UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Química Biológica

Desarrollo de técnicas de bioinformática estructural para la evaluación y comprensión de variantes y su aplicación en el diagnóstico y estudio de los mecanismos moleculares subyacentes de Trastornos del Crecimiento y Desarrollo Puberal

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área de Química Biológica

Franco Gino Brunello

Director de Tesis: Dr. Marcelo Adrián Martí
Co-Director de Tesis: Dr. Rodolfo Rey
Consejero de Estudios: Dr. Javier Santos

Lugar de trabajo: Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Defensa: 21 de Agosto de 2025, Buenos Aires, Argentina.

Resumen

Las Enfermedades Poco Frecuentes (EPoFs) representan un desafío clínico y diagnóstico debido a su diversidad genética y baja prevalencia. Dentro de este grupo, el Hipogonadismo Hipogonadotrópico Congénito (CHH) es un trastorno caracterizado por una deficiencia en la secreción de GnRH, que afecta la maduración sexual del individuo. En esta tesis, desarrollamos un enfoque multidimensional basado en tecnologías de secuenciación masiva (TSM), la revisión sistemática de variantes asociadas a CHH, y el modelado estructural de proteínas, con el objetivo de optimizar el diagnóstico molecular de esta patología y otras EPoFs con base genética.

En el primer capítulo, realizamos una revisión sistemática del conocimiento acumulado en los últimos años sobre variantes en genes asociados a CHH, lo que nos permitió optimizar la aplicación de los criterios del American College of Medical Genetics (ACMG) y mejorar la precisión en la clasificación de variantes de significado incierto (VUS). En el segundo capítulo, analizamos una cohorte de pacientes con CHH mediante secuenciación de exoma completo y paneles dirigidos, logrando un diagnóstico molecular en el 50% de los casos. Además, identificamos casos con posibles modelos de herencia oligogénica, resaltando la complejidad genética de la enfermedad y la necesidad de enfoques analíticos integradores.

En el tercer capítulo, aplicamos herramientas de inteligencia artificial, específicamente AlphaFold2, para el modelado estructural de complejos motivo-dominio afectados por variantes causales de enfermedades mendelianas, particularmente aquellas localizadas en motivos lineales cortos de proteínas (SLiMs). Mediante este enfoque, logramos predecir el impacto funcional de estas variantes en estructuras proteicas clave, proporcionando un marco racional para la interpretación de patogenicidad y la priorización de variantes en el diagnóstico clínico.

Los hallazgos de esta tesis enfatizan la importancia de la integración entre genómica clínica, bioinformática estructural y fisiología en la identificación de variantes causales en CHH y otras EPoFs. La combinación de TSM, criterios refinados de clasificación, modelado proteico basado en IA y conocimiento fisiopatológico establece un paradigma de trabajo más preciso y eficiente para el diagnóstico molecular de enfermedades genéticas, facilitando su implementación en la práctica clínica.

Development of structural bioinformatics techniques for the evaluation and understanding of variants and their application in the diagnosis and study of the underlying molecular mechanisms of Growth and Pubertal Development Disorders

Abstract

Rare Diseases (RDs) pose a clinical and diagnostic challenge due to their genetic diversity and low prevalence. Within this group, Congenital Hypogonadotropic Hypogonadism (CHH) is a disorder characterized by a deficiency in GnRH secretion, affecting an individual's sexual maturation. In this thesis, we developed a multidimensional approach based on massive sequencing technologies (MST), a systematic review of CHH-associated variants, and structural protein modeling to optimize the molecular diagnosis of this condition and other genetically based RDs.

In the first chapter, we analyzed a cohort of CHH patients using whole-exome sequencing and targeted panels, achieving a molecular diagnosis in 50% of cases. Additionally, we identified cases with possible oligogenic inheritance models, highlighting the genetic complexity of the disease and the need for integrative analytical approaches. In the second chapter, we conducted a systematic review of the knowledge accumulated in recent years on variants in CHH-associated genes, which allowed us to optimize the application of the American College of Medical Genetics (ACMG) criteria and improve the accuracy of classifying variants of uncertain significance (VUS).

In the third chapter, we applied artificial intelligence tools, specifically AlphaFold2, for the structural modeling of motif-domain complexes affected by causal variants of Mendelian diseases, particularly those located in short linear motifs (SLiMs) of proteins. Through this approach, we predicted the functional impact of these variants on key protein structures, providing a rational framework for pathogenicity interpretation and variant prioritization in clinical diagnosis.

The findings of this thesis emphasize the importance of integrating clinical genomics, structural bioinformatics, and physiology in identifying causal variants in CHH and other RDs. The combination of MST, refined classification criteria, AI-based protein modeling, and pathophysiological knowledge establishes a more precise and efficient workflow paradigm for the molecular diagnosis of genetic diseases, facilitating its implementation in clinical practice.

Índice

Capítulo 1: Introducción

Sección 1: Enfermedades Poco Frecuentes, Tecnologías de Secuenciación Masiva y Genómica Clínica

Las Enfermedades Poco Frecuentes (EPoFs) -o también llamadas "Enfermedades Raras (ERs)" son condiciones de salud con patrones específicos de signos, síntomas y hallazgos clínicos definidos, que afectan a un escaso número de personas en una población determinada. En Argentina, se considera como EPoFs a aquellas patologías cuya prevalencia en la población es igual o inferior a 1 persona cada 2.000 habitantes, según lo establece la Ley N° 26.689 de "Cuidado integral de la salud de las personas con EPOF y sus familias", promulgada en Junio 2011 y reglamentada en el año 2015 por el Decreto 794/15 [1]. Si bien individualmente las EPoFs afectan pocos individuos, como son miles, en conjunto afectan a un número significativo de la población. Solo en nuestro país, afectan a alrededor de 3,6 millones de habitantes. Aproximadamente, el 70% de las mismas hacen su presentación clínica a edad pediátrica, aunque dependiendo del caso podrían debutar incluso a edad adulta. El 80% de las mismas poseen un origen genético definido, pudiendo implicar a uno o más genes, mientras que otras son causadas por infecciones, alergias, factores degenerativos, proliferativos o teratógenos (como sustancias químicas o radiación), o aún no se conoce su origen. La Organización Mundial de la Salud (OMS) reporta que entre y el 4% y el 8% de la población mundial se ve afectado por enfermedades de origen genético, siendo la mayoría de las mismas causas por trastornos en un único gen - es decir, son de origen monogénico- [2,3] también llamadas mendelianas. La mayoría de las mismas se encuentran reportadas en el catálogo OMIM (por sus siglas en inglés, Online Mendelian Inheritance in Man), donde ascienden a un número de más de 8.000 [4]. Cuando en un paciente se identifica el gen y su cambio subyacente (técnicamente llamado variante) responsables de la patología, es que se ha realizado un diagnóstico genético o molecular. La mayoría de las enfermedades mendelianas son de naturaleza crónica y degenerativa, siendo la causa de diversos tipos de discapacidad y, en casos extremos, poniendo en riesgo la vida del paciente de no contar con un adecuado diagnóstico y tratamiento. En este contexto, se pone de manifiesto que lograr un diagnóstico molecular temprano en la vida del paciente resulta de vital importancia a la hora de pensar en acciones terapéuticas que pueden mejorar exponencialmente su desarrollo y calidad de vida. Por otro lado, no debemos olvidar que esta información es absolutamente decisiva para su familia, que en su poder, puede tomar decisiones con respecto a tópicos como la planificación familiar.

En general, las enfermedades genéticas mendelianas se heredan siguiendo los conocidos patrones de herencia: autosómico recesivo (AR), autosómico dominante (AD) o ligado al cromosoma X (XL). Las mismas se originan debido a variantes genéticas que generan efectos deletéreos a nivel de su expresión, el procesamiento del RNAm en el proceso de splicing u otros mecanismos regulatorios, o alteraciones del producto proteico final. Debido al curso que tomó el estudio del genoma y sus genes, la mayoría de variantes patogénicas causales se encuentran localizadas en ADN codificante para proteína o regiones intrónicas flanqueando exones. Menos numerosas son las variantes patogénicas detectadas en regiones intrónicas profundas o secuencias regulatorias. Y a su vez, existen casos más complejos, asociados a fenómenos epigenéticos, como los trastornos de imprinting, que suman complejidad y dificultad al análisis de la transmisión de caracteres.

En la práctica asistencial, solemos coincidir en que se ha arribado a un diagnóstico molecular cuando se ha identificado una o un pequeño conjunto de variantes que se interpretan como causales en un gen asociado al cuadro en estudio y siguiendo una cigosidad acorde al modelo de herencia para dicho gen. Sin embargo, esta tarea dista

mucho de ser lo simple que pudiese aparentar, explicada esta complejidad por factores como la ausencia de algunas relaciones fenotipo-genotipo para algunos cuadros, el hecho que algunos fenotipos pueden originarse por variantes en una diversidad de genes, un gen pudiendo desencadenar una diversidad de fenotipos, o simplemente no contar con las herramientas suficientes para argumentar el rol causal de una determinada variante en la fisiopatogenia de una enfermedad. De la misma manera, no podemos dejar de comentar que la etapa previa al estudio molecular es igualmente compleja, no siendo siempre aparente la etiología genética del cuadro, y donde muchas veces el paciente y su familia transitan un proceso de años, donde consultan con cantidad de profesionales y se realizan un sinfín de procedimientos médicos, muchas veces costosos e innecesarios. Este período desde la manifestación clínica del cuadro hasta su diagnóstico molecular es denominado "Odisea Diagnóstica", y en la Argentina posee una media alrededor de los 5 años [5]. En todo ese período, la familia se encuentra realizando esfuerzos fútiles, se desgasta económica y psicológicamente y pierde tiempo precioso a la hora de aplicar acciones correctivas.

En esta tesis, abordaremos diferentes facetas en lo referido al diagnóstico molecular de pacientes de una EPoF Particular, el Hipogonadismo Hipogonadotrófico Congénito (CHH). Por un lado, realizamos un estudio general del conocimiento acumulado las últimas décadas en lo referido a diagnóstico molecular en este tipo de pacientes, es decir una revisión sistemática de variantes. En segundo lugar, nos enfocamos al estudio de una cohorte de pacientes con sospecha diagnóstica de este cuadro que concurrieron a nuestro hospital en busca de una respuesta. Por último, tomamos un problema biológico como lo es el análisis de variantes residentes en motivos proteicos y analizamos cómo el conocimiento biológico nos puede llevar a mejores diagnósticos moleculares en pacientes que presenten las mismas.

La estrategia llevada adelante por muchos años en el diagnóstico de CHH consistía en iniciar el estudio de los genes más frecuentemente mutados asociados al diagnóstico presuntivo mediante secuenciación por el método de Sanger: si se detectaban las variantes compatibles con ser causales, se confirmaba el diagnóstico molecular; si esto no ocurría, se continuaba con la secuenciación de otros genes asociados hasta poder hallar -o no- las variantes causales de la enfermedad, lo cual en términos de tiempo y dinero resultaba absolutamente ineficiente. El avance de las Tecnologías de Secuenciación Masiva, también conocidas como Secuenciación de Próxima Generación (NGS, por sus siglas en inglés *Next-Generation Sequencing*) permitió analizar simultáneamente una gran cantidad de genes asociados al fenotipo en cuestión, facilitando la realización de este tipo de diagnósticos. Para que esto sea factible de implementar en la práctica clínica nosocomial, es necesario desarrollar toda una serie de protocolos de biología molecular (preparación de muestras), de secuenciación, y de bioinformática (asociados al procesamiento y análisis de los datos).

En la presente tesis se plantea la implementación de NGS en el contexto hospitalario para lograr el diagnóstico molecular preciso de pacientes con diagnóstico clínico de CHH bajo el enfoque de la genómica clínica. Y de la misma manera, trataremos de desarrollar un conjunto de herramientas y enfoques orientados a generar mejores diagnósticos, para estos pacientes como así también para otros aquejados por otras enfermedades.

El camino traslacional del NGS

El 14 de Abril de 2003 se produjo el anuncio público de la culminación exitosa del Proyecto Genoma Humano (PGH), realizado en manera conjunta por el Instituto Nacional de Investigación del Genoma Humano (NHGRI), el Departamento de Energía (DOE) y sus socios del Consorcio Internacional para la Secuenciación del Genoma Humano. Este constituyó un hito en la historia de la Biología y la Genética, si bien muy rápidamente los propios investigadores se percataron de que la complejidad del mismo era mucho mayor de lo que podrían haber imaginado. No obstante, el hecho de contar con un boceto global del genoma humano constituyó un punto de partida de gran valor para el desarrollo venidero de investigaciones y desarrollos asociados[7,8]. Debemos aclarar que hasta este punto, si bien la Genética había realizado progresos formidables utilizando la Secuenciación de Sanger (como identificar relaciones genotipo-fenotipo como BRCA1-cáncer hereditario o CFTR-Fibrosis Quística), aún presentaba limitaciones significativas en lo referido a costos, tiempo, rendimiento y resolución. Su aplicación a escala del genoma y masivamente en la población requería de un salto cualitativo.

Históricamente, la genética había tomado al gen como su objeto de estudio, y dedicó gran parte de sus esfuerzos a establecer relaciones uno-a-uno entre los mismos y sus fenotipos asociados. El surgimiento del NGS como producto colateral de la carrera entre los consorcios público y privado durante el recorrido del PGH, permitió abaratar exponencialmente los costos de la secuenciación, haciendo mucho más económico el costo por nucleótido. Esto abrió muchísimo el horizonte del paradigma de estudio del paciente, donde se pasó de secuenciar por Sanger secuencialmente genes asociados a un fenotipo (e incluso sólo determinadas partes de exones donde se encontraban las variantes conocidas) a un modelo donde se estudiaban simultáneamente a cientos de los mismos y donde se obtenía un panorama relativamente parejo de los mismos a lo largo de su extensión. Eran los inicios de la medicina basada en la información genética particular y personal del paciente, a la cual se le dio nombres como medicina personalizada, o, más aceptada por los profesionales médicos, medicina de precisión [9,10].

A diferencia de la secuenciación tradicional de Sanger, que utiliza un número limitado de reacciones de secuenciación con longitudes de lectura relativamente largas (500–1000 pb) y genera una cantidad moderada de datos, el NGS se basa en la ejecución simultánea de millones de reacciones de secuenciación en paralelo. Estas reacciones se llevan a cabo en volúmenes extremadamente pequeños, producen lecturas de menor longitud (típicamente entre 100 y 250 pb) y generan gigabases (GB) de datos por experimento. Los pasos a seguir para un típico experimento NGS en plataforma Illumina incluye: (i) la fragmentación enzimática del ADN a secuenciar, (ii) la adición de adaptadores específicos en ambos extremos de los fragmentos, (iii) la amplificación clonal mediante la fijación de los fragmentos de ADN, a través de los adaptadores, a una placa de secuenciación o flow cell, (iv) la secuenciación de los fragmentos por síntesis, con detección simultánea de las bases de forma masiva y en paralelo, (v) la adquisición de datos crudos en una captura de imágenes a través de un detector óptico de fluorescencia, y (vi) la conversión de estos datos crudos en secuencias de nucleótidos (es decir, secuencias de las lecturas en archivos FASTQ) [11].

Las lecturas (o “reads”) pueden ensamblarse utilizando herramientas bioinformáticas. Este proceso permite unir progresivamente las lecturas para formar las secuencias más largas que les dieron origen y lograr la cobertura completa de un genoma o de una región específica, en un enfoque conocido como “ensamblado de novo”. Dado que este proceso es largo y computacionalmente muy costoso, en el caso de genomas previamente ensamblados como el humano, los fragmentos generados se mapean y alinean

contra un genoma de referencia, con el objetivo de identificar diferencias con respecto al mismo. Esto acelera sideralmente los tiempos y reduce exponencialmente los costos computacionales. Dado que múltiples copias del genoma del paciente son ingresadas como *input* del experimento, y que en las mismas la fragmentación se realiza al azar, contaremos con muchas lecturas que se solapan en cubrir la misma posición pero que difieren ligeramente en la secuencia específica del genoma de referencia que cubren. Como resultado, una determinada posición nucleotídica se verá representada en algunos centenares de lecturas, cuyo origen, idealmente, es independiente entre ellas. Esta representación en una cantidad de lecturas determinada es lo que normalmente entendemos como profundidad.

El costo de un experimento de NGS es directamente proporcional a la cantidad de nucleótidos trifosfatados reversibles marcados con fluorescencia utilizados, y por lo tanto a la cantidad de bases deseadas secuenciar a una determinada profundidad. Entender esto nos permite adaptar la técnica a las necesidades del proyecto. Por ejemplo, el NGS puede utilizarse para secuenciar genomas completos o limitarse a regiones específicas de interés [12], como las regiones codificantes de los aproximadamente 21.000 genes humanos [13], en lo que se conoce como secuenciación de exoma completo, o a un conjunto reducido de genes individuales, conocido como panel de genes. En aplicaciones diagnósticas, es común recurrir a la secuenciación del exoma completo del paciente [14], ya que este enfoque se centra en secuenciar únicamente el 1-2% del genoma que contiene aproximadamente el 85% de las variantes genéticas conocidas como causantes de enfermedades [15,16].

Aunque la secuenciación de Sanger sigue siendo el estándar de calidad en secuenciación, resulta ineficiente para abordar exomas o genomas completos, e interesantemente, para detectar variantes genéticas presentes en proporciones bajas [17]. La secuenciación selectiva en NGS representa un enfoque útil para obtener la secuencia de un número reducido de genes o de áreas genómicas específicas a altas profundidades. Este proceso puede ser asistido por tecnologías de enriquecimiento, como las desarrolladas por Roche/NimbleGen y RainDance Technologies, entre otras [18,19]. Por tanto, el NGS se posiciona como la opción ideal para la identificación y el diagnóstico de enfermedades genéticas, incluso aquellas que se presentan como mosaicos genéticos [20].

