAN AQUELLOS QUE TIENEN LOS MISMOS VALORES PARA AMBOS GRUPOS. (B) LAS MEDIAS DE LAS VARIANTES MISSENSE DEL GRUPO DE 91 GENES IMPORTANTES CONTROLES VS CVID. LOS PUNTOS QUE REPRESENTAN LOS GENES (AZUL) AUMENTAN DE TAMAÑO AL AUMENTAR LA DENSIDAD DE GENES CON ESE VALOR.

Se realizó un análisis bibliográfico de los genes que resultaron más influyentes en el análisis las variantes *missense* H de genes importantes (Figura 8C). En general, las proteínas que codifican son de un tamaño menor que los observados en el estudio que solo incluía a los pacientes. Las funciones de sus tránscritos tienen un fuerte impacto en el sistema inmune y, en algunos casos concretos, con el fenotipo CVID en específico. Este resultado presenta un valor biológico más coherente que lo concluido en la fase previa. A continuación, se aportan las características de dichos genes obtenidas tras un análisis de la literatura.

- **AK2**. La Adenilato quinasa 2 (26kDa) cataliza la transferencia reversible de un fosfato entre ATP y AMP, regulando la homeostasis de ATP, indicativa para la supervivencia, proliferación y función de las células inmunes. Ciertas mutaciones provocan la inmunodeficiencia combinada severa (SCID) más grave, la disgenesia reticular, con ausencia total de respuesta inmune (Campos, A. & Mendes, P.M. 2020).

- **CD3E**. Es el precursor de la glicoproteína de membrana de células T CD3-epsilon (20kDa). Su función es la transducción de señales y su deficiencia es crítica para la inmunidad adaptativa, estando directamente relacionada con la inmunodeficiencia 18 y la SCID B+T- (de Saint Basile, G. et al, 2004).

- **CIITA.** El transactivador del complejo mayor de histocompatibilidad clase II (CIITA) (123kDa). Defectos en el gen pueden desencadenan varios fenotipos clínicos, como el síndrome de linfocitos desnudos tipo II, debido a que MHC II es indispensable para la presentación de antígenos extraños a las células T CD4 + y lograr así una respuesta inmune adaptativa exitosa (Steimle, V. et al, 1993).

- **IL7R.** El receptor de interleucina-7 es una proteína que se encuentra en la superficie celular (53*kDa)* involucrada en el desarrollo de los linfocitos. Defectos en su expresión se relacionan con una SCID (Puel, A. et al1998).

- **RNF168**. Ligasa de Ubiquitina asociada a la cromatina (65kDa) implicada en la reparación de rotura DSB. Mutaciones en el gen provocan radiosensibilidad e inmunodeficiencia, entre otros (Chinn, I.K. et al, 2017).

- **STAT3** Factor de transcripción (80kDa) que actúa en multitud de procesos celulares como la proliferación celular o la apoptosis de linfocitos. Por ello no sorprende que se hayan descrito variantes en pacientes CVID y que se apunte como posible gen de diagnóstico monogénico (Russel, M.A. et al, 2018).

- **TCN2**. Codifica la transcobalamina, un transportador de la vitamina B12 (47kDa*).* Su deficiencia provoca un trastorno que puede comprometer el sistema inmune a través de la agammaglobulinemia y la pancitopenia (Zhan, S. et al, 2020).

- **TNFRSF13C**. Codifica el receptor BAFF o 'miembro de la superfamilia de receptores de factor de necrosis tumoral 13C' (33kDa). Su función está estrechamente relacionada con la maduración y

supervivencia de los linfocitos B. Se ha descrito la presencia de variantes en pacientes CVID (Losi, C.G. et al, 2005).