Finalmente, cabe destacar el desarrollo y auge de las tecnologías de secuenciación de tercera generación, que eliminan el paso de amplificación clonal de los fragmentos en la célula de flujo, lo que reduce significativamente la duración y costos del experimento de secuenciación. Estas tecnologías permiten la secuenciación de moléculas únicas (single-molecule sequencing) y se han implementado en plataformas como Pacific Biosciences (PacBio), que emplea detección por fluorescencia sin amplificación previa, y Oxford Nanopore, que utiliza nanoporos para leer la secuencia de la hebra de ADN a medida que esta pasa a través de un poro biológico [21,22,23].

Bioinformática de "nueva generación"

La transición desde la secuenciación Sanger hacia plataformas NGS impulsó el desarrollo de nuevos algoritmos bioinformáticos y métodos de análisis para la correcta manipulación, almacenamiento e interpretación de los datos generados. De hecho, varios autores señalan que la utilidad clínica de las tecnologías de NGS está más limitada por nuestra capacidad para analizar los datos generados que por nuestra capacidad para generarlos [42,43].

El incremento sostenido en la producción de datos ha superado ampliamente los recursos bioinformáticos disponibles, tanto en términos de personal especializado como de infraestructura computacional para almacenarlos y procesarlos. Esto ha convertido al análisis bioinformático en un recurso escaso y altamente valorado. Además, factores inherentes a la tecnología, como la tasa de error (0,1-1%, o incluso más en plataformas de tercera generación), la corta longitud de las lecturas (100-150 pb) y la alta profundidad requerida en la secuenciación, hacen que el análisis de datos de NGS, especialmente en humanos, sea computacionalmente costoso, intensivo y técnicamente complejo. En este contexto, resulta indispensable el desarrollo de herramientas bioinformáticas que permitan aplicar con éxito la secuenciación masiva al diagnóstico molecular en la práctica clínica [44].

A diferencia del proceso de generación de datos, que sigue protocolos estandarizados y estables según la plataforma utilizada, el análisis bioinformático carece de un modelo estándar universal (“gold standard”). La combinación de diferentes herramientas de software y su integración con bases de datos diversas puede producir resultados muy variables. Sin embargo, la última década vio cómo la comunidad científica genómica se alineó homogéneamente detrás de la implementación de las recomendaciones de trabajo vertidas en el Genome Analysis Toolkit, desarrollado por el Broad Institute [45]. La misma incluye recomendaciones y herramientas estandarizadas para el procesamiento y análisis de exomas y genomas completos. En el caso de genomas o exomas humanos, uno de los pasos más críticos y sensibles es la identificación y caracterización del conjunto de variantes presentes en una muestra, ya que este proceso determina en gran medida el éxito de la secuenciación masiva en el diagnóstico molecular [46].

El procesamiento de lecturas generadas por un secuenciador se realiza mediante un pipeline bioinformático que, en términos generales, incluye cuatro etapas básicas: 1) control de calidad, 2) mapeo-alineamiento, 3) llamado de variantes y 4) anotación de las mismas. Primero, se lleva a cabo un control de calidad para controlar que los estándares mínimos de calidad del experimento de secuenciación se hayan cumplido. Luego, las lecturas se mapean contra el genoma humano de referencia utilizado (prestando especial importancia a la versión elegida) y se realizan alineamientos para optimizar el mismo y corregir discrepancias. Finalmente, se identifica el conjunto de las diferencias entre la muestra secuenciada y el genoma de referencia (es decir, las variantes), en un proceso conocido como llamado de variantes. En la etapa final, se agrega información funcional y clínica relevante a las variantes detectadas, integrando datos provenientes de diferentes bases de datos biológicas, en un proceso denominado anotación de variantes.

El concepto de variante

El genoma de un individuo puede representarse como una cadena de caracteres (ATCG) en la que la mayoría de las posiciones (99,9%) se conservan entre todos los miembros de la misma especie. A fines de facilitar la práctica diaria en la tarea genética, comúnmente toda la comunidad científica trabaja en función de un consenso de genoma humano, denominado genoma humano de referencia [47]. En ciertas posiciones, ciertos individuos presentan diferencias en estas letras respecto del genoma de referencia, conocidas como variantes. La mínima diferencia de una sola letra se denomina variante de nucleótido único o SNV (Single Nucleotide Variant, por sus siglas en inglés). Estas SNVs generalmente serán benignas y no ocasionarán efectos perjudiciales en la fisiología del individuo, pero eventualmente, de alterar un producto génico proteico o un mecanismo de regulación como el splicing, tienen mayor probabilidad de influir en la etiopatogenia de una

enfermedad. Estas variantes son menos frecuentes, evidenciando la acción de la selección natural contra los alelos perjudiciales o deletéreos a lo largo de la evolución humana.

Desde la publicación del primer genoma humano de referencia en 2003, uno de los principales intereses de los genetistas ha sido caracterizar la variabilidad del genoma en diferentes poblaciones. El Consorcio Internacional HapMap ("International HapMap Consortium", en inglés) fue pionero en este ámbito, enfocándose en catalogar las variantes más comunes del genoma humano. Este proyecto recopiló variantes con una frecuencia de alelos minoritarios (MAF, del inglés minor allele frequency) de al menos 5% en uno o más grupos étnicos. Para 2008, el catálogo del proyecto HapMap contenía aproximadamente 3.5 millones de variantes, generalmente denominadas polimorfismos.

Posteriormente, con el fin de incluir variantes menos frecuentes y ampliar el conocimiento sobre aquellas asociadas a patologías, fue necesario adoptar NGS. Estas herramientas no solo permitieron identificar variantes asociadas a estudios de asociación genética, sino también mejorar el entendimiento evolutivo y sentar las bases para establecer predisposiciones a enfermedades. En este contexto, surgió el Proyecto 1000 Genomas (1000 Genomes Project), cuyo objetivo principal fue proporcionar una base de datos de referencia de secuencias genómicas y caracterizar variaciones profundas del genoma humano para estudiar la relación entre genotipo y fenotipo [48].

El Proyecto 1000 Genomas concluyó en 2015 tras reconstruir los genomas de aproximadamente 2.500 individuos provenientes de 26 poblaciones diversas, empleando una combinación de secuenciación de genoma completo a baja cobertura, secuenciación de exoma y genotipificación mediante MicroArrays. Como resultado, se lograron caracterizar más de 88 millones de variantes, incluyendo 84,7 millones de polimorfismos de un solo nucleótido (SNPs, del inglés Single Nucleotide Polymorphism), 3,6 millones de inserciones/deleciones cortas (Indels) y 60.000 variantes estructurales.

En años posteriores, diferentes iniciativas analizaron, reorganizaron y armonizaron los datos de numerosos proyectos de secuenciación exómica a gran escala, haciendo estos datos accesibles a la comunidad científica. Entre los recursos más destacados se encuentran los conjuntos de datos del Proyecto 1000 Genomas, el Proyecto de Secuenciación del Exoma (ESP) [49], el Consorcio de Agregación del Exoma (ExAC) [50], y la Base de datos de Agregación del Genoma (gnomAD) [51]. Estos recursos ofrecen información de gran relevancia tanto para investigación clínica como básica.

En particular, la frecuencia y tipo de las variantes es una de las mejores herramientas para modelar perfiles de su potencial efecto deletéreo, tanto a nivel individual como a nivel génico. Por ello, los genetistas clínicos utilizan estos datos para distinguir variantes con potencial patogénico de polimorfismos francamente benignos, inferir la función de genes y variantes (por ejemplo, si un gen es esencial o haploinsuficiente), y realizar análisis de genética poblacional [52]. Estos proyectos publican datos en bruto en archivos de formato denominado VCF (del inglés Variant Call Format que es el resultado del llamado de variantes), cuya interpretación y manipulación requiere avanzados conocimientos bioinformáticos. Sin embargo, herramientas como navegadores del genoma (por ejemplo, UCSC Genome Browser, ExAC y GnomAD) han facilitado enormemente la visualización de estos datos.

Además, estos navegadores permiten realizar análisis a nivel de genes, evaluar variantes individuales (SNPs, InDels, variantes del número de copias), y proporcionar información detallada sobre las variantes, incluyendo sus anotaciones y métricas de calidad. Igualmente, ofrecen la posibilidad de visualizar la ausencia de variación, lo que puede ser indicativo de una falta de cobertura, baja calidad de los datos o regiones filtradas por otras

razones [53]. En definitiva, estas herramientas simplifican el proceso de filtrado y priorización de variantes, proporcionando información esencial para el análisis genómico y la interpretación clínica.

Frecuencias alélicas - Magnitud del efecto: una relación de idas y vueltas

En el año 2009, Manolio trabajó un reconocido gráfico donde logró caracterizar bidimensionalmente a las variantes genéticas en función de dos de sus propiedades más sobresalientes: la frecuencia alélica poblacional y su efecto molecular en algún proceso relativo a la normal fisiología y desarrollo de un organismo [54]. La primera se encuentra representada en el eje X de la Figura 1 y es arbitrariamente clasificada en rangos según los valores de la misma: variantes comunes (MAF ≥ 5%), variantes de baja frecuencia (MAF < 5%), variantes raras (MAF < 0,5%) y variantes muy poco frecuentes (MAF < 0,1%). En el eje Y está representada la magnitud del efecto que estas variantes producen. Esta magnitud suele ser medida a través del Odds Ratio (OR), que compara las Odds o chances (es decir, la razón entre la probabilidad de que ocurra un evento y la probabilidad de que no ocurra) en individuos con el alelo de riesgo frente a aquellos individuos que no lo poseen [55].

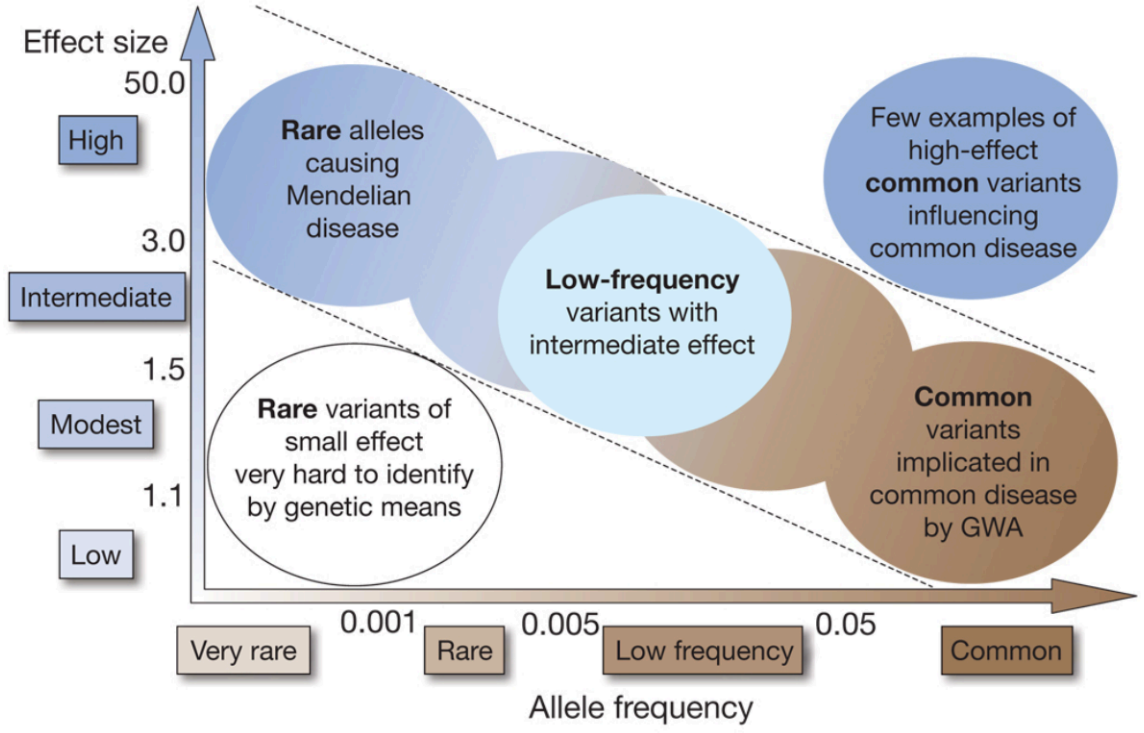

Figura 1. Frecuencia alélica poblacional y magnitud de efecto. Relación entre la magnitud de efecto molecular (OR, eje Y) y las frecuencias alélicas (eje X). Extraída de Manolio et al. [54].

El OR cuantifica la asociación entre una variante genética y una característica que podemos encontrar en un individuo, como el desarrollo de una enfermedad. Un OR = 1 indica que la presencia de la variante no influye en la probabilidad de desarrollar la enfermedad. Un OR > 1 implica una asociación positiva entre la presencia de la variante y el desarrollo de la patología, mientras que un OR < 1 sugiere una asociación negativa, es decir, un efecto protector contra la misma. En los conocidos estudios de asociación GWAS

para enfermedades poligénicas, es normal encontrar valores de OR que oscilan entre 1 y 2, mientras que en las variantes de alto impacto que se heredan con patrón mendeliano y determinan el desarrollo de EPoFs, los OR presentan valores altísimos, técnicamente infinitos.

El principal objetivo de la genómica clínica es identificar variantes causales de diversas patologías o que incrementen el riesgo de padecerlas. En este contexto, se presta especial atención a la región del gráfico delimitada por líneas punteadas en la Figura 1. En esta región, las variantes raras o muy raras con una alta magnitud de efecto (tamaño de efecto superior a 4-5) tienden a ser responsables de enfermedades monogénicas de herencia mendeliana, como se observa en el cuadrante superior izquierdo de la figura. Por otro lado, los alelos comunes, identificados principalmente a través de estudios de asociación genómica (GWAS), tienen generalmente un efecto bajo (OR relativamente cercano a 1). Aunque su efecto es limitado, estas variantes comunes contribuyen al desarrollo de enfermedades complejas y comunes, como la hipertensión o la diabetes, como se muestra en el cuadrante inferior derecho de la figura.

Existen pocos ejemplos de variantes frecuentes con una gran contribución a patologías comunes (cuadrante superior derecho). Aquellas relacionadas a cuadros mendelianos como hemocromatosis o fiebre familiar mediterránea en general han sido harto estudiadas y se ha llegado a ciertos consensos en su interpretación [56]. Otras enfermedades neurodegenerativas, como el Alzheimer, también posicionan aquí a algunas de sus variantes representantes [57]. A su vez, las variantes muy infrecuentes con un bajo efecto son difíciles de identificar, y aún más de adjudicar un rol en la etiopatogenia, ya que justamente el escaso número de individuos dificulta obtener resultados estadísticamente significativos. Como podemos apreciar, el espectro de la variación genética es extraordinariamente complejo, desde los efectos de las variantes individuales hasta las epistasis entre genes y cómo variantes simultáneas en los mismos pueden interaccionar, lo que sugiere que la arquitectura genómica subyacente de muchas enfermedades y rasgos aún dista de ser totalmente comprendido.

En esta tesis, nos centraremos en variantes raras de elevado efecto molecular, ya que, como se mencionó anteriormente, son las que generalmente terminan siendo compatibles con ser causantes de EPoFs siguiendo un típico patrón de herencia mendeliana. Por otro lado, abordaremos superficialmente otras variantes de frecuencias levemente superiores y de efectos moleculares intermedios en el contexto de modelos oligogénicos.

Bases moleculares de las enfermedades de etiología genética

Los organismos vivos experimentan variaciones en su genoma que pueden transmitir a su descendencia, pudiendo resultar deletéreas o beneficiosas para su supervivencia o fitness reproductivo. Las variantes de secuencia generalmente corresponden a variantes de sustitución de un nucleótido por otro o a pequeñas deleciones e inserciones de los mismos. En promedio, cada individuo presenta entre 3 y 5 millones de variantes de secuencia en su genoma. La mayoría de estas se localizan en regiones intergénicas, repetitivas o no-codificantes en un sentido general, no teniendo en la mayoría de los casos un impacto directo en la salud. Sin embargo, las variantes ubicadas dentro de secuencia codificante en un gen, o en regiones regulatorias de los mismos, pueden tener un impacto significativo, al afectar su función o expresión por medio de la enorme cantidad de mecanismos fisiológicos que se conocen.

El término “enfermedad molecular”, introducido a mitad de siglo pasado por Pauling, describe aquellas patologías cuya causa principal radica en la alteración —heredada o adquirida— de uno o varios genes, afectando su función y/o expresión [59]. Aunque la mayoría de las enfermedades monogénicas conocidas se originan por variantes en la región codificante del ADN, existen excepciones notables que han valido enorme reconocimiento a sus descubridores, como enfermedades causadas por variantes en genes de ARN no-codificante (ncRNA), incluyendo a los microRNAs (miRNAs), long non-coding RNAs (lncRNAs), circular RNA (circRNAs) y small nucleolar RNA (snoRNAs). Entender las enfermedades de etiología genética a nivel molecular es clave, pues constituye la base para el desarrollo de terapias racionales.

En esta tesis, el enfoque estará dirigido principalmente a enfermedades monogénicas causadas por defectos o alteraciones en la región codificante de los genes asociados a CHH, mientras que abordaremos superficialmente el universo oligogénico a fines de plantear posibles epistasis entre genes, siempre analizando variantes en secuencia codificante.

Distintos tipos de variantes y sus efectos moleculares

Según el dogma central de la biología molecular, el ADN es transcripto en un ARN mensajero (ARNm), que luego es procesado hasta alcanzar su forma madura. Posteriormente, este ARNm migra de núcleo a citosol donde es traducido en un polipéptido, que atravesará diversas modificaciones y procesos biológicos hasta convertirse en una proteína funcional madura que desempeñe su correspondiente rol fisiológico. Cualquier variante genética posee potencialmente capacidad de interferir alguna de estas etapas, dando lugar a cambios cuantitativos o cualitativos de la proteína producida. Las variantes de secuencia localizadas en genes son en general clasificadas en función del efecto molecular que generan.

Entre las variantes de secuencia, las sustituciones representan uno de los tipos más frecuentemente analizados en casos de genómica clínica. Algunas de estas son sinónimas o silenciosas, lo que significa que, como consecuencia de la degeneración del código genético, el cambio de nucleótido a nivel de codón no altera el aminoácido codificado. Por esta razón, suelen carecer de un impacto perceptible en la proteína resultante, aunque existen casos donde la consecuencia a nivel de proteína enmascara la verdadera etiología a nivel ARN [62]. Sin embargo, otras variantes, conocidas como no-sinónimas o missense, producen un cambio en el aminoácido codificado, lo que puede, o no, modificar la estructura y función de la proteína. En muchos casos, estas alteraciones conducen a alteraciones significativas en su función, por medio de un abanico de posibilidades como su desestabilización, imposibilidad de experimentar modificaciones post-traduccionales o disrupción de interacciones con ligandos, ADN o proteínas.

La pérdida de función de una proteína es un fenómeno que también puede estar relacionado con otras variantes, como aquellas que desplazan el marco de lectura del ARNm (frameshift) o generan un codón de terminación prematuro (nonsense), particularmente cuando se produce decaimiento del mismo por medio del mecanismo de control ARN *Nonsense-Mediated Decay* (NMD) [62]. Asimismo, existen variantes que eliminan un codón de terminación (stop loss) o afectan el inicio de la traducción (start loss). Usualmente, la gravedad de una enfermedad causada por la pérdida de función de una proteína depende directamente de la cantidad de actividad residual que conserve de la

misma. En los casos en que permanece un pequeño porcentaje de función, las consecuencias fenotípicas suelen ser menos severas.

Por otro lado, algunas variantes pueden conferir una ganancia de función. En estos casos, la proteína adquiere actividad extra, ya sea mediante un aumento en su producción –a menudo debido a duplicaciones cromosómicas que incrementan el dosaje génico o a incrementos en la tasa de transcripción– o mediante cambios estructurales que incrementan su actividad normal. Este tipo de variantes son inusuales, ya que la mayoría de las sustituciones aminoacídicas afectan negativamente la estabilidad o funcionalidad de la proteína.

Además de las variantes que afectan la secuencia proteica, aquellas que alteran la expresión génica pueden tener consecuencias igualmente importantes. Por ejemplo, una mutación en la región regulatoria de un gen puede desencadenar su expresión en momentos o tejidos inadecuados, como ocurre con los oncogenes que, al activarse de forma anómala, promueven la proliferación celular descontrolada en el cáncer.

Asimismo, las variantes localizadas en regiones intrónicas pueden interferir con el proceso de splicing, clave en la generación de isoformas proteicas funcionales por medio de eliminación selectiva de intrones y exones. Variantes en las regiones 5’ UTR o 3’ UTR también pueden alterar la interacción con elementos reguladores, como los miRNAs o proteínas de unión al ARN (RBPs). Estas alteraciones pueden afectar la estabilidad, localización, traducción o modulación del ARNm, perturbando vías moleculares y procesos celulares que pueden culminar en enfermedades.

El impacto molecular de una variante de secuencia no depende exclusivamente de su localización en el gen o naturaleza. Para analizar sus efectos, es fundamental considerar factores como la frecuencia alélica poblacional, la conservación evolutiva, las propiedades fisicoquímicas de los aminoácidos implicados y el posible efecto sobre la estructura y función de la proteína. Estos análisis se complementan con estudios funcionales in vitro o in vivo que permiten evaluar de manera más precisa el impacto molecular de la variante.

En el contexto clínico, también es crucial determinar el modelo de herencia para verificar si el genotipo de la variante en el paciente se corresponde con lo esperado para la enfermedad. Esto incluye evaluar su segregación en familiares sanos y afectados, investigar asociaciones previas en la literatura científica y explorar la etiopatogenia involucrada en la patología en cuestión.

La interpretación de variantes: un arte sutil

En la última década, la determinación precisa de la patogenicidad de una variante genética se ha convertido en una tarea central y crítica para alcanzar diagnósticos certeros. Con el objetivo de abordar este desafío, en 2013, el Colegio Americano de Genética Médica y Genómica (ACMG) conformó un grupo de trabajo que incluyó representantes de la Asociación de Patología Molecular (AMP) y del Colegio de Patólogos Americanos (CAP). Su objetivo era el de establecer un marco metodológico para valorar y combinar distintas evidencias asociadas a una variante genética, lo que permite clasificarla en función de su patogenicidad. El resultado fue la emisión de una serie de recomendaciones, de uso voluntario en la práctica asistencial, que en su conjunto son conocidas como "Guías ACMG" [63].

Las recomendaciones de la ACMG/AMP se basan en el análisis de 28 criterios de evidencia de las variantes que abarcan aspectos como la frecuencia alélica en la población, resultados de análisis funcionales, predicciones mediante herramientas *in silico*, análisis de

segregación familiar, reportes en bases de datos curadas y la relación genotipo-fenotipo, entre otros. Cada tipo de evidencia es evaluado con un código alfanumérico que refleja su peso relativo, y estos códigos (también llamados etiquetas) se combinan de manera sistemática para asignar la variante a una de cinco categorías: Patogénica (Pathogenic, P), Probablemente patogénica (Likely pathogenic, LP), de Significado Incierto (Variant of Uncertain Significance, VUS), Probablemente benigna (Likely benign, LB) o Benigna (Benign, B).

La interpretación y clasificación de variantes según las guías ACMG/AMP requiere una comprensión interdisciplinaria que incluye conocimientos en medicina, genética, bioquímica y bioinformática. Esto se debe a la necesidad de analizar datos genómicos desde múltiples perspectivas y en el contexto de la enfermedad estudiada. Eso ya de por sí marcaba una realidad de que diferentes profesionales de la salud con diferente grado y perfil de formación podían interpretar la misma variante de manera diversa. Y por otro, muchos de los criterios que las Guías ACMG planteaban en sus recomendaciones de 2015 se encontraban enunciados de una manera un poco ambigua, lo que hacía que empezara a proliferar una reproducibilidad disminuída, agravada por el etiquetado agresivo de variantes que se “percibían” causales. Con el objetivo de perfeccionar y refinar la asignación de criterios, ClinGen (Clinical Genome Resource), financiado por el Instituto Nacional de Salud (NIH), inició, a través de un conjunto de grupos disciplinarios, una cruzada en pos de lograr una mayor reproducibilidad en la implementación de criterios acompañado de un enfoque mucho más conservador en el etiquetado [65]. A su vez, se busca crear un recurso centralizado que defina la relevancia clínica de genes y variantes para determinadas patologías, lo cual implica un esfuerzo extra en la curación de la relación genotipo-fenotipo, la cual se realiza por medio de grupos de expertos [66].

Estos refinamientos en la implementación de criterios abarcan aquellos relacionados a la frecuencia alélica poblacional [67,68], al análisis del efecto de molecular de variantes de pérdida de función de alto impacto [69] especificando aquellas de splicing [70], predicción computacional de efecto molecular [71], análisis de variantes de novo [72] o heredadas en trans con otras variantes para modelo de herencia autosómico recesivo [73], y reportes previos en bases de datos biológicas [74]. Cada uno de estos trabajos busca adjudicar el grado justo de evidencia a cada criterio, con el fin de prevenir uno de los peores riesgos de la genómica clínica, sobreestimar la patogenicidad de una variante.

A fines de lograr un manejo y una comunicación más fluida y precisa del grado de confianza de patogenicidad de las variantes de secuencia, se tradujo la misma en una implementación bayesiana que genera una probabilidad (entre 0 y 1) de la patogenicidad de la misma [75]. Como muchos integrantes de la comunidad no se encontraban cómodos con el manejo de probabilidades, también se dio la posibilidad de expresar la misma en un sistema de puntos [76]. Paralelamente a estos avances en variantes de secuencia, se desarrollaron recomendaciones que buscaban en forma análoga realizar la interpretación de variantes del número de copias (CNVs) [77]. De esta manera, la interpretación de variantes se transformó en una práctica cotidiana que implicaba la comprensión de todo un conjunto de estudios subyacentes, que a su vez se actualizan permanentemente y demandan una actualización continua de los analistas genómicos.

Miedo a las Variantes de Significado Incierto (VUS)?

En la genómica clínica, cuando se busca un diagnóstico molecular preciso que explique la etiopatogenia de una enfermedad, a menudo nos enfrentamos al escenario en el

que el análisis genómico identifica una o más variantes de significado incierto (VUS, por sus siglas en inglés, Variant of Uncertain Significance). La mayoría de las variantes pueden clasificarse fácilmente como benignas gracias a su alta frecuencia poblacional. Sin embargo, en la práctica solemos encontrar variantes nunca antes detectadas poblacionalmente o en individuos afectados, ante las cuales debemos iniciar una búsqueda exhaustiva de evidencia que evalúe su potencial patogenicidad, utilizando los criterios previamente establecidos. En muchos casos, estas búsquedas no aportan datos concluyentes que permitan determinar la relevancia clínica de estas variantes poco frecuentes.

Dado que el diagnóstico molecular persigue objetivos como la confirmación o corrección del diagnóstico clínico presuntivo, lo cual concluye con la odisea diagnóstica que atraviesa la familia y le permite hacer planificación familiar, y además en muchos casos permite la selección de un abordaje terapéutico individualizado, el resultado de una VUS suele generar insatisfacción tanto en los médicos como en los pacientes y sus familiares. Párrafo aparte para la incomodidad que genera en los profesionales médicos su comunicación. Por esta razón, gran parte del esfuerzo en el campo de la genómica se centra en determinar si estas variantes finalmente serán interpretadas como benignas o como patogénicas.

Según los criterios del ACMG/AMP enunciados en su publicación de 2015, una variante se clasifica como VUS cuando:

a) La combinación de las evidencias evaluadas no es suficiente para clasificarla como probablemente patogénica o probablemente benigna; o

b) La variante presenta evidencias contradictorias.

Muchas veces nos encontramos con el caso de variantes que cumplen con el requisito mínimo de presentar una frecuencia alélica poblacional baja compatible con la incidencia de la patología, pero donde posteriormente, al analizar su efecto molecular, factores como el tipo de variante (variantes missense a la cabeza), su posición en el gen o proteína, la falta de estudios funcionales y las predicciones bioinformáticas inconsistentes pueden dificultar su clasificación como probablemente patogénica.

En estos casos, el análisis del fenotipo del paciente y la correlación con datos clínicos previos resultan cruciales. Si el fenotipo del paciente, junto con la historia familiar, es altamente específico y coincide con el observado en otros pacientes previamente reportados con variantes en el mismo gen, puede ser adecuado reportar la variante como una VUS dado que constituye una información absolutamente relevante para el paciente y su cuadro. Así mismo, se intentará profundizar en cuestiones como la segregación familiar, y por qué no, hasta podría ser el disparador de la realización de ensayos funcionales que pongan en evidencia el efecto molecular de la variante. A veces estos estudios no son abordables o arrojan resultados de dificultosa interpretación. No obstante, debe indicarse que, aunque formalmente se clasifique como VUS, la evidencia al momento del estudio sugiere una posible patogenicidad y se sugiere realizar un reanálisis de la misma en el futuro tratando de valerse de los posteriores avances en el conocimiento molecular y médico que termine de mover la clasificación de la variante en una u otra dirección.

Este proceso de reevaluación debe ser indicado en el informe y acompañado de un seguimiento adecuado. En todos los casos analizados y descritos en la presente tesis, se recomendó y, en algunos casos, se llevó a cabo la validación de la segregación de las variantes identificadas en los integrantes disponibles del grupo familiar.

Establecer la relación genotipo-fenotipo en EPoFs

El principal objetivo de la genética, tanto en el pasado como en el presente, es establecer la relación entre el genotipo del paciente y el fenotipo que presenta. En el contexto de la genómica clínica, particularmente cuando se analizan variantes en múltiples genes, el desafío radica en identificar cuál de estos genes resulta compatible con un fenotipo que se solapa con el observado en el paciente.

La relación entre el genotipo y el fenotipo suele ser compleja, y esta complejidad tiene diversos orígenes. Por ejemplo, en ciertas enfermedades genéticas, una determinada variante no siempre produce un fenotipo aberrante en todos los individuos que la portan, fenómeno conocido como penetrancia. En otros casos, los individuos con la misma variante presentan una amplia gama de fenotipos que varían en severidad, lo que se conoce como expresividad variable. Mientras que la penetrancia describe la proporción de individuos afectados dentro de un grupo de portadores, la expresividad se refiere a la variabilidad fenotípica entre individuos.

La penetrancia y la expresividad se encuentran influenciadas por el background genético de cada individuo como así también por factores ambientales. Además, la heterogeneidad fenotípica, que se refiere a la asociación de un único gen con múltiples fenotipos clínicos, contribuye a la diversidad de presentaciones clínicas observadas en una enfermedad genética.

Por otra parte, la secuenciación de exoma completo ha revelado en los últimos años niveles significativos de heterogeneidad alélica (variantes distintas en un mismo gen que producen un fenotipo similar), y heterogeneidad genética (variantes en genes diferentes que conducen a un mismo fenotipo clínico), eventos comunes en las EPoFs de origen genético.

Buena parte de lo que se conoce sobre la relación genotipo-fenotipo se ha derivado del estudio de variantes poco comunes que causan fenotipos específicos y patologías de herencia mendeliana [78]. A partir de un diagnóstico clínico tentativo, resulta fundamental evaluar todos los genes potencialmente causales del fenotipo observado, a fines de lograr el mejor rendimiento diagnóstico posible. La cantidad de genes implicados en una patología particular suele depender de los genes y proteínas involucrados en el proceso fisiológico afectado.

Un aspecto crucial en el análisis genotipo-fenotipo es determinar si existen variantes previamente reportadas en el mismo gen que producen fenotipos clínicos similares al observado en el paciente. Para ello, resulta indispensable contar con bases de datos exhaustivas, curadas y confiables que recopilen información sobre genes humanos, enfermedades genéticas asociadas y variantes patogénicas reportadas, respaldadas por bibliografía relevante.

En este contexto, destaca la labor pionera del Dr. Victor A. McKusick, de la Universidad Johns Hopkins, quien desarrolló el catálogo Online Mendelian Inheritance in Man (OMIM) a partir de sus publicaciones sobre Herencia Mendeliana en el Hombre (MIM) [79]. Este recurso, ampliamente accesible, ha promovido la investigación y apoyado la educación en genética humana. Establecer un vínculo causal entre genotipo y fenotipo habilita posteriormente la detección de portadores sanos, el cribado poblacional y el diagnóstico directo. Estos avances contribuyen, además, al conocimiento de la función génica, la regulación genética y los mecanismos biológicos que podrían ser aprovechados para desarrollar nuevas terapias.

Sección 2: Hipogonadismo Hipogonadotrófico Congénito

Una enfermedad endocrinológica

En la presente tesis se desarrollan tanto casos de pacientes que contaban con un diagnóstico clínico de hipogonadismo hipogonadotrófico y que buscaban un diagnóstico molecular, como también una revisión sistemática que buscó valerse del conocimiento disponible al día de la fecha en publicaciones científicas sobre dichos pacientes. Realizar un diagnóstico molecular es mucho más que tildar características de una variante genética, y por eso resulta fundamental comprender la fisiología y la fisiopatología que encierra el cuadro. En varios textos se retrata al CHH como “una enfermedad de la neurona GnRH”. Pues bien, veamos de qué se trata la misma, y para ello, haremos un breve planteo sobre un típico eje endócrino como el que podemos observar en la Figura 2. Dado que abordaremos el estudio de pacientes masculinos, esta introducción la enfocaremos en el estudio de la fisiología reproductiva masculina.