- **XIAP**. Codifica el Inhibidor de la apoptosis ligado al cromosoma X (57kDa), el más potente de la familia y con un gran impacto en la homeostasis de linfocitos. Variantes provocan el síndrome proliferativo ligado al cromosoma X (XLP), que en ocasiones es mal diagnosticado como CVID. La diferencia entre ambas es que la hipogammaglobulinemia de XLP tiene origen desconocico, puede tener un inicio tardío e incluso ser transitoria (Rigaud, S et al, 2011).

5.3 Análisis de las variantes potencialmente patogénicas

Finalmente, se decidió analizar más en profundidad la contribución de las variantes que se han considerado patológicas. Observamos que los predictores SIFT y PolyPhen daban un número demasiado elevado de ellas, tendencia consistente con la tasa de falsos positivos propia de estas dos técnicas. Por este motivo, centramos nuestro análisis en las predicciones realizadas con REVEL, una técnica con especificidad y sensibilidad más comparables y una mayor tasa de aciertos (Ioannidis, N.M. et al, 2016). Los resultados obtenidos coincidieron casi en su totalidad con los casos previamente diagnosticados genéticamente por el departamento de Inmunología. Sin embargo, llamó la atención el caso de un individuo cuyo diagnóstico genético había resultado no concluyente. Era el único de estas características que aparecía en el análisis de variantes patogénicas según puntuación de REVEL del grupo de 91 genes importantes (Figura 11) . Se comprobó que el varón era portador hemicigoto de una variante patogénica MAGT1, un gen ligado al cromosoma X que codifica la proteína transportadora de magnesio I. Su deficiencia desencadena una forma leve de inmunodeficiencia combinada (CID) denominada XMEN, solo presente en varones, caracterizada por infecciones recurrentes del virus Epstein-Barr (VEB) (Ravell, J. et al 2014). Como consecuencia de nuestro análisis, el diagnóstico de este individuo va a ser revisado por el grupo de Inmunología.

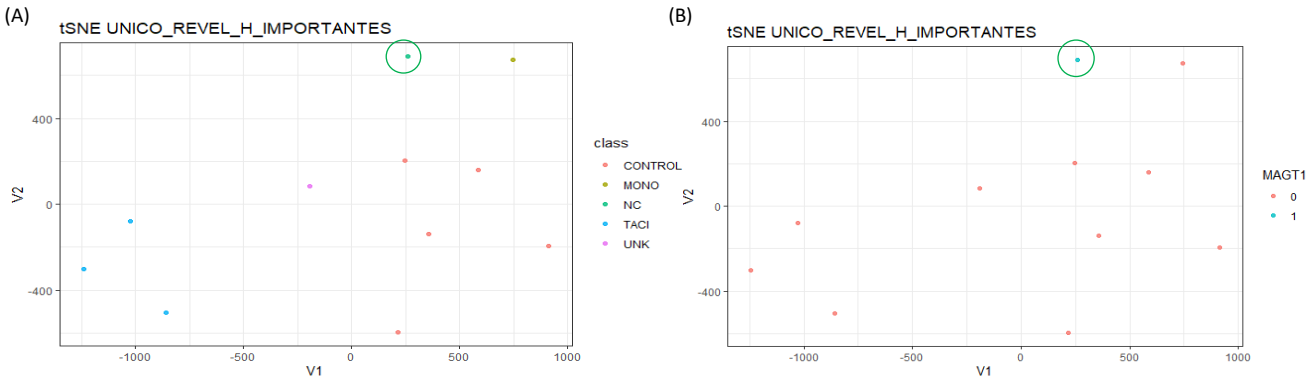


FIGURA 11. REPRESENTACIÓN TSNE DE VARIANTES DE CONTROLES VS CVID, DEL GRUPO H/IMPORTANTES CON PUNTUACIÓN EN REVEL SUPERIOR A 0.5. (A) AGRUPADOS POR DIAGNÓSTICO GENÉTICO. (B) AGRUPADOS POR NÚMERO DE VARIANTES EN EL GEN MAGT1. SE OBSERVA UN ÚNICO INDIVIDUO CON LA PRESENCIA DE VARIANTE PATOLÓGICA EN EL GEN (SEÑALADO EN VERDE EN AMBAS GRÁFICAS).

6 DISCUSIÓN

El flujo de trabajo propuesto, que pretende resolver el objetivo inicial de caracterizar los genes asociados a CVID, entre otros, se ha ido adaptando en el transcurso del mismo. Durante el proceso de obtención de resultados se ha realizado un cribado descartando aquellos métodos o herramientas que no aportaban información relevante para alcanzar el objetivo, incluyendo en este informe final solo unos pocos de los análisis efectuados en este trabajo. A nivel técnico, se ha buscado facilitar el uso posterior de esta pipeline bioinformática para que sea aplicable a grupos de datos diversos. Para ello se han conservado en los programas creados las herramientas no usadas finalmente en el análisis, ya que podrían ser útiles en el estudio de otros datos.

Dentro del transcurso del mismo proyecto se ha evaluado la efectividad del flujo de trabajo, poniéndola a punto en el proceso de anotación, predicción y análisis conjunto de los grupos de variantes de pacientes. Este proceso duró aproximadamente tres meses. Posteriormente, se aplicó el flujo de trabajo a la información

de los controles, disponible tres meses y medio después de iniciar el proyecto. A pesar de que esta cohorte contaba con 6 veces más individuos que la cohorte CVID, que su secuenciación se había obtenido mediante WGS y que estaba comparado con otro genoma de referencia, la duración del análisis se redujo de los tres meses iniciales a 5 días.

En lo que respecta a los resultados propios de la investigación, se ha conseguido caracterizar algunos de los genes con una incidencia de variantes superior en los pacientes CVID en España y también aquellos que podrían distinguirlos del grupo de controles. Estos genes han sido listados junto con las características que los vinculan a CVID en el apartado de resultados. En general, la reducción de dimensionalidad por PCA y su posterior análisis gráfico ha sido la técnica empleada que ha resultado más informativa, ya que de ella hemos podido obtener estos genes con número de variantes predominantes.

También se ha comprobado qué características epidemiológicas propias del individuo como el sexo o la edad no parecen tener un efecto en el número de variantes, y extendiendo nuestros objetivos iniciales, se ha encontrado un caso individual que podía preceder al diagnóstico genético de uno de los pacientes que formaban nuestra cohorte de estudio.

Finalmente, señalar que los resultados obtenidos no confirman que la diferencia en el número de variantes con impacto molecular en los genes descritos permita discriminar entre pacientes CVID e individuos sanos. La mayor dificultad a la que nos hemos visto enfrentados a la hora de abordar el estudio ha sido la segregación tan pronunciada entre los grupos CVID y control que crecía al aumentar el volumen de datos estudiados, procedente del artefacto producido por la diferencia del método de secuenciación. Esto nos confirma la necesidad de unificar la forma de extracción y procesado de los datos para poder estudiar su naturaleza de forma veraz.

7 CONCLUSIONES

La CVID es una enfermedad inmune con una variabilidad clínica, etiológica y genética pavorosa. Por ello, el hecho de que se haya identificado un número considerable de genes que en españoles caucásicos CVID tienen un número mayor de variantes que en controles sanos resulta un avance valioso. Algunos de estos genes podrían ser candidatos diana para un diagnóstico monogénico, igual que los que se presentaban en el diagnóstico del departamento de Inmunología, o podrían formar parte de un panel más acotado, útil (i) para el diagnóstico clínico de CVID, y (ii) para la obtención de datos de controles sanos obtenidos de la misma forma que los pacientes. Esta segunda función surge de la necesidad de una mayor coherencia en las cohortes de análisis, ya que la diferencia entre ambas ha impedido la correcta interpretación de los análisis realizados.