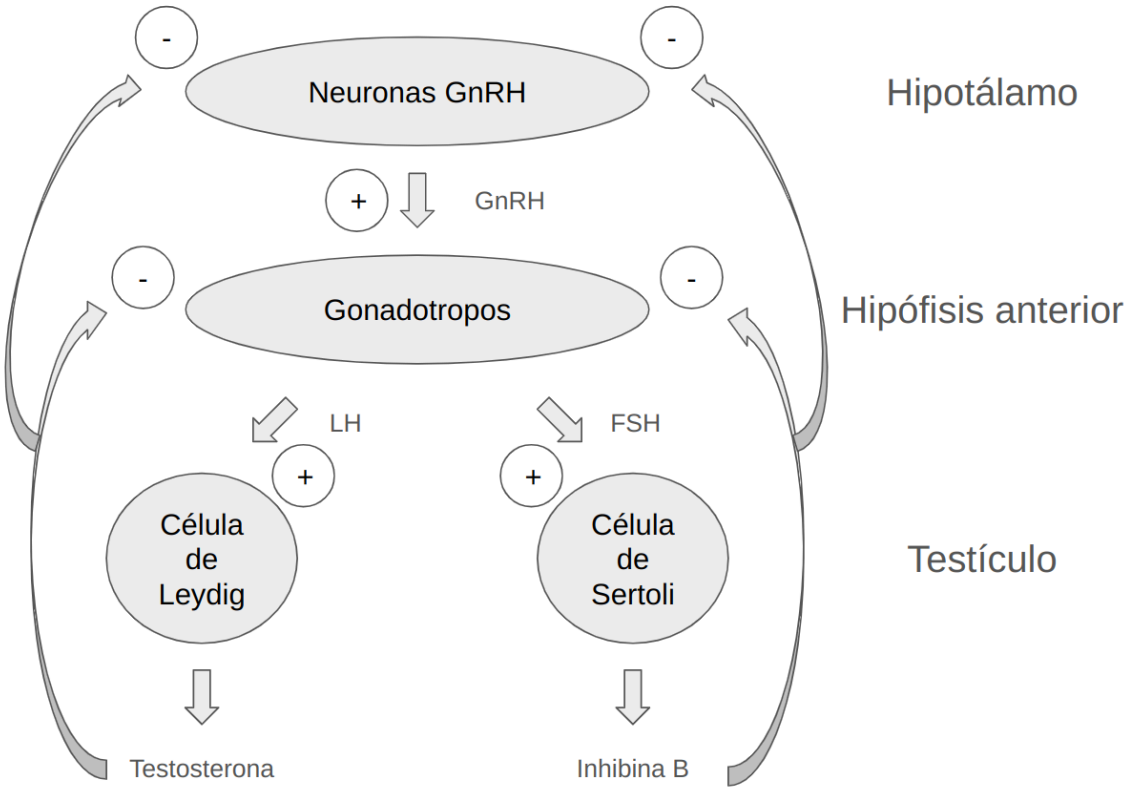

Figura 2. Esquema del eje endócrino hipotálamo-hipófiso-gonadal masculino humano.

El esquema de la Figura 2 podría comprenderse perfectamente si nos situamos en la típica situación de un adulto humano masculino que ha adquirido ya la madurez sexual. El eje se encuentra comandado por un pequeño grupo de células, alrededor de 1000, situadas en el área preóptica hipotalámica. Son las neuronas GnRH, que como su nombre lo indica, son capaces de sintetizar y secretar la hormona liberadora de gonadotrofinas (GnRH, por sus siglas en inglés). Estas neuronas reciben toda una diversidad de aferencias nerviosas (como el hipotálamo en general), las cuales integran y terminan modulando su frecuencia y amplitud de liberación de la mencionada hormona. La misma es liberada por las

prolongaciones axonales neuronales a nivel de la eminencia media en el contexto del sistema porta (sistema sanguíneo de doble capilarización) hipotálamo-hipofisario. De esta manera será que la GnRH podrá llegar vía sanguínea a la hipófisis anterior o adenohipófisis donde encuentra a sus células diana, los gonadotropos, que cuentan con los receptores pertinentes (GnRHR) en su membrana para reconocerla. Al ser el mismo un receptor de la familia de los siete dominios transmembrana acoplado a proteína Gq, la interacción con su ligando tiene como resultado final la generación de un pulso de calcio intracelular que desencadena la síntesis y liberación de las hormonas LH y FSH. Las mismas viajan por circulación sistémica hasta encontrar sus dianas en testículo. En el intersticio testicular encontramos a las células de Leydig, que responden a la LH con la síntesis y liberación de testosterona, responsable de los caracteres sexuales secundarios, como así también de promover la espermatogénesis y espermiogénesis. Por otro lado, la FSH actúa a nivel de las células de Sertoli, responsable del mantenimiento de la arquitectura del túbulo seminífero y de la nutrición y soporte de la línea germinal, los espermatogonios, de los cuales finalmente se derivarán los espermatozoides. La estimulación de la célula de Sertoli tiene también como consecuencia la liberación a circulación de Inhibina B, que junto a la testosterona, cierren el eje en un típico ejemplo de mecanismo de retroalimentación negativa a fines de modular la funcionalidad del mismo.

El análisis de este pequeño esquema ya puede brindarnos informaciones de muchísima utilidad. Por ejemplo, podemos hipotetizar sobre los escenarios donde algún fallo se ha producido a nivel del sistema nervioso central (SNC) o a nivel gonadal. Si algún problema en el desarrollo o funcional se ha producido a nivel del SNC, entonces veremos valores disminuídos de testosterona e inhibina B como así también de las gonadotropinas LH y FSH que no encuentran el estímulo para ser sintetizadas y secretadas. Es el caso del hipogonadismo hipogonadotrópico. Por el contrario, si se hubiera producido un fallo gonadal, entonces la testosterona y la inhibina B se encontrarán disminuídas en sangre pero la intención compensatoria del SNC, que no percibe la retroalimentación negativa dada por la testosterona y la Inhibina B, tendrá como consecuencia que observemos valores elevados de gonadotropinas en sangre. Estamos hablando del caso del hipogonadismo hipergonadotrófico.

Un buen disparador para abordar el estudio de la patología que nos compete, es el de entender cuál es la presentación típica de los pacientes que realizan consultas a su médico clínico o pediatra. En general, se trata de adolescentes que llegan a edades tales donde se dan cuenta que su desarrollo y maduración sexual no acompaña la de sus pares, y por lo tanto comienza a sospecharse de la presencia de un cuadro médicamente relevante. Esta demora en el inicio del desarrollo sexual la conocemos como pubertad retrasada (DP, por sus siglas en inglés). La misma es definida formalmente como la falta de signos puberales (telarca en mujeres y volumen testicular ≥ 4 ml en hombres) a una edad 2 - 2,5 desviaciones estándar por encima de la media poblacional, tradicionalmente fijada en 13 años para las mujeres y en 14 años para los hombres. También puede corresponderse con una progresión alterada respecto de los nomogramas puberales. La DP afecta aproximadamente al 2% de los individuos en edad puberal.

No todas las DP presentan la misma etiología. Como podemos ver en la Figura 3, y acompañando lo previamente planteado, una falla a nivel gonadal, de origen genético o no, puede desencadenar un hipogonadismo hipergonadotrófico. En caso que las gonadotrofinas se encuentren disminuídas, se abre un abanico de posibilidades, lo que pone en relevancia el valor de una buena anamnesis médica. La DP puede darse en el contexto de diversas causas orgánicas como pueden ser tratamientos crónicos con corticoides, o regímenes

transfusionales crónicos, en cuyo caso habremos encontrado la etiología. Por otro lado, diversas situaciones como el ejercicio físico excesivo o los trastornos alimentarios pueden desencadenar los denominados hipogonadismos funcionales. Junto con los mismos, con cierto grado de emparentamiento, podemos encontrar a las denominadas pubertades retrasadas auto-limitadas o retrasos constitucionales del crecimiento y la pubertad (CDGP). Las mismas cuentan con un importante componente genético como se refleja en el hecho de que existe una transmisibilidad familiar en la edad de inicio de la pubertad siguiendo un patrón dominante.

Por último, nos encontramos con los casos que más nos interesan en esta tesis, aquellos pacientes que presentan características clínicas desde el nacimiento como micropene o criptorquidia, muchas veces trastornos en la olfacción, e incluso en ocasiones fenotipo no-gonadal no-olfatorio como sordera, paladar hendido, anomalías esqueléticas, entre otras. Estos casos son claros candidatos para su estudio por NGS pues se correlacionan en alto grado con la presencia de variantes genéticas de alto impacto marginal en la génesis o función de la neurona GnRH. Se trata de pacientes donde la terapia de inducción de pubertad adquiere gran relevancia.

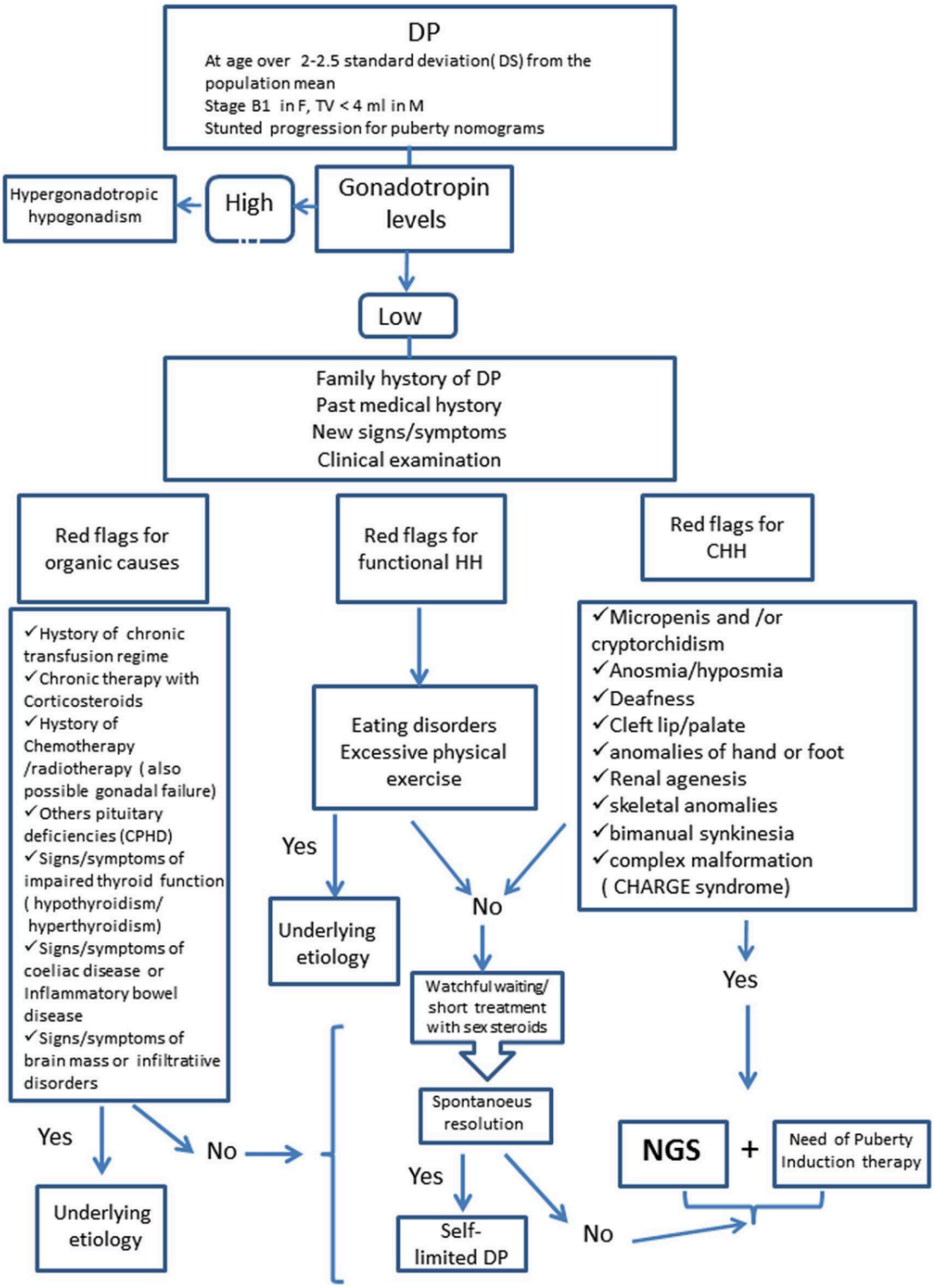

Figura 3. Algoritmo diagnóstico para pubertad retrasada (DP). Extraída de Festa et al. [86]

El hallazgo de una variante diagnóstica para este tipo de pacientes implica un diagnóstico molecular que no sólo es relevante al manejo terapéutico y seguimiento del paciente, sino que además permite un mejor asesoramiento de la familia en su planificación familiar, tanto del paciente como de sus padres. La gran pregunta es dónde vamos a encontrar las mismas. Pues bien, la respuesta es clara: iremos a buscarlas a aquellos genes implicados en el proceso de génesis y funcionamiento de las neuronas GnRH. Para ello es fundamental que tengamos en mente cómo es este proceso.

Como podemos observar en la Figura 4, existen múltiples procesos necesarios para que un individuo termine presentando un eje hipotálamo-hipófiso-gonadal sano [87]. Durante la embriogénesis temprana (semana 14), un grupo de neuronas se diferencian en la placoda nasal, pasando a constituir las neuronas GnRH precursoras. Estas deberán iniciar un proceso migratorio que, cruzando la placa cribiforme, las llevará a su destino final en el hipotálamo, área preóptica. No realizarán este recorrido solas, pues utilizarán como andamio los axones de los nervios vomeronasal y olfatorio, grupos neuronales que comparten muchas señales para su correcto desarrollo. Una vez que las neuronas GnRH arriban a su destino final, deben emitir sus axones hacia la eminencia media, donde verterán el decapéptido GnRH a la circulación portal. Por otro lado, la neurona GnRH establecida en el área preóptica debe asentarse, consolidar sinapsis con otros grupos neuronales cercanos que modulan su actividad, asegurando un buen fitness neuronal. Todos estos procesos que mencionamos previamente podrían fallar, y todos ellos, como vemos en la Figura 4, son comandados por genes que pueden potencialmente ser responsables debido a la presencia de variantes genéticas disruptivas.

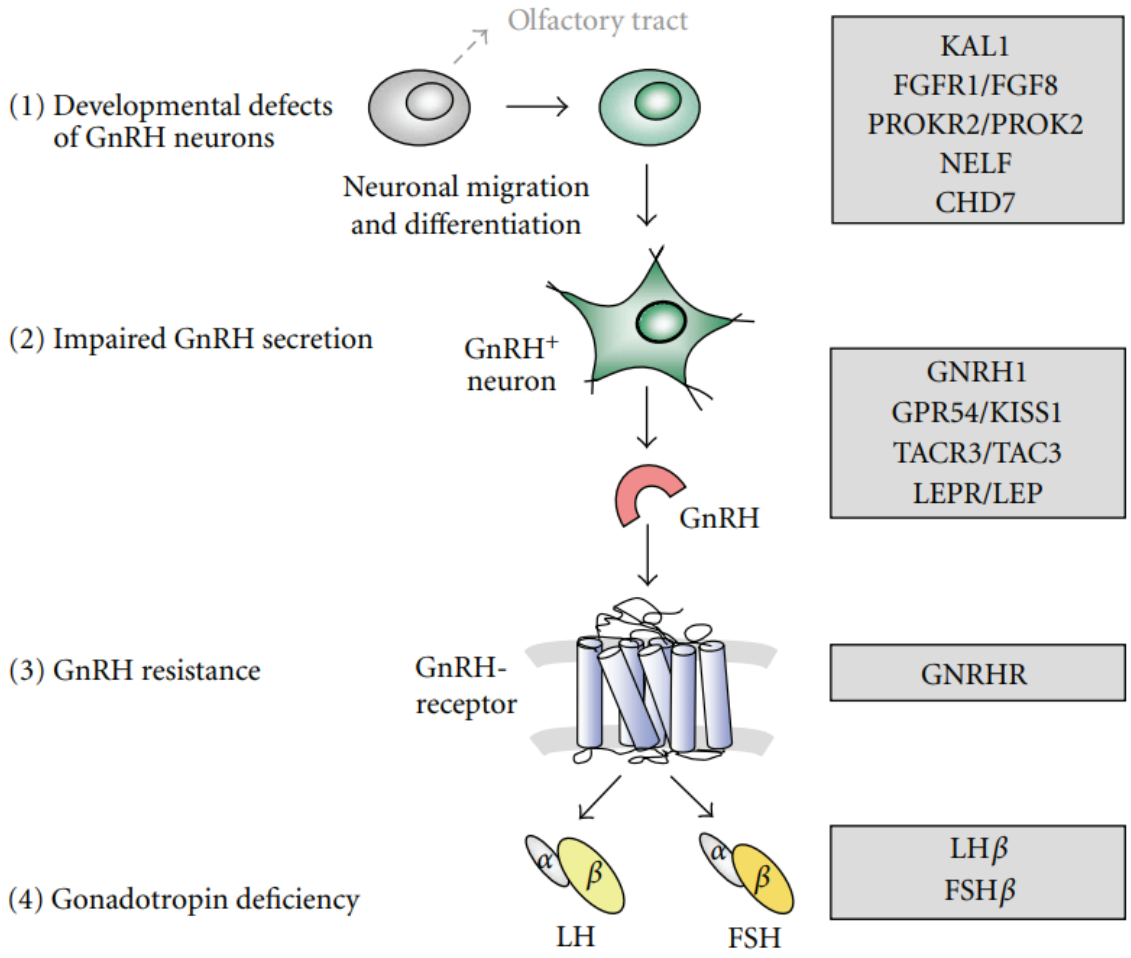

Figura 4. Control genético del desarrollo puberal. Extraído de Karges et al. [87].

Como bien dijimos, existen sistemas compartidos entre las neuronas GnRH y los nervios olfatorios que sirven a los procesos migratorios de ambos. Esto es lo que nos explica que muchos de los genes implicados tanto en la especificación del destino celular de

neurona GnRH como en la migración de las mismas, tengan la posibilidad de dar lugar tanto a un fenotipo gonadal (el hipogonadismo hipogonadotrófico) como olfatorio (anosmia o hiposmia). Esta entidad que combina ambos fenotipos es comúnmente denominada en la comunidad como Síndrome de Kallmann (KS). Es muy importante contar con el dato de trastorno de olfacción (por medio de pruebas olfatométricas o incluso por técnicas de imágenes como RMN), pues su combinación con el fenotipo gonadal eleva marcadamente la especificidad del cuadro, dado que orienta la sospecha diagnóstica a una etiopatogenia muy particular. Por otro lado, muchos de estos genes se encuentran implicados en la embriogénesis en otros procesos, lo que explica la posibilidad de otros fenotipos no-gonadales no-olfatorios, como las anomalías esqueléticas, sordera, etc. Quizás uno de los casos más icónicos es el de la presencia de un cuadro complejo generado por variantes en el gen *CHD7*, denominado Síndrome CHARGE. El mismo incluye la presencia de coloboma ocular, anomalías cardíacas, atresia de coanas, retraso del crecimiento y desarrollo, malformaciones genitales y anomalías en el oído. Por último, no podemos dejar de analizar el caso de que los procesos (y genes asociados) afectados impliquen el direccionamiento axonal de las proyecciones de la neurona GnRH hacia la eminencia media y la llegada de aferencias nerviosas hacia la neurona GnRH, o la liberación apropiada de una GnRH funcional que impacte a nivel de adenohipófisis produciendo una apropiada liberación de gonadotropinas funcionales. En estos casos, como la diferenciación y migración no fueron afectadas, veremos predominantemente fenotipo gonadal, es decir, un hipogonadismo hipogonadotrófico congénito normósmico (nCHH).

Habiendo planteado el desarrollo del eje, y aquellos procesos implicados con sus respectivos genes protagonistas, podemos abordar mejor el porqué de determinados signos como así también los beneficios de diagnósticos tempranos en la implementación de terapias de reemplazo. Para ello debemos comprender, al menos a modo ilustrativo, la progresión de actividad que sufre este eje a lo largo de la vida. Contrario a lo que el público general pudiera inferir, el eje hipotálamo-hipófiso-gonadal sufre tres activaciones a lo largo de la vida, que podríamos plantear cómo tres pubertades [88]. La Figura 5 nos ayudará a entenderlo.

La primera se da durante el período fetal, donde para la semana 6 ya se detecta testosterona generada por las células de Leydig precursoras gracias al estímulo de la hCG placentaria. Esta testosterona promueve la diferenciación de los conductos de Wolff en epidídimo, conductos deferentes y vesículas seminales y la virilización de los genitales externos. Luego de la semana 15 serán la GnRH y la LH fetales quienes comanden la producción de testosterona, aunque hacia el final de la gestación esto se irá apagando como consecuencia de la inhibición del eje por los estrógenos placentarios. En los fetos masculinos, esta testosterona es fundamental para dos procesos claves: el descenso testicular a las bolsas escrotales y el desarrollo peneano. El hecho de encontrar en un recién nacido criptorquidia y micropene podría hablarnos de una falla en estos procesos, constituyendo motivos de sospecha de un CHH. Por otro lado, la FSH impacta en las células de Sertoli haciendo que estas proliferen, y sinteticen y liberen Inhibina B y Hormona Antimulleriana (AMH). Esta última es clave para provocar la regresión de los conductos de Muller antes de la semana 10 (aunque en este período fetal temprano su expresión es desencadenada por SOX9 y regulada positivamente por otros factores como SF1, GATA4 y WT1, pero independientemente de gonadotrofinas).

La segunda activación se registra luego del nacimiento y dura en los hombres unos seis meses. Al nacer, se interrumpe la inhibición por los estrógenos placentarios lo cual genera una gran síntesis y liberación de GnRH en el hipotálamo y, en consecuencia, de LH

y FSH en la adenohipófisis. La primera genera una gran producción de testosterona que contribuye a finalizar el descenso testicular y/o mantener a los testículos en localización escrotal. Por su parte, la FSH genera proliferación de las células de Sertoli (aumentando el volumen del testículo) provocando la liberación de Inhibina B y AMH. Estas dos se consolidan en este momento como marcadores bioquímicos fundamentales para evaluar el funcionamiento del eje y la correcta prosecución de esta mini-pubertad postnatal. La proliferación de las células de Sertoli en esta etapa se cree fundamental para asegurar una adecuada capacidad reproductiva futura, ya que son estas células las que aseguran la arquitectura del túbulo seminífero y dan soporte a los espermatogonios que en última instancia darán lugar a los espermatozoides. Esto despierta interés a fines terapéuticos, pues una terapia de reemplazo con gonadotrofinas en esta etapa podría mimetizar esta mini-pubertad mejorando las previsiones reproductivas.

La tercera y última activación es, sin lugar a dudas, la más conocida. Luego de que el eje se mantenga quiescente durante la infancia, entre los 9 y 14 años de los individuos masculinos se produce el inicio de la pubertad. Esta se da por el cambio en el balance de un complejo sistema regulatorio a nivel hipotalámico conformado por diferentes sistemas (kisspeptina, neurokinina, dinorfina) que genera como consecuencia el inicio de pulsos de GnRH. El primer signo bioquímico de la pubertad es la aparición de los pulsos de LH, inicialmente nocturnos, mientras que el primer signo clínico es el aumento del volumen testicular por encima de los 4 ml. Este se da como consecuencia del estímulo de la FSH sobre las células de Sertoli promoviendo su división mitótica. Recordemos que cada célula de Sertoli da soporte trófico y nutrición a aproximadamente 20 células germinales. Por su parte, la LH generará una gran liberación de testosterona desde las células de Leydig, que además de ser necesaria para desencadenar la espermatogénesis, también será determinante de los caracteres sexuales secundarios. Esto, desde el punto de vista terapéutico, nos explica que si un individuo masculino no inició la pubertad en su debido tiempo, se puede optar por administrar una terapia de reemplazo con testosterona exógena, aunque la administración de gonadotrofinas representa la primera opción a la hora de lograr una pubertad más fisiológica.

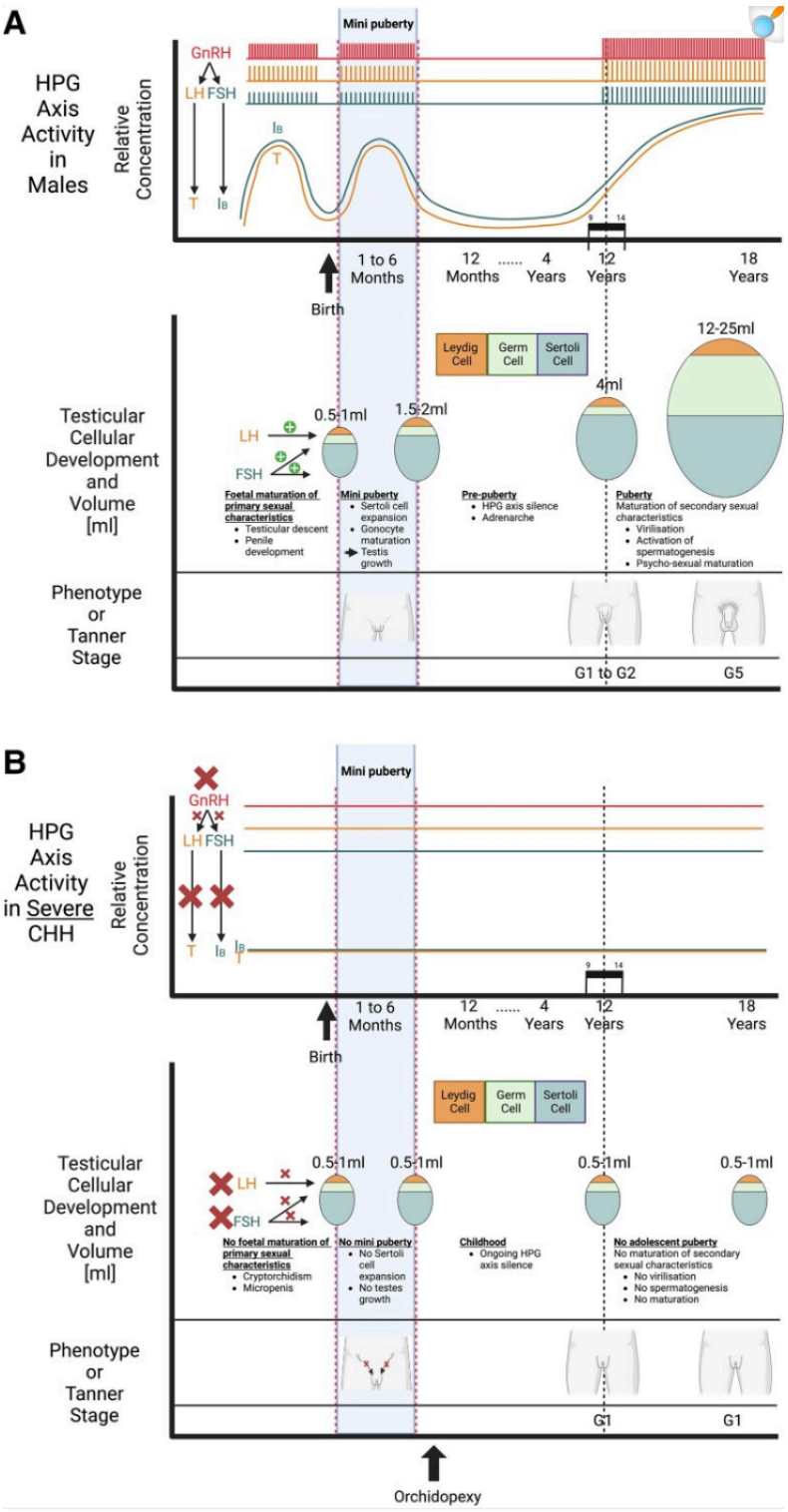

Figura 5. A) Olas de activación del eje hipotálamo-hipófiso-gonadal, indicando las consecuencias en el crecimiento testicular y la composición celular del mismo. B) Olas de activación en el caso de un profundo CHH. Extraído de Rohayem et al. [88].

Volviendo a la genómica clínica, como ya hemos dicho, existe una cantidad de genes implicados en diferentes procesos de la embriogénesis y normal fisiología de la neurona GnRH, en los cuales la presencia de variantes genéticas puede alterar estos

procesos generando patología. Podemos observar la Tabla 1 para ilustrarlos junto a su modelo de herencia.

Gen	Rol	Frecuencia estimada en CHH	KS o nCHH	Modelo de herencia
ANOS1 (KAL1)	Codifica la proteína anosmina-1, vital para la migración de las neuronas olfatoria y GnRH	7,1%	KS	Recesivo ligada al X
CHD7	Codifica la proteína helicasa cromodominio de unión al ADN 7, asociada con Síndrome de CHARGE	16,0%	Ambos	Autosómico dominante
FGF8	Codifica la proteína FGF8, factor de señalización involucrado en la especificación de destino neuronal GnRH y el desarrollo del sistema olfatorio	1,3%	Ambos	Autosómico dominante/ recesivo
FGFR1	Codifica el receptor involucrado en la especificación del destino neuronal GnRH y el desarrollo del sistema olfatorio	9,0%	Ambos	Autosómico dominante
FSHB	Codifica la subunidad β de la proteína FSH	< 1,0%	nCHH	Autosómico recesivo
GNRH1	Codifica la proteína GnRH	1,7%	nCHH	Autosómico recesivo
GNRHR	Codifica el receptor GnRH	4,7%	nCHH	Autosómico recesivo
IL17RD	Codifica la proteína IL17RD, la cual interacciona con FGF8	2,1%	KS	Autosómico recesivo
KISS1R	Codifica el receptor de kisspeptina, siendo la kisspeptina el principal regulador de la activación y secreción de la neurona GnRH	2,0%	nCHH	Autosómico recesivo
KLB	Codifica la proteína Klotho-β, co-receptor obligado para FGF21	4,0%	Ambos	Autosómico dominante
LEP	Codifica la hormona leptina, la cual regula la función neuroendocrina y reproductiva	< 1,0%	nCHH	Autosómico recesivo
LEPR	Codifica el receptor de leptina, involucrándose en la función neuroendocrina y reproductiva	< 1,0%	nCHH	Autosómico recesivo
LHB	Codifica la subunidad β de la proteína LH	< 1,0%	nCHH	Autosómico recesivo
NDNF	Codifica el factor neurotrófico derivado de neuronas, involucrado en la supervivencia y migración neuronal	1,7%	KS	Autosómico dominante

PLXNA3	Codifica el receptor de plexina, Plexina-A3, involucrado en la transducción de la señal de plexina y en la guía del axón neuronal	2,3%	Ambos	Recesivo ligado al X
PROK2	Codifica la prokineticina 2, involucrada en la migración y regulación de la neurona GnRH y la formación del bulbo olfatorio	2,1%	Ambos	Autosómico recesivo
PROKR2	Codifica el receptor de PROK2, involucrado en la migración y regulación de la neurona GnRH y la formación del bulbo olfatorio	5,2%	Ambos	Autosómico recesivo
SEMA3F	Codifica la semaforina-3F, involucrada en la topografía del mapa olfatorio	4,6%	Ambos	Autosómico dominante
SOX10	Codifica la proteína de la caja Y de la región determinante del sexo 10, reguladora del desarrollo de la cresta neural	1,5%	nCHH	Autosómico dominante
TAC3	Codifica la neurokinina B, un péptido involucrado en la regulación de la liberación de GnRH	1,0%	nCHH	Autosómico recesivo
TACR3	Codifica el receptor para neurokinina B, un péptido involucrado en la regulación de la liberación de GnRH	2,6%	nCHH	Autosómico recesivo
WDR11	Codifica WDR11, involucrado en la expresión génica de la vía hedgehog y en la producción de GnRH	0,9%	Ambos	Autosómico dominante

Tabla 1. Genes tradicionalmente estudiados en CHH. Extraído y traducido de Rohayem et al. [88].

Muchos de estos genes han sido estudiados ya por varias décadas y numerosos casos de pacientes con CHH han sido reportados para cada uno en literatura. Durante muchos años la estrategia de estudio fue por medio de secuenciación de Sanger, exón por exón, mientras que el advenimiento del NGS determinó un cambio de paradigma otorgando la posibilidad de estudiar simultáneamente al conjunto involucrado. Si bien pocos han utilizado WGS, muchos grupos se han decantado por WES o paneles customizados, como es nuestro caso.

En cuanto al rendimiento diagnóstico, una de las cohortes de pacientes más importante ha sido la de Cassatella et al. [89]. En la misma, se registró una tasa de éxito del 50% en pacientes con un diagnóstico clínico de CHH, levemente superior en pacientes con KS en comparación con pacientes con nCHH. En la mayoría de los casos, el diagnóstico molecular implicó el hallazgo de variantes mono o bialélicas en un solo gen. Sin embargo, se produjeron hallazgos relevantes de combinaciones de dos o más variantes en dos o más genes. Tomando en consideración los registros de penetrancia, expresividad variable y comienzo de la patología, estos resultados fundamentan el surgimiento de la hipótesis de que la etiopatogenia del CHH podría estar asociada a acciones sinérgicas entre genes que presenten epistasis, donde las variantes en los mismos realizan sus contribuciones marginales al cuadro. Es la denominada oligogenicidad, la cual se registró en un 15% de los casos estudiados, y que desarrollaremos a continuación.

La oligogenicidad como objeto de estudio

En muchos cursos de genética, uno puede tomar el temario a desarrollar y suele encontrarse con varias clases iniciales orientadas al estudio de desórdenes monogénicos, en general analizando las EPoFs que generan, y hacia el final del mismo un par de clases sobre riesgo genético, junto a enfermedades poligénicas que cuentan con una prevalencia alta en la población. Desde nuestro laboratorio siempre tuvimos la convicción que ver ambos escenarios como compartimentos estancos representaba un error garrafal, pues privaba al alumno de hacerse preguntas muy valiosas. ¿No será que existe acaso un hilo conductor entre las mismas? Las enfermedades monogénicas mendelianas se dan debido a una (o dos según modelo de herencia) variante en un solo gen que presenta un impacto marginal alto desencadenando la etiopatogenia por sí sola, donde el factor ambiental debe considerarse, pero presenta un aporte menor. En las enfermedades poligénicas la integración del efecto de un conjunto de variantes de bajo impacto marginal localizadas en un conjunto definido de genes son los responsables de la etiopatogenia de un cuadro como la hipertensión arterial, donde además hay que ponderar el factor ambiental.

Pues bien, en la oligogenicidad lo que se plantea es la posibilidad de un escenario intermedio, donde la presencia de variantes en dos (o más) genes, que presentan una epistasis, generen una sinergia para que sus efectos marginales moderados se traduzcan en el desencadenamiento de la etiopatogenia. Analizaremos principalmente el caso de variantes en dos genes (patrón digénico), el más sencillo de todos, el cual nos será sumamente útil para analizar los modelos de herencia asociados, aspecto común a todos los cuadros de marcado origen genético.

En la Figura 6 observamos los tres modelos de herencia digénicos que podríamos plantear [90]. El primero corresponde a lo que comúnmente denominamos un “Verdadero digénico”. Lo que observamos es un paciente afectado que presenta variantes heterocigotas en dos genes. Dichas variantes fueron heredadas por separado de sus padres, los cuales no se encuentran afectados. Nótese que es una situación análoga a una enfermedad monogénica mendeliana autosómica recesiva, solo que en este caso los dos alelos se encuentran en genes diferentes.

En el segundo modelo de herencia, podemos observar que el padre del probando ya contaba con una variante en un gen y se encontraba afectado, pero al haber transmitido esta variante a su progenie, y sumado a una nueva variante en otro gen, heredada de la madre no-afectada, determina que el probando se encuentra también afectado, pero que en su caso el inicio de la patología sea más temprano o sus síntomas más severos. Este escenario lo definimos como “Monogénico más modificador”.

Por último, nos encontramos con el caso donde el probando directamente presenta dos enfermedades monogénicas simultáneas, donde el fenotipo resultante emergerá del solapamiento de los fenotipos individuales. Es el caso del denominado “Diagnóstico molecular dual”, y constituye de los casos menos frecuentes.