Por otro lado, el flujo de trabajo y los programas creados han mostrado un esquema lógico con unos resultados robustos y reproducibles que permitieron alcanzar las metas técnicas propuestas. Las grandes diferencias entre el modo de obtención de los datos de las cohortes, ha permitido comprobar la adaptabilidad de las herramientas. Estos mismos programas, con modificaciones menores, serán una herramienta valiosa para proseguir esta investigación desde enfoques diferentes; por ejemplo, llevando el punto de atención a aquellas variantes raras de baja frecuencia de la población española, especialmente las que no han sido identificadas. Este punto de vista es interesante para una continuación del estudio debido a la naturaleza descrita previamente en la enfermedad.

Tras mi estancia realizando este trabajo final de máster con el equipo de Bioinformática Clínica y Traslacional del HVDH puedo concluir que se han logrado gran parte de los objetivos descritos al inicio del proyecto. El camino para comprender y manejar las enfermedades poligénicas es desconocido, complicado y largo, pero es necesario emprenderlo y cada paso en aras de su descubrimiento siempre será un avance.

8 BIBLIOGRAFÍA

Abecasis, G., Auton, A., Brooks, L., DePristo, M., Durbin, R., Handsaker, R., . . . 1000GenomesProjectConsortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*, 55-65. doi:10.1038/nature11632

Adzhubei, I., Jordan, D., & Sunyaev, S. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*(7). doi:10.1002/0471142905.hg0720s76

Allaoui, M., Kherfi, M., & Cheriet, A. (2020). Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. *Lecture Notes in Computer Science*. doi:10.1007/978-3-030-51935-3_3

Ameratunga, R., Lehnert, K., Woon, S., Gillis, D., Bryant, V., Slade, C., & Steele, R. (2018). Review: Diagnosing Common Variable Immunodeficiency Disorder in the Era of Genome Sequencing. *Clin Rev Allergy Immunol.*, 261-268. doi:10.1007/s12016-017-8645-0

Baldin, F. (2018). The role of Sp110 in human T cell apoptosis and immunopathology. *Doctoral Thesis*. University of Basel, Faculty of Science.

Bisgin, A., Boga, I., Yilmaz, M., Bingol, G., & Altintas, D. (2018). The utility of next-generation sequencing for primary immunodeficiency disorders: experience from a clinical diagnostic laboratory. *Biomed. Res. Int.* doi:10.1155/2018/9

Bisgin, A., Sonmezler, O., Boga, I., & Yilmaz, M. (2021). The impact of rare and low-frequency genetic variants in common variable immunodeficiency (CVID). *Sci Rep 11*. doi:10.1038/s41598-021-87898-1

Bogaert, D., Dullaers, M., Lambrecht, B., Vermaelen, K., De Baere, E., & Haerynck, F. (2016). Genes, associated with common variable immunodeficiency: one diagnosis to rule them all? *J Med Genet*(53), 575-590. doi:10.1136/jmedgenet-2015-103690

Bonilla, F., Barlan, I., Chapel, H., Costa-Carvalho, B., Cunningham-Rundles, C., de la Morena, M., . . . Warnatz, K. (2016). International Consensus Document (ICON): common variable immunodeficiency disorders. *J Allergy Clin Immunol Pract., 4*, 38-59. doi:10.1016/j.jaip.2015.07.025

Bush, W., & Moore, J. (2012). Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology, 8*(12). doi:10.1371/journal.pcbi.1002822

Campos, A., & Mendes, P. (2020). AK2 deficiency: An awkward tale for B cells. *The Journal of Allergy and Clinical Immunology*, 74-76. doi:10.1016/j.jaci.2020.04.060

Chinn, I., Sanders, R., Stray-Pedersen, A., Coban-Akdemir, Z., Kim, V., Dadi, H., . . . Hanson, C. (2017). Novel Combined Immune Deficiency and Radiation Sensitivity Blended Phenotype in an Adult with Biallelic Variations in ZAP70 and RNF168. *Front Immunol., 576*(8). doi:10.3389/fimmu.2017