Un análisis muy interesante realizado en el contexto de la definición de los modelos de herencia es el de evaluar el impacto relativo de las variantes implicadas en los mismos. En el modelo verdadero digénico, en general, ambas variantes presentan un impacto molecular moderado. En el monogénico más modificador, el impacto de la variante líder es alto, mientras que el de la modificadora es moderado. En el diagnóstico molecular dual el impacto molecular de ambas variantes es alto, concordante con la presencia simultánea de dos fenotipos monogénicos. Esto validaba de alguna manera los modelos de herencia que se habían estudiado en función de segregaciones familiares de casos. A su vez, esto dio pie

al diseño de una herramienta de machine-learning que en función de dos variantes *inputs* pudiera predecir con buena confianza el modelo de herencia que las mismas pudieran seguir, considerando características de sus genes, la interconectividad entre ellos en redes génicas y de las variantes mismas.

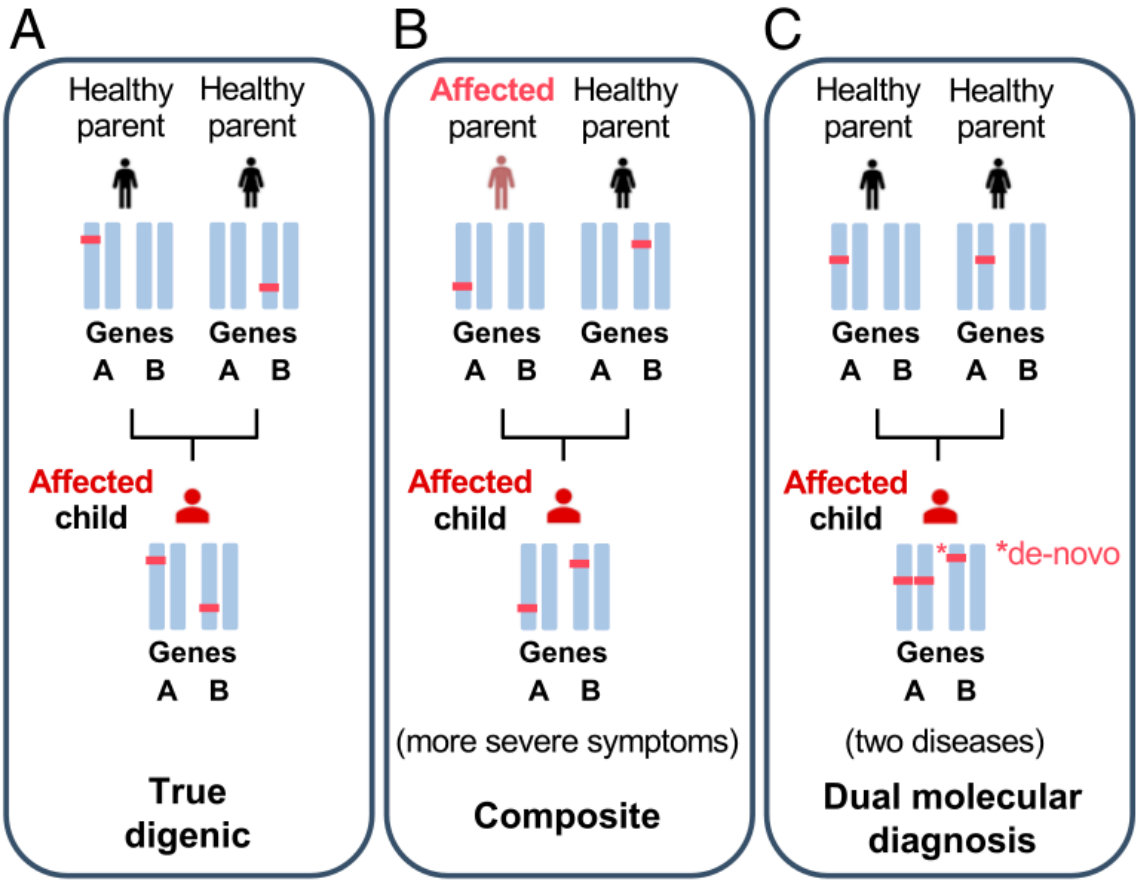

Figura 6. Posibles modelos de herencia oligogénicos. Extraído de Versbraegen et al. [90].

La modelización de los modelos de herencia que pueden seguir los casos oligogénicos dio un marco teórico importante para abordar los mismos. Esto redundó en el hecho de que como consecuencia, muchos grupos de trabajo comenzaron a reportar los mismos con mayor confianza en publicaciones. Estos casos intentaron ser recopilados y depositados en bases de datos, siendo la primera DIDA, que luego evolucionó en OLIDA [91,92]. OLIDA, en cuanto a base de datos, presenta una arquitectura extraordinariamente interesante, donde podemos acceder a entradas de variante, gen, combinación de variantes, combinación de genes, enfermedad o referencia bibliográfica de origen. De esta manera, constituye un recurso muy valioso para evaluar antecedentes de combinaciones que nos proponemos estudiar.

Uno de los agregados de valor más grande que tiene OLIDA en tanto base de datos, es la clasificación de las combinaciones de variantes depositadas en cuatro categorías siguiendo un score numérico de confianza que toma valores enteros entre 0 y 3, siendo el 3 el mayor valor de confianza posible de asignar. Este score surge a partir de un conjunto de criterios y de árboles de decisión que evalúan justamente aspectos de las entradas

previamente mencionadas. Estos criterios se encuentran debidamente explicados tanto en el paper de OLIDA como en la publicación de Papadimitriou et al. [93].

¿Cuáles son los criterios a indagar en el estudio de una combinación digénica/oligogénica? Pues bien, dicha evidencia se ha dividido en una primera instancia en dos tipos. Por un lado encontramos la denominada “evidencia genética”, que profundiza en dos aspectos claves. El primero es el estudio de la presencia tanto de las variantes como de la combinación de las mismas en cohortes de individuos convocados con motivo del estudio o en bases de datos de frecuencia alélica poblacional. Esto constituye actualmente un problema, pues si bien la frecuencia alélica poblacional de variantes individuales es trivialmente consultable en bases de datos, la combinación de las mismas no lo es. Pensemos en que la base de datos debería contar con un número de registros igual a la cantidad total de combinaciones de las variantes que registra, que ya de por sí es un número elevadísimo (es decir, pasar de N registros a (NxN)/2 registros), constituyendo un desafío para indexar y realizar búsquedas. El segundo aspecto que aborda la evidencia genética es la segregación familiar de las variantes con el fenotipo en el pedigree familiar. Recordemos que muchas veces disponer de todos los individuos relacionados en primer y segundo grado con el paciente es extremadamente dificultoso. Sin embargo, cuando los tenemos a disposición, brindan una información clave para establecer una coherencia en la aparición del fenotipo correspondiente. El segundo tipo de evidencia que se aborda es la denominada “Evidencia funcional”. Aquí nuevamente realizaremos una separación de criterios. Por un lado la evidencia in vivo/in vitro, que representa la realización de ensayos funcionales. Sin embargo tenemos que tener en claro las dimensiones que este trabajo conlleva, puesto que deben estudiarse tanto los efectos a nivel génico como a nivel de las variantes, y ambos tanto a nivel individual como a nivel de sus combinaciones. Dada la complejidad, recursos e infraestructura que estos ensayos funcionales involucran, los ensayos para combinaciones de variantes individuales no abundan, pero el lado positivo es que, si un grupo ya ha estudiado la combinación de genes, este es un dato que puede extrapolarse a una nueva combinación de variantes. El segundo criterio abordado por la evidencia funcional es la evidencia in silico, que se vale de herramientas bioinformáticas como predictores de patogenicidad de variantes (a nivel combinación de variantes) o bases públicas de interacciones proteína-proteína, vías de señalización o procesos biológicos, coexpresión, localización celular, etc. (a nivel combinación de genes). Obsérvese cómo fuimos introduciendo en la ecuación a todo un abanico de criterios que en su conjunto se orientan a ponderar la confianza con que una combinación de variantes sea causal del cuadro en estudio o no, al mejor estilo de las recomendaciones de la ACMG, que vale aclararlo, no posee recomendaciones para combinaciones oligogénicas vigentes al día de la fecha. Algo importante de aclarar es que aún no existen recomendaciones orientadas a cuándo informar o no una combinación oligogénica, y probablemente falte algún tiempo aún para llegar a su introducción firme en la tarea asistencial.

Para finalizar, debemos plantear una situación de vida real: si ya disponemos seleccionada una combinación oligogénica para estudiar, probablemente podamos seguir los pasos en la determinación del score de confianza para la misma tal como lo define ORVAL. Sin embargo, cuando recibo el VCF de mi paciente y no encuentro una variante que explique el cuadro en forma monogénica mendeliana, ¿cómo hago para discernir la combinación adecuada para proseguir el estudio? Recordemos que en muchos casos estamos hablando de variantes con impactos sumamente moderados pero que en sinergia logran desencadenar el cuadro. Se hace necesaria una herramienta de priorización de combinaciones, y ese vacío busca llenarlo VarCoPP [94].

VarCoPP es un algoritmo de machine learning (específicamente, un algoritmo de random forest) que fue entrenado a partir de aquellas combinaciones de alta confianza de OLIDA (controles positivos) y combinaciones al azar de variantes en el Proyecto 1000 Genomas (controles negativos). El algoritmo toma en cuenta una diversidad de características a nivel de variantes (predictores de patogenicidad de variantes), genes (scores de sensibilidad a la haploinsuficiencia, a desarrollar patología por diversos modelos de herencia, presión de selección medida por la relación entre variantes sinónimas y no-sinónimas) y combinaciones de genes (diversas bases de datos de redes génicas). Esta herramienta toma como *input* un VCF, arma todas las combinaciones digénicas posibles y las scorea en función de este algoritmo, siendo dicho score un valor entre 0 y 1, donde 1 representa la certeza de efecto deletéreo de la combinación. Uno puede comenzar entonces a analizar las combinaciones que pasen un determinado umbral de patogenicidad ya establecido en su validación. Estos valores de corte se calculan para valores de confianza del 50, 99 y 99.9%. Recordemos que en la genómica en general, y en este caso puntual también, siempre nos encontraremos con más casos negativos (no deletéreos, benignos) que positivos, por lo cual no debemos confiar ciegamente en métricas de exactitud desprendidas de la validación y creer que son algoritmos casi infalibles. Repetimos, esta es una herramienta de priorización, no de interpretación y diagnóstico.

Sección 3: Motivos Lineales Cortos: un problema más general

Los Motivos Lineales Cortos (SLiMs, por sus siglas en inglés) son elementos funcionales clave en la fisiología celular, consistentes en secuencias cortas de aminoácidos (típicamente entre los 3 y 12 aminoácidos de longitud) frecuentemente localizados en regiones intrínsecamente desordenadas de las proteínas o en bucles flexibles expuestos. Estas características permiten que los SLiMs interactúen con sus socios de unión (dominios proteicos) de manera reversible y transitoria, desempeñando un papel esencial en la estabilidad y localización subcelular de las proteínas, así como en la formación de complejos proteicos dinámicos que regulan múltiples procesos celulares [95]. Además, los SLiMs controlan modificaciones postraduccionales al facilitar el reconocimiento de sustratos por enzimas como fosfatasas y quinasas, integrando señales que regulan la toma de decisiones celulares. También son determinantes en la localización subcelular de proteínas, lo que organiza y define la función de los orgánulos.

A pesar de su relevancia, la identificación precisa de SLiMs sigue siendo un desafío biológico importante. Normalmente se encuentran definidos por una expresión regular que establece los residuos permitidos en cada una de sus posiciones. Se estima que el proteoma humano contiene más de un millón de estos motivos [96], pero su relevancia fisiológica es difícil de determinar sin confirmación experimental. Tradicionalmente, los SLiMs se han caracterizado mediante enfoques de biología celular y biofísica, lo que presenta limitaciones para su estudio en modelos in vivo cuando forman parte de ensamblajes multiproteicos complejos. Recientemente, se han desarrollado herramientas bioinformáticas como el Eukaryotic Lineal Motif (ELM), y muchas otras más, para predecir SLiMs [97]. Estas herramientas utilizan filtros lógicos basados en conservación evolutiva, estructura secundaria y localización subcelular para mejorar la precisión de sus predicciones. Sin embargo, la especificidad sigue siendo baja debido al alto número de falsos positivos, originados por la naturaleza degenerada de los motivos, la inmensidad del proteoma humano y la falta de suficientes instancias previas validadas experimentalmente.

La relación entre las mutaciones en SLiMs y diversas enfermedades humanas pone de manifiesto su relevancia clínica. Por ejemplo, la interrupción del motivo basado en tirosina del receptor de lipoproteínas de baja densidad (LDLR) conduce a hipercolesterolemia familiar [98], mientras que la alteración del motivo de unión PDZ del síndrome de Usher tipo 1G afecta la mecanotransducción en las células ciliadas cocleares [99]. De esta manera, variantes localizadas en SLiMs pueden constituir la base molecular de la etiopatogenia que da lugar a un cuadro monogénico mendeliano en el contexto de una EPoF. Pensemos por un momento que el desafío para un analista de variantes será doble, pues primero tendrá que percatarse que su variante compromete a un SLiM funcional y en segundo lugar entender porque la variación puede ser deletérea a su función. En este sentido, se precisa un enfoque que no solo contribuya a la identificación de motivos funcionales, sino que también contribuya a la clasificación de variantes en un marco clínico.

Una hipótesis interesante es si pueden combinarse datos de variantes de secuencia con análisis estructural para mejorar la predicción de SLiMs funcionales [100]. Existen estructuras cristalográficas de complejos motivo-dominio depositadas en el Protein Data Bank (PDB) para analizar el impacto termodinámico de sustituciones de aminoácidos individuales (SAS) en la estabilidad del complejo, empleando por ejemplo el software FoldX [101, 102]. Pueden utilizarse variantes patogénicas y benignas de ClinVar y gnomAD localizadas en SLiMs funcionales para evaluar patrones de tolerancia a la variación genética [103, 104]. Aquí reside un enorme potencial para avanzar en el diagnóstico molecular y la comprensión de las bases estructurales de las enfermedades relacionadas con SLiMs. Sin embargo, este enfoque se vería limitado a clases de motivos con estructuras cristalográficas representativas depositadas en el PDB.

Con los avances recientes en predicción estructural mediante inteligencia artificial, como AlphaFold2 (AF2), ahora es posible superar estas limitaciones [105]. AF2 permite generar modelos de alta precisión para estructuras de proteínas basándose en alineamientos múltiples de secuencia y contactos residuo-residuo, incluso para pares de interacción motivo-dominio sin estructuras cristalográficas previas. Esto sienta las bases para ampliar significativamente las posibilidades para analizar nuevas clases de SLiMs funcionales.

En la presente tesis, utilizamos las estructuras cristalográficas depositadas en el PDB y luego integramos las capacidades de predicción estructural de AlphaFold2 con datos de variantes genéticas y filtros lógicos para refinar las predicciones de SLiMs funcionales en el proteoma humano. Al construir matrices de tolerancia a sustituciones de aminoácidos para cada motivo, evaluamos el potencial impacto patogénico de variantes missense en cada posición dentro de estas clases de motivos. Estas matrices fueron confeccionadas de manera que constituyan un recurso accesible a la comunidad que realiza interpretación de variantes. Además, este enfoque no solo permitió identificar un mayor número de SLiMs funcionales con alta confianza, sino también descartar una proporción significativa de falsos positivos. Aunque el método actual está limitado a motivos con variantes conocidas y estructuras de complejos disponibles, anticipamos que su aplicación se expandirá significativamente con la creciente disponibilidad de datos genómicos y avances en biología estructural.

Objetivos

El objetivo general de esta tesis fue desarrollar y aplicar en casos reales un protocolo bioinformático para la priorización y el análisis de variantes genéticas derivadas de tecnologías de secuenciación masiva, que nos permitiera llegar al diagnóstico molecular de pacientes con diagnóstico clínico de Hipogonadismo Hipogonadotrófico Congénito (CHH). A partir de la capitalización del conocimiento adquirido, y en contexto más amplio, buscamos además utilizar recientes desarrollos en biología estructural para profundizar el análisis y la predicción de efecto deletéreo de variantes, principalmente aquellas residentes en motivos lineales, para el espectro total de Enfermedades Poco Frecuentes.

Específicamente nos propusimos:

- Exponer los resultados de una revisión sistemática de variantes genéticas relacionadas con HHC realizada sobre datos de bibliografía y mostrar cómo su utilización permite optimizar la priorización e interpretación de variantes genéticas halladas en pacientes.

- Realizar análisis genómicos a un conjunto de individuos que concurrieron al Hospital de Niños Ricardo Gutiérrez, con sintomatología compatible con CHH. Desarrollaremos cinco casos representativos de la mencionada cohorte para ejemplificar el trabajo realizado.

- Integrar los datos provenientes de bases de datos de variantes de interés clínico y frecuencia alélica poblacional junto con estructuras cristalográficas producto de experimentos de cristalografía o modelos generados a partir de inteligencia artificial con el fin de generar predicciones sobre el efecto deletéreo de variantes missense en motivos lineales proteicos aptas para su utilización en el contexto clínico.