Converse, P. J., & Kniffin, C. L. (2 de 5 de 2007). *IMMUNODEFICIENCY 33; IMD33*. Recuperado el 2021, de OMIM: https://www.omim.org/entry/300636#12

Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M., . . . 1000GenomesProjectAnalysisGroup. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

de Saint Basile, G., Geissmann, F., Flori, E., Uring-Lambert, B., Soudais, C., Cavazzana-Calvo, M., . . . Le Deist, F. (2004). Severe combined immunodeficiency caused by deficiency in either the δ or the ε subunit of CD3. *J Clin Invest, 114*, 1512-1517. doi:10.1172/JCI22588.

Dežman, K., Korošec, P., Rupnik, H., & Rijavec, M. (2017). SPINK5 is associated with early-onset and CHI3L1 with late-onset atopic dermatitis. *Int J Immunogenet, 44*, 212-218. doi:10.1111/iji.12327

Flanagan, S., Patch, A., & Ellard, S. (2010). Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers, 14*(5), 533-537. doi:10.1089/gtmb.2010.0036

GO:0000724 Double-strand break repair via homologous recombination. (2009). Obtenido de amiGO: http://amigo.geneontology.org/amigo/term/GO:0000724

Grimbacher, B. (2014). The European Society for Immunodeficiencies (ESID). *Clin. Exp. Immunol., 178*, 18-20. doi:10.1111/cei.12496

Gupta, A., & Gupta, U. (2013). Next Generation Sequencing and Its Applications. *Animal Biotechnology: Models in Discovery and Translation.*, 345-367. doi:10.1016/B978-0-12-416002-6.00019-5

Hinton, G. &. (2002). Stochastic neighbor embedding. *NIPS, 15*, 833-840.

Holm, A., Aukrust, P., Damås, J., Müller, F., Halvorsen, B., & Frøland, S. (2004). Abnormal interleukin-7 function in common variable immunodeficiency. *Blood, 105*(7), 2887-2890. doi:10.1182/blood-2004-06-2423

Ioannidis, N., Rothstein, J., Pejaver, V., Middha, S., McDonnell, S., Baheti, S., . . . Lange, E. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics, 99*(4), 877-885. doi:10.1016/j.ajhg.2016.08.016

Kniffin, C. L. (19 de 11 de 2014). *AUTOIMMUNE LYMPHOPROLIFERATIVE SYNDROME, TYPE V; ALPS5*. Recuperado el 08 de 2021, de ONIM: https://www.omim.org/entry/616100

Knight, S., Vulliamy, T., Morgan, B., Devriendt, K., Mason, P., & Dokal, I. (2001). Identification of novel DKC1 mutations in patients with dyskeratosis congenita: implications for pathophysiology and diagnosis. *Human Genetics, 108*, 299-303. doi:10.1007/s004390100494

Kobak, D. &. (2019). UMAP does not preserve global structure any better than t-SNE when using the same initialization. *bioRxiv*.

Leonardi, L., Lorenzetti, G., Carsetti, R., Ferrari, S., Di Felice, A., Cinicola, B., & Duse, M. (2019). Rare TACI Mutation in a 3-Year-Old Boy With CVID Phenotype. *Frontiers in Pediatrics, 7*(418). doi:10.3389/fped.2019.00418

Li, W., Cerise, J., Yang, Y., & and Henry Han. (2017). Application of t-SNE to human genetic data. *Journal of Bioinformatics and Computational Biology, 15*(4). doi:10.1142/S0219720017500172

Liu, X., Li, C., Mou, C., Dong, Y., & Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine, 12*(103). doi:10.1186/s13073-020-00803-9

Lopez-Herrera, G., Tampella, G., Pan-Hammarström, Q., Herholz, P., Trujillo-Vargas, C., Phadwal, K., . . . Dideberg, V. (2012). Deleterious Mutations in LRBA Are Associated with a Syndrome of Immune Deficiency and Autoimmunity. *The American Journal of Human Genetics, 90*(6), 986-1001. doi:10.1016/j.ajhg.2012.04.015