Capítulo 2: Materiales y métodos

Revisión sistemática

Este estudio se llevó a cabo siguiendo los principios de las guías PRISMA [113]. Las mismas definen los pasos estandarizados a seguir en pos de mejorar la transparencia, precisión, exhaustividad y frecuencia de los protocolos de revisiones sistemáticas y metanálisis documentados. Para esta revisión sistemática, se seleccionaron estudios publicados como artículos originales según los siguientes criterios:

- Diseños de estudio: Se incluyeron ensayos clínicos prospectivos y retrospectivos, estudios de cohortes, casos y controles, estudios transversales, series de casos y reportes de casos.

- Participantes: Se consideraron estudios que incluían pacientes con diagnóstico clínico de HH que provocaron retraso puberal, en quienes se encontraron variantes de secuencia génica en asociación con el diagnóstico. El diagnóstico de CHH se basó en los criterios definidos por los autores de cada artículo. Se incluyeron casos con CHH aislado o deficiencia combinada de hormonas pituitarias (CPHD). Se excluyeron los pacientes menores de 14 años, dada la imposibilidad de diagnosticar un retraso puberal, y los pacientes con hipogonadismo hipogonadotrófico (HH) de inicio en la edad adulta (>18 años). No se aplicaron restricciones con respecto a la etnicidad. Para ser elegible, se aceptó la asociación de la variante genética con el diagnóstico de CHH según los criterios establecidos por los autores del artículo. No se requirió la confirmación de la patogenicidad de las variantes para la elegibilidad. Se excluyeron los polimorfismos, las variantes sinónimas, las variantes benignas o probablemente benignas, según la clasificación de los autores. También se excluyeron las variantes genéticas sin detalles adecuados del mapeo genómico. La adecuación de los detalles del mapeo genómico fue evaluada de forma independiente por el presente autor y otro colega de investigación, ambos con un perfil de analista genómico, con discrepancias mediadas por un tercer miembro del grupo.

- Contexto: No se aplicaron restricciones en cuanto al tipo de entorno del estudio.

Cumpliendo con las recomendaciones de PRISMA, la búsqueda bibliográfica de artículos originales publicados se realizó de manera independiente por quien escribe y un colega investigador adicional el 5 de octubre de 2022, utilizando PubMed para examinar MEDLINE (National Library of Medicine, Bethesda, MD, EE. UU.). Solo se consideraron artículos completos; los resúmenes de congresos no fueron tomados en cuenta. No se aplicaron restricciones iniciales en cuanto a la fecha. Tras una búsqueda preliminar de términos, se utilizó la siguiente estrategia: "((hypogonadotropic hypogonadism OR Kallmann) AND (sequencing OR mutation OR variant))", limitada a "Humans" y "English". Los títulos y resúmenes de todos los artículos fueron revisados por cuatro integrantes del grupo, tanto aquellos de perfil más clínico como molecular, para identificar aquellos que serían incluidos en la revisión de texto completo. En caso de desacuerdo, se llevó a cabo una discusión de consenso con la participación de otros dos integrantes. Se evaluaron los textos completos y datos complementarios para seleccionar los estudios. Los criterios de exclusión incluyeron artículos en idiomas distintos al inglés, datos clínicos insuficientes para confirmar el

diagnóstico de pubertad ausente o incompleta, y datos insuficientes sobre las variantes génicas asociadas con CHH reportadas en los mismos.

La recolección de datos clínicos se realizó de manera independiente por dos investigadores de perfil clínico, y cualquier desacuerdo se resolvió mediante mediación de un tercer integrante. Los datos genéticos fueron recolectados de manera independiente por tres integrantes de perfil molecular, con mediación de un cuarto en caso de discrepancias. No se intentó contactar a los autores de los artículos originales para obtener información adicional, aunque ocasionalmente se extrajeron datos previamente publicados de citas disponibles. La información se recopiló utilizando un formulario estandarizado, a partir del manuscrito principal y material complementario. En los estudios familiares con varios casos elegibles, solo se incluyó información del caso índice. Los casos reportados en más de un estudio se identificaron mediante el ID de muestra y la información se incluyó solo una vez tras una curación manual.

Se extrajo la siguiente información clínica de los estudios publicados: ID del artículo según PubMed (PMID) e ID del paciente en el artículo, año de publicación, sexo del paciente (masculino, femenino), pubertad espontánea (completa/incompleta/ausente), criptorquidia, micropene y/o microorquidismo (presente/ausente), alteraciones del olfato (presente/ausente), alteraciones en la vía olfatoria mediante resonancia magnética (presente/ausente), CPHD (presente/ausente), compromiso de otros órganos o sistemas (cardiovascular, urinario, neurológico, adrenal, auditivo, visual, manos/pies, cara, dentición, integumentos, otros). En el formulario estandarizado, se creó un ID compuesto para cada paciente utilizando el PMID y el ID del paciente en el artículo, para evitar duplicaciones.

La información genómica de variantes presumiblemente patogénicas o causales según los autores fue extraída de los estudios publicados (la información cruda de la variante de secuencia se registró tal como fuera proporcionada en los mismos). La lista de variantes se ensambló en un archivo de Excel y se realizó un mapeo de las mismas sobre genoma de referencia humano GRCh38 / UCSC hg18 (se anotaron el cromosoma, la posición, la referencia y el alelo alternativo). Primero, las variantes fueron consultadas en la base de datos ClinVar para recopilar coordenadas genómicas, lo que permitió mapear casi la mitad de las variantes. El resto de las coordenadas genómicas fueron recuperadas manualmente utilizando los motores de búsqueda Franklin [114] y VarSome [115]. Las variantes con datos insuficientes para su remapeo fueron eliminadas debido a la incertidumbre de su locus.

La biblioteca de Python vcfpy se utilizó tanto para la lectura como para la escritura de un archivo VCF con las variantes seleccionadas[116]. Para mejorar la precisión del análisis, la anotación y la nomenclatura final de las variantes se basaron en transcritos clínicamente relevantes de cada gen. La lista de transcritos MANE (Matched Annotation from NCBI and EMBL-EBI) Select se obtuvo de Entrez y se incorporó en nuestro script de Python desarrollado internamente [117].

La predicción del efecto se llevó a cabo con SnpEff build 5.1f utilizando datos de gnomAD 3.1.2, ClinVar (2023-04-24), REVEL, SpliceAI, entre otros. Las variantes fueron puntuadas y categorizadas según el nivel de evidencia de su patogenicidad. Se utilizó el método bayesiano de cálculo de probabilidad de patogenicidad de Tavtigian et al. mencionado en la introducción, y dichas probabilidades se agruparon en las siguientes categorías:

< 0,001 = Benigna
0,001 - 0,051 = Probablemente benigna

0,100 - 0,188 = VUS (variante de significado incierto) de baja probabilidad de patogenicidad
0,325 - 0,500 = VUS de media probabilidad de patogenicidad
0,675 - 0,812 = VUS de alta probabilidad de patogenicidad
0,900 - 0,988 = Probablemente patogénica
0,994 - 0,999 = Patogénica

La anotación de las variantes, fue finalmente revisada utilizando todos los documentos de actualización que ClinGen emitió en los últimos años, previamente mencionados en la introducción de esta tesis, para criterios como la frecuencia alélica poblacional, veredictos de predictores bioinformáticos, reportes previos de casos con la variante, etc.

Cálculo de una frecuencia alélica poblacional para utilizar como valor de corte en casos de CHH

Se utilizó el algoritmo de Whiffin tal cual fue descrito en su publicación de 2017 [67]. El mismo realiza un uso inteligente de datos característicos de la patología (modelo de herencia, prevalencia, penetrancia) y del estudio de casos de la misma (máximo porcentaje de casos atribuibles a un solo gen y alelo causal más frecuente en el mismo) a fines de modelar una máxima frecuencia alélica poblacional creíble (MCPAF). En pocas palabras, nuestro objetivo será calcular una frecuencia alélica poblacional (la MCPAF), tal que cualquier otra variante cuya AF se encuentre por encima de la misma, su responsabilidad como causal no será compatible con los datos sobre casos previos de la patología recabados al día de la fecha. La fórmula utilizada para su determinación fue la correspondiente a patologías que siguen un modelo de herencia autosómico recesivo:

$$MCPAF = \sqrt{Pr} \times (MCA+ESMCA) \times \sqrt{(MCG+ESMCG)} \times 1/\sqrt{Pe}$$

Donde:

Pr = Prevalencia máxima de la patología de base genética (en nuestro caso CHH) hallada en bibliografía.

MCG = Máxima contribución genética. Es el máximo porcentaje de casos atribuibles a variantes en un único gen.
ESMCG = Error estándar de la máxima contribución genética. Se calcula como $\sqrt{(MCG)*(1-MCG)/n}$.

MCA = Máxima contribución alélica. Es el máximo porcentaje de casos atribuibles al gen de la MCG que son a su vez atribuibles a un determinado alelo.
ESMCA = Error estándar de la máxima contribución alélica. Se calcula como $\sqrt{(MCA)*(1-MCA)/n}$.

Pe = Penetrancia

Análisis de redes génicas y enriquecimiento en el set de genes de la revisión sistemática de CHH

A fines de explotar la información genética relevada para confeccionar paneles de genes que consideran nuevos aspectos de los mismos, se procedió a utilizar el software Cytoscape, stringApp, librería que posee montada toda la información de la base de datos de interacciones proteína-proteína STRING [118].

Se utilizaron tres bases de datos, todas incluídas en el paquete de stringApp, que ofrece la posibilidad de realizar el análisis de enriquecimiento en un set de genes (GSEA) incluyendo a las mismas. Las bases de datos fueron seleccionadas de manera de considerar un perfil más clínico (DISEASES), fenotípico (Monarch Phenotype) y molecular (GO Biological Process). Las categorías que se utilizaron en el análisis fueron las siguientes:

DISEASES:
Endocrine system disease, Hypogonadotropic hypogonadism, Kallmann syndrome, Gonadal disease, Genetic disease, Pituitary gland disease, Hypopituitarism, Monogenic disease, Disorders of sexual development, Gonadal dysgenesis.

Monarch Phenotype:
Hypogonadotropic hypogonadism, Delayed puberty, Puberty and gonadal disorders, Abnormality of the endocrine system, Decreased testicular size, Abnormal circulating hormone concentration, Abnormality of the hypothalamus-pituitary axis, Abnormality of reproductive system physiology, Micropenis, Anosmia.

GO Biological Process:
Neurogenesis, Axon guidance, Generation of neurons, Forebrain development, Nervous system development, Cell differentiation, Axonogenesis, Chemotaxis, Cell migration, Neuron migration.

Cohorte de pacientes

Selección y criterios de inclusión de pacientes

Los pacientes han concurrido a los consultorios de Endocrinología del Hospital de Niños Ricardo Gutiérrez, algunos por propia voluntad mientras que otros fueron derivados por colegas médicos de otros nosocomios. Los pacientes masculinos que presentaban un retraso puberal fueron estudiados clínicamente e interrogados profundamente en sus antecedentes a fines de descartar otras posibles etiologías no congénitas. Aquellos que presentaban antecedentes al nacer como micropene o criptorquidia, un estadio de Tanner compatible (G1 y VP1, es decir, genitales y vello público prepuberales) y un perfil endocrinológico compatible (LH, FSH, T y AMH bajas) fueron indicados para realizarse un ensayo funcional bioquímico, la prueba de infusión de GnRH, donde se le administra la mencionada hormona y se mide la respuesta en sangre de los niveles de LH y FSH en el tiempo (basal, 15, 30, 45, 60, 120 minutos) [106]. La misma permite orientar el diagnóstico clínico hacia un posible CHH o un CDGP. Además, en varios casos se solicitó estudios de imágenes por resonancia magnética nuclear a fines de observar la preservación de los bulbos olfatorios y otras estructuras relacionadas y algunos pacientes lograron acceder a realizarse una olfatometría. Aquellos individuos que presentaron perfiles compatibles con

CHH, fueron seleccionados para practicarles una extracción de sangre para proceder a la obtención de ADN y la secuenciación genómica.

Consentimiento informado

Todos los casos analizados formaron parte de un protocolo de investigación aprobado por el Comité de Ética en Investigación del Hospital de Niños Ricardo Gutiérrez, con los consentimientos informados correspondientes. El médico responsable del caso explica las características del estudio y toma el consentimiento informado al paciente y/o a sus padres o tutores, quienes consienten de manera informada sobre la realización del estudio.

Preparación de bibliotecas de NGS

El ADN genómico fue extraído de células sanguíneas venosas periféricas utilizando el Kit Gentra Puregene Blood (Qiagen). La cuantificación del ADN se realizó mediante un espectrofotómetro de microvolumen de alto rendimiento, Nanophotometer® NP60 (Implen Inc.), y la concentración de ADN se normalizó a 10 ng/µl usando un fluorómetro Qubit® 3.0 (Invitrogen). La pureza del ADN se evaluó midiendo la relación de absorbancia 260/280 nm; el procesamiento posterior de las muestras de ADN se realizó únicamente si la relación se encontraba entre 1.8 y 2.1. Se realizaron dos métodos diferentes para la preparación de la librería de ADN. En primer lugar, se utilizó el panel de secuenciación TruSight One® (TSO) (Illumina), que cubre 4.813 genes asociados con trastornos genéticos mendelianos conocidos (~12 Mb de contenido genómico). La calidad de la fragmentación del ADN genómico se controló utilizando un sistema capilar Fragment AnalyzerTM (Advanced Analytical). La secuenciación de próxima generación por síntesis con terminadores reversibles fluorescentes de desoxinucleótidos trifosfatados se realizó utilizando un sistema NextSeq 500® (Illumina) en la Unidad de Medicina Traslacional del Hospital de Niños de Buenos Aires (Unidad de Medicina Traslacional, Hospital de Niños Ricardo Gutiérrez, Buenos Aires). Cuando no se pudo priorizar ninguna variante en el estudio TSO, posteriormente se realizó la Secuenciación de Exoma Completo (WES) por 3Billion, Inc. (Seúl, República de Corea). Se capturaron todas las regiones de los exones de todos los genes humanos (~22,000) mediante el Panel de Investigación de Exoma xGen v2 (*Integrated DNA Technologies*, Coralville, Iowa, USA). Las regiones capturadas del genoma fueron secuenciadas con Novaseq 6000 (Illumina, San Diego, CA, USA). Posteriormente, se contrató un servicio profesional de la empresa Twist BioScience (San Francisco, California, Estados Unidos) para realizar un panel customizado que cubría aquellos genes relacionados con CHH como así también otras patologías de relevancia para el nosocomio.

Análisis bioinformático

Nuestro pipeline está basado en las buenas prácticas de Genome Analysis Toolkit (GATK) del Broad Institute [107]. El primer paso del análisis bioinformático consistió en un control de calidad de los archivos FASTQ (que contienen las lecturas) obtenidos tras la secuenciación. En este formato, cada lectura de ADN incluye información de la secuencia de nucleótidos y metadata asociada, incluyendo una métrica de calidad, representada en escala Phred, que indica la precisión de cada base leída. Para evaluar la calidad general de las lecturas, se utilizó el programa FastQC, que permitió analizar parámetros como la

longitud promedio de las lecturas, el contenido de GC, la cantidad de lecturas repetidas, entre otros.

Con las lecturas preprocesadas y de calidad controlada, el siguiente paso fue el mapeo y alineamiento contra el genoma humano de referencia, versión GRCh38. Este proceso permitió ubicar cada lectura en su posición correspondiente respecto del genoma de referencia. El mapeo se realizó mediante el programa BWA basado en el algoritmo de alineamiento Burrows-Wheeler, dando como resultado archivos en formato SAM, los cuales contienen información detallada sobre las lecturas alineadas, como su posición y calidad de mapeo. Para optimizar el almacenamiento y facilitar el análisis posterior, los archivos SAM fueron ordenados por coordenadas y convertidos a su versión binaria comprimida, denominada BAM. Los mismos fueron utilizados paralelamente para realizar el screening de CNVs por medio del software DECoN [108].

Con las lecturas mapeadas, se procedió al llamado de variantes, es decir, a identificar diferencias entre las secuencias de la muestra y el genoma de referencia. Estas variantes, que incluyen polimorfismos de nucleótido único (SNP) e inserciones y deleciones pequeñas (InDels), fueron evaluadas según criterios de calidad, como la profundidad de cobertura y el número de lecturas independientes que respaldan su presencia. En el medio se dan dos procesos importantes, la recalibración de los puntajes de calidad de las bases (BQSR), que permite ajustar las estimaciones iniciales de calidad generadas por el secuenciador, y la recalibración de los puntajes de calidad de las variantes (VQSR), que pondera diversas características de las mismas, como calidad de las bases, profundidad, mapeo, etc. Estos procesos mejoran significativamente la precisión del llamado de variantes. Los resultados fueron almacenados en archivos en formato VCF (del inglés Variant Call Format), que contenían información detallada sobre cada variante, como su posición, tipo y posibles efectos.

Finalmente, las variantes detectadas fueron anotadas utilizando bases de datos biológicas para agregar información relevante que permitiera su filtrado y priorización. Entre las bases de datos utilizadas contamos con gnomAD (frecuencia alélica poblacional), ClinVar (variantes de interés clínica), SnpEff (anotación funcional de genes), OMIM (asociación a cuadros clínicos congénitos), REVEL (metapredictor bioinformático de patogenicidad de variantes missense), SpliceAI (predictor bioinformático de variantes de splicing) [109-111]. Además anotamos las variantes con el software InterVar, que brinda una clasificación automatizada preliminar siguiendo los lineamientos de la ACMG [112]. Este flujo de procesamiento, desde los archivos FASTQ iniciales hasta los archivos VCF finales, se resume esquemáticamente en la Figura 7.