Losi, C., Silini, A., Fiorini, C., Soresina, A., Meini, A., Ferrari, S., . . . Plebani, A. (2005). Mutational analysis of human BAFF receptor TNFRSF13C (BAFF-R) in patients with common variable immunodeficiency. *J Clin Immunol*, 496-502. doi:10.1007/s10875-005-5637-2

Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences, 19*(3), 303-342. doi:10.1016/0098-3004(93)90090-R

Makalowska, I., Lin, C. F., & Makalowski, W. (2005). Overlapping genes in vertebrate genomes. *Computational biology and chemistry 29 (1)*, 1-12. doi: 10.1016/j.compbiolchem.2004.12.006

Martinez-Gallo, M., Radigan, L., Almejun, M., Martinez-Pomar, N., Matamoros, N., & Cunningham-Rundles, C. (2013). TACI mutations and impaired B-cell function in subjects with CVID and healthy heterozygotes. *J Allergy Clin Immunol, 131*(2), 468-476. doi:10.1016/j.jaci.2012.10.029

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Graham, R. R., Thormann, A., . . . Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology 17 (1), 112*(17), 1-14. doi:10.1186/s13059-016-0974-4

Morganti, S., Tarantino, P., Ferraro, E., D'Amico, P., Viale, G., Trapani, D., . . . Curigliano, G. (2018). Complexity of Genome Sequencing and Reporting: Next generation sequencing (NGS) technologies and implementation of Precision Medicine in Real Life. *Critical Reviews in Oncology / Hematology*, 171-182. doi:10.1016/j.critrevonc.2018.11.008

National Center for Biotechnology Information. (2021). *PubChem Protein Summary for NCBI Protein P27986*. Recuperado el 08 de 2021, de https://pubchem.ncbi.nlm.nih.gov/protein/P27986.

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

Obón-Santacana, M., Vilardell, M., Carreras, A., Duran, X., Velasco, J., Galván-Femenía, I., . . . de Cid, R. (2018). GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open 10.1136/bmjopen-2017-018324*. doi:10.1136/bmjopen-2017-018324

Park, M., J.T., L., Hagan, J., Maddox, D., & Abraham, R. (2008). Common variable immunodeficiency: a new look at an old disease. *Lancet, 372*, 489-502. doi:10.1016/S0140-6736(08)61199-X

Preite, S., Gomez-Rodriguez, J., Cannons, J., & Schwartzberg, P. (2019). T and B-cell signaling in activated PI3K delta syndrome: From immunodeficiency to autoimmunity. *Special Issue: Signaling and Signal Diversification in Immune Cells, 291*(1), 154-173. doi:10.1111/imr.12790

Puel, A., Ziegler, S., Buckley, R., & Leonard, W. (1998). Defective IL7R expression in T-B+NK+ severe combined immunodeficiency. *Nat Genet, 20*, 394-397. doi:10.1038/3877

Ravell, J., Chaigne-Delalande, B., & Lenardo, M. (2014). X-linked immunodeficiency with magnesium defect, Epstein-Barr virus infection, and neoplasia disease: a combined immune deficiency with magnesium defect. *Current opinion in pediatrics, 26*(6), 713-771. doi:10.1097/MOP.0000000000000156

Resnick, E., Moshier, E., Godbold, J., & Cunningham-Rundles, C. (2012). Morbidity and mortality in common variable immune deficiency over 4 decades. *Blood, 119*, 1650-1657. doi:10.1182/blood-2011-09-377945

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Rehm, H. (2015). Standars 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*, 405-424. doi:10.1038/gim.2015.30

Rigaud, S., Lopez-Granados, E., Sibéril, S., Gloire, G., Lambert, N., Lenoir, C., . . . Latour, S. (2011). Human X-linked variable immunodeficiency caused by a hypomorphic mutation in XIAP in association with a rare polymorphism in CD40LG. *Blood, 118*(2), 252-261. doi:10.1182/blood-2011-01-328849

Rip, J., Van Der Ploeg, E., Hendriks, R., & Corneth, O. (2018). The Role of Bruton's Tyrosine Kinase in Immune Cell Signaling and Systemic Autoimmunity. *Crit Rev Immunol., 38*(1), 17-62. doi:10.1615/CritRevImmunol.2018025184

Ruiz-García, R., Lermo-Rojo, S., Martínez-Lostao, L., Mancebo, E., Mora-Díaz, S., Paz-Artal, E., & Allende, L. M. (2014). A case of partial dedicator of cytokinesis 8 deficiency with altered effector phenotype and impaired CD8+ and natural killer cell cytotoxicity. *Journal of Allergy and Clinical Immunology*, 218-221. doi:10.1016/j.jaci.2014.01.023

Russel, M., Pigors, M., M.E., H., Manson, A., Kelsell, D., Longhurst, H., & Morgan, N. (2018). A novel de novo activating mutation in STAT3 identified in a patient with common variable imminodeficiency (CVID). *Clinical Immunology, 187*, 132-136. doi:10.1016/j.clim.2017.11.007

Saudi Mendeliome Group. (2015). Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. *Genome Biology, 16*(134). doi:10.1186/s13059-015-0693-2

Schroeder, C., Faust, U., Sturm, M., Hackmann, K., Grundmann, K., Harmuth, F., . . . Rump, A. (2015). HBOC multi-gene panel testing: Comparison of two sequencing centers. *Breast Cancer Rest Treat.*, 129-136. doi:10.1007/s10549-015-3429-9

Sklarz, T., Hurwitz, S., Stanley, N., Juusola, J., Bagg, A., & Babushok, D. (2020). Aplastic anemia in a patient with CVID due to NFKB1 haploinsufficiency. *Cold Spring Harb Mol Case Stud., 6*(6). doi:10.1101/mcs.a005769

Steimle, V., Otten, L., Zufferey, M., & Mach, B. (1993). Complementation cloning of an MHC class II transactivator mutated in hereditary MHC class II deficiency (or bare lymphocyte syndrome). *Cell, 75*, 135-146.

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*, 1061–1073. doi:10.1038/nature09534

Tuijnenburg, P., Allen, H., Burns, S., Greene, D., Jansen, M., Staples, E., . . . de Bree. (2018). Loss-of-function nuclear factor κB subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans. *Journal of Allergy and Clinical Immunology, 142*(4), 1285-1296. doi:10.1016/j.jaci.2018.01.039.

Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

Velliangiri, S., Alagumuthukrishnan, S., & Iwin Thankumar joseph, S. (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science, 165*, 104-111. doi:10.1016/j.procs.2020.01.079

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to use t-SNE effectively. *Distill, 1*(10).

Yazdani, R., Habibi, S., Sharifi, L., Azizi, G., Abolhassani, H., Olbrich, P., & Aghamohammadi, A. (2019). Common Variable Immunodeficiency: Epidemiology, Pathogenesis, Clinical manifestations, Diagnosis, Classification and Management. *Journal of Investigational Allergology and Clinical Immnunology*, 30. doi:10.18176/jiaci.0388

Yoshida, N., Sakaguchi, H., Muramatsu, H., Okuno, Y., Song, C., Dovat, S., . . . Kojima, S. (2017). Germline IKAROS mutation associated with primary immunodeficiency that progressed to T-cell acute lymphoblastic leukemia. *Leukemia, 31*, 1221-1223. doi:10.1038/leu.2017.25

Zhan, S., Cheng, F., He, H., Hu, S., & Feng, X. (2020). Identification of transcobalamin deficiency with two novel mutations in the TCN2 gene in a Chinese girl with abnormal immunity:a case report. *BMC Pediatr, 20*(246). doi:10.1186/s12887-020-02357-6