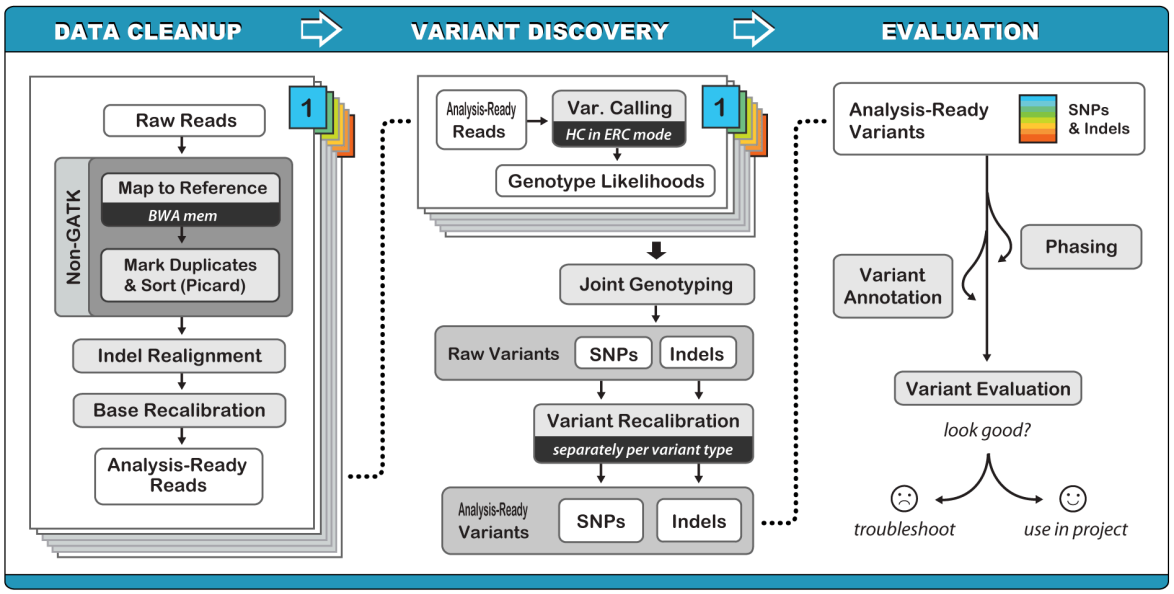

Figura 7. Flujo de trabajo de los datos de secuenciación desde la obtención de los archivos FASTQ del secuenciador hasta la obtención del VCF listo para analizar, tomado de la página de GATK del Broad Institute.

Predicción de patogenicidad de variantes en motivos lineales cortos: MotSASi

Clases de motivos ELM y expresiones regulares

Las clases de motivos (se denomina Clase a cada tipo de motivo definido de acuerdo a una expresión Regular) de la base datos del ELM fueron seleccionadas según los siguientes criterios: debían tener una longitud fija, al menos una instancia validada experimentalmente en humanos (verdadero positivo), y un mínimo de cinco variantes con significancia clínica establecida (patogénicas o benignas). Se excluyeron las clases asociadas con modificaciones postraduccionales. Las clases seleccionadas se subdividieron en aquellas que contaban con una estructura cristalográfica representativa de la interacción con su receptor previamente depositada en el Protein Data Bank (PDB) y aquellas que no. El primer grupo se utilizó en el procedimiento de validación de predicción de estructuras de complejos motivo-dominio mediante AlphaFold2. Las expresiones regulares que definen estas clases de motivos fueron utilizadas según lo especificado en la base de datos ELM.

Identificación de SLiMs en el proteoma humano

El proteoma humano completo se definió como las 20.435 secuencias de proteínas canónicas con estado curado en la base de datos UniProt (es decir, Swiss-Prot) (Septiembre 2024) [119]. Las coincidencias de las expresiones regulares de motivos en el proteoma fueron halladas utilizando un script de Python desarrollado internamente, que escaneó las secuencias completas, permitiendo múltiples coincidencias por secuencia. Las coincidencias se definieron como las ocurrencias de la expresión regular del motivo dentro de la proteína analizada, caracterizadas por la secuencia específica, su posición y su

longitud. El conjunto completo de todas estas coincidencias identificadas en el proteoma humano para una dada clase de motivo lo denominamos el “Set Inicial”.

Identificación de variantes dentro de cada motivo

Las variantes missense clínicamente relevantes involucradas en el análisis de SLiMs fueron extraídas de las bases de datos ClinVar (FullRelease, Septiembre 2024) y gnomAD (v4.0 - Exomas, Noviembre 2023). Solo se consideraron las variantes clasificadas en ClinVar como Patogénicas, Patogénicas/Probablemente Patogénicas, Probablemente Patogénicas, Benignas, Benignas/Probablemente Benignas o Probablemente Benignas (es decir, se descartó del análisis las VUS). Las variantes benignas de cambio de sentido que excedían un umbral predefinido de frecuencia alélica fueron recuperadas de gnomAD. Para los genes con al menos 10 variantes patogénicas, el umbral de frecuencia alélica se determinó en base a la variante patogénica más frecuente reportada en ClinVar. Para los genes con menos de 10 variantes patogénicas, se aplicaron umbrales basados en modelos de herencia autosómica recesiva (AR) y autosómica dominante (AD) (10^-4,1 y 10^-4,28, respectivamente), obtenidos a partir de las distribuciones de frecuencia alélica de variantes patogénicas y benignas en ClinVar. Finalmente, para los genes que aún no se asocian con ninguna enfermedad, se aplicó un umbral fijo de 10^-4,1. Los valores de umbral variaron de 0,7637 a 0,000015, y son los comúnmente utilizados en nuestro grupo para la anotación de variantes, siguiendo los criterios del ACMG [120].

Conjunto de datos de estructuras

Las estructuras cristalográficas asociadas con las instancias de motivos del ELM fueron recuperadas de PDB. Para las clases ELM sin estructuras cristalográficas asociadas, se generaron 10 modelos predictivos utilizando AlphaFold2, con las secuencias de motivos y dominios de instancias humanas validadas experimentalmente (es decir, verdaderos positivos) de la base de datos ELM como entrada. Estos modelos se calcularon utilizando dos parámetros diferentes para el número de ciclos de iteración: 24 (cinco modelos por clase) y 72 (cinco modelos por clase).

Cálculo del cambio de energía libre de Gibbs de estabilidad

El cambio de energía libre de Gibbs de estabilidad (△△G) para cada variante missense fue calculado utilizando el software FoldX en cada estructura cristalográfica representativa del complejo motivo-dominio correspondiente (derivado de PDB o AF2). Los cálculos se realizaron utilizando el comando PositionScan. Los valores resultantes de △△G para cada clase de motivo se presentan como matrices de sustitución. Cuando hubieren múltiples estructuras cristalográficas disponibles para un complejo dado, los valores de la matriz representan el promedio de △△G a través de todas las estructuras.

Conversión de la significancia clínica, frecuencia alélica poblacional y cambio de energía libre de Gibbs en scores de confianza

La significancia clínica de las variantes, según lo informado en ClinVar, se convirtió en un score de confianza, donde cada estrella del estado de revisión aportó 1 punto y cada submission a la base añadió 0,1 puntos extra. Las frecuencias alélicas poblacionales

obtenidas de gnomAD y los valores de cambio de energía libre de estabilidad de FoldX también se transformaron en puntajes de confianza utilizando ecuaciones sigmoideas respectivas. Esto permite que a medida que las variantes de las tres fuentes vayan alcanzando las matrices finales, podamos compararlas entre sí de una manera razonable a través de sus scores de confianza.

Cálculo del score de conservación

Se construyeron alineamientos múltiples de secuencia (MSA) utilizando las secuencias del clúster UniRef50 asociadas con cada proteína que contenía coincidencias de la expresión regular. Los alineamientos se realizaron utilizando el software MAFFT. Posteriormente, estos alineamientos se utilizaron para calcular el score de conservación de la secuencia correspondiente al motivo, el cual fue determinado mediante el algoritmo de Divergencia de Jensen-Shannon [121,122].

Predicción de la estructura secundaria de proteínas

La predicción de la estructura secundaria de las proteínas se llevó a cabo utilizando el software Scratch, proporcionando la secuencia completa como entrada. El resultado consistió en una clasificación por residuo en las categorías de hélice, hebra (también llamada hoja) u "otra" [123].

Cálculo de la exposición de residuos

El área de superficie accesible al solvente (SASA, por sus siglas en inglés) se calculó utilizando la biblioteca FreeSASA. Este cálculo se aplicó a las predicciones de estructuras proteicas generadas por AlphaFold2 (AF2) para el proteoma humano de SwissProt disponible en su sitio web [124].

Términos de la Gene Ontology (GO)

Los términos de la Gene Ontology (GO) se obtuvieron del recurso The Gene Ontology Resource [125].

Visualización de estructuras proteicas y matrices

Las visualizaciones de estructuras proteicas y las matrices de sustitución se generaron utilizando el software Visual Molecular Dynamics (VMD) y la biblioteca Seaborn [126,127].

Análisis y filtrado de Motivos Lineales Cortos (SLiMs)

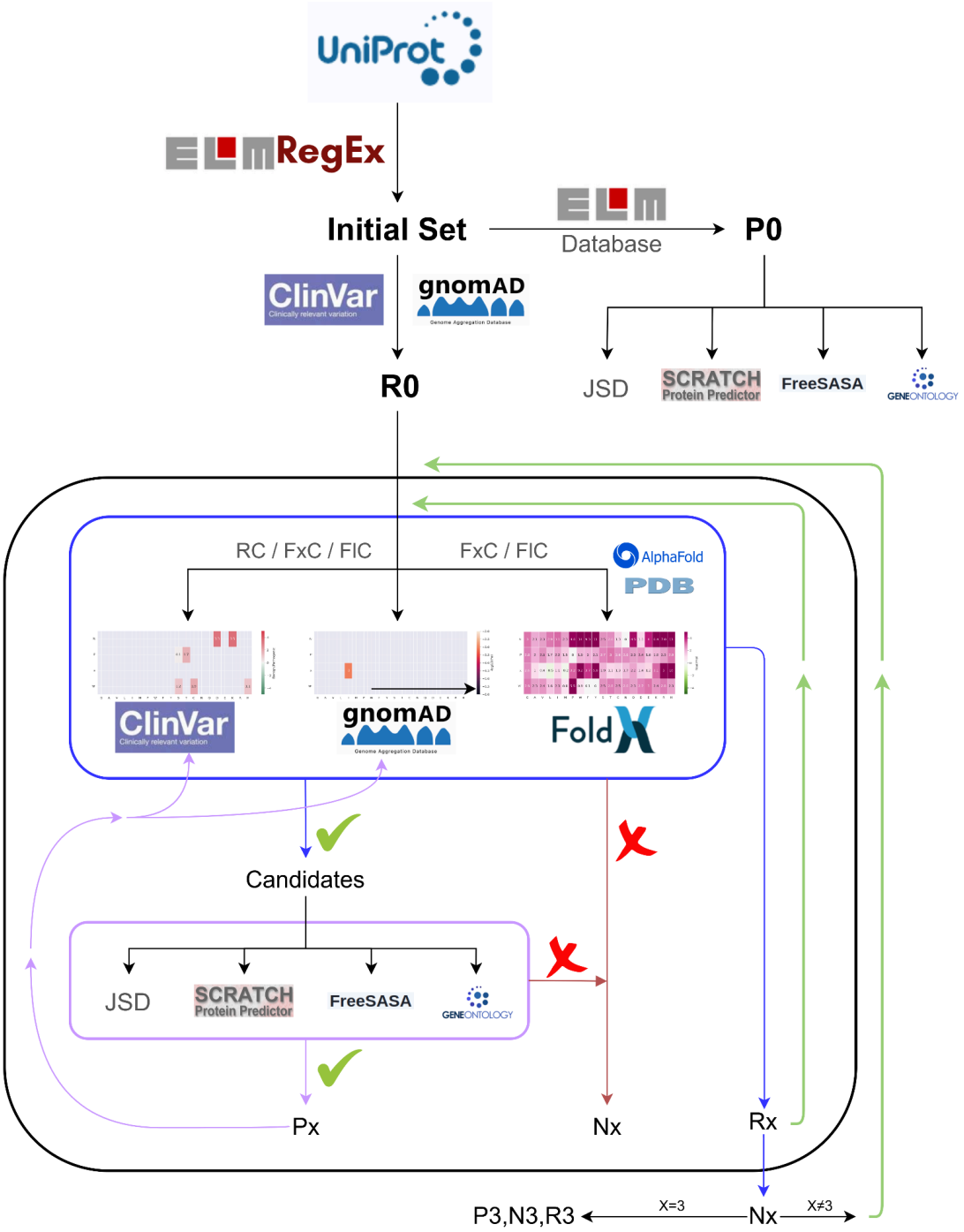

Figura 8: Diagrama de flujo del pipeline de MotSASi, que integra estructuras cristalográficas del PDB y modelos predictivos generados por AlphaFold2.

El conjunto inicial de SLiMs (conocido como “Set Inicial”) se filtró para retener sólo aquellos motivos que contenían al menos una variante registrada en ClinVar o gnomAD, formando con los motivos resultantes el Set 0 (S0). Además, las instancias de motivos confirmadas experimentalmente en la base de datos ELM fueron clasificadas como Set Positivo 0 (P0), representando los verdaderos positivos. Estos motivos fueron analizados en términos de estructura secundaria, exposición al solvente, conservación y términos GO

asociados. A partir de los datos de variantes para cada proteína en P0, se construyeron matrices de significancia clínica, mientras que las matrices de estabilidad estructural (o tolerancia) se generaron aplicando FoldX a las estructuras obtenidas del PDB o de AlphaFold2 (AF2), según correspondiera.

Ciclo de filtrado de tres pasos

Después de construir las matrices, los motivos fueron filtrados mediante un ciclo iterativo de tres pasos:

1. Ciclo de significancia clínica

En este paso, todas las variantes en los motivos de S0 fueron clasificadas como patogénicas o benignas utilizando datos de significancia clínica de ClinVar y los umbrales de frecuencia alélica de gnomAD descritos en métodos. Los motivos que no presentaban contradicciones con el Set Positivo (es decir, que pasaron tanto la evaluación de las matrices de variantes como la de características que ya mencionamos para P0) fueron asignados al Set Positivo 1 (P1).
Los motivos con variantes discordantes fueron asignados al Set Negativo 1 (N1), mientras que los casos no resueltos se clasificaron en el Set Remanente 1 (R1). Las variantes de los motivos en P1 se usaron para refinar las matrices de ClinVar y gnomAD. Este ciclo se iteró hasta que no se pudieron identificar nuevos motivos positivos.

2. Ciclo de análisis estructural

Los motivos en R1 fueron evaluados comparando sus variantes con las matrices de estabilidad obtenidas de los complejos motivo-dominio correspondientes mediante FoldX. En nuestro caso definimos como cambios missense tolerados a aquellos cuyos valores de ΔΔG están por debajo de umbrales predefinidos: 2,1 kCal/mol para estructuras del PDB y 1,4 kCal/mol para modelos de AF2 (las calibraciones de estos umbrales se muestran más adelante). Los motivos cuyas variantes no contradecían la matriz de tolerancia resultante se asignaron al Set Positivo 2 (P2). Los casos discordantes y no resueltos se transfirieron al Set Negativo 2 (N2) y al Set Remanente 2 (R2), respectivamente. Al igual que en el ciclo de significancia clínica, los motivos en P2 se utilizaron para refinar las matrices de variantes de gnomAD y ClinVar. En este ciclo, sólo se analizaron las posiciones de residuos bien definidas presentes en la expresión regular. Las posiciones flexibles (como x o ^P) se analizaron en el último ciclo.

3. Ciclo de posiciones flexibles

Finalmente, los motivos en R2 fueron filtrados evaluando la (in)tolerancia de las sustituciones missense en las posiciones flexibles de las expresiones regulares, utilizando los valores de ΔΔG calculados con FoldX. La metodología reflejó la de los ciclos anteriores y resultó en los conjuntos finales:

- Set Positivo 3 (P3): motivos designados como funcionales con alta confianza.
- Set Negativo 3 (N3): motivos clasificados como no funcionales con igual nivel de confianza.

- Set Remanente 3 (R3): casos no resueltos, siempre confluye a un set sin candidatos pues al usar las matrices completas, ya debería haber analizado todas las variantes en ellos.

Capítulo 3: Resultados

Sección 1: Revisión sistemática.

En esta sección nos abocaremos al estudio de los resultados que arrojó la revisión sistemática de variantes construída en base a la búsqueda bibliográfica realizada en la base de datos PubMed, la cual nos permitió seleccionar una cantidad de artículos científicos relacionados a CHH, de los cuales se extrajo información tanto de pacientes como de las variantes que los mismos portaban.

La revisión sistemática llevada adelante entregó como resultado 352 trabajos científicos que fueron curados, donde se encontraban documentados 1.938 pacientes, de los cuales 1.380 eran de sexo masculino, 412 de sexo femenino y no se pudo determinar el sexo de 147 individuos en función de los datos de bibliografía. La deficiencia hipofisaria era aislada en 1.197 individuos, mientras que era combinada en 39 y no se pudo determinar en 703 de los mismos. Estos individuos en su totalidad sumaban una cantidad de 2.602 variantes genéticas, que de ser deduplicadas representan 1.518 variantes únicas. Las mismas se encuentran localizadas en un total de 143 genes. En la Figura 15a podemos observar como es la distribución global de las variantes en los mismos, mientras que en la Figura 15b podemos observar el gráfico de barras correspondiente a los 41 genes portadores de la mayor cantidad de variantes. Como podemos comenzar a vislumbrar, existe una gran disparidad en la cantidad de variantes identificadas por gen.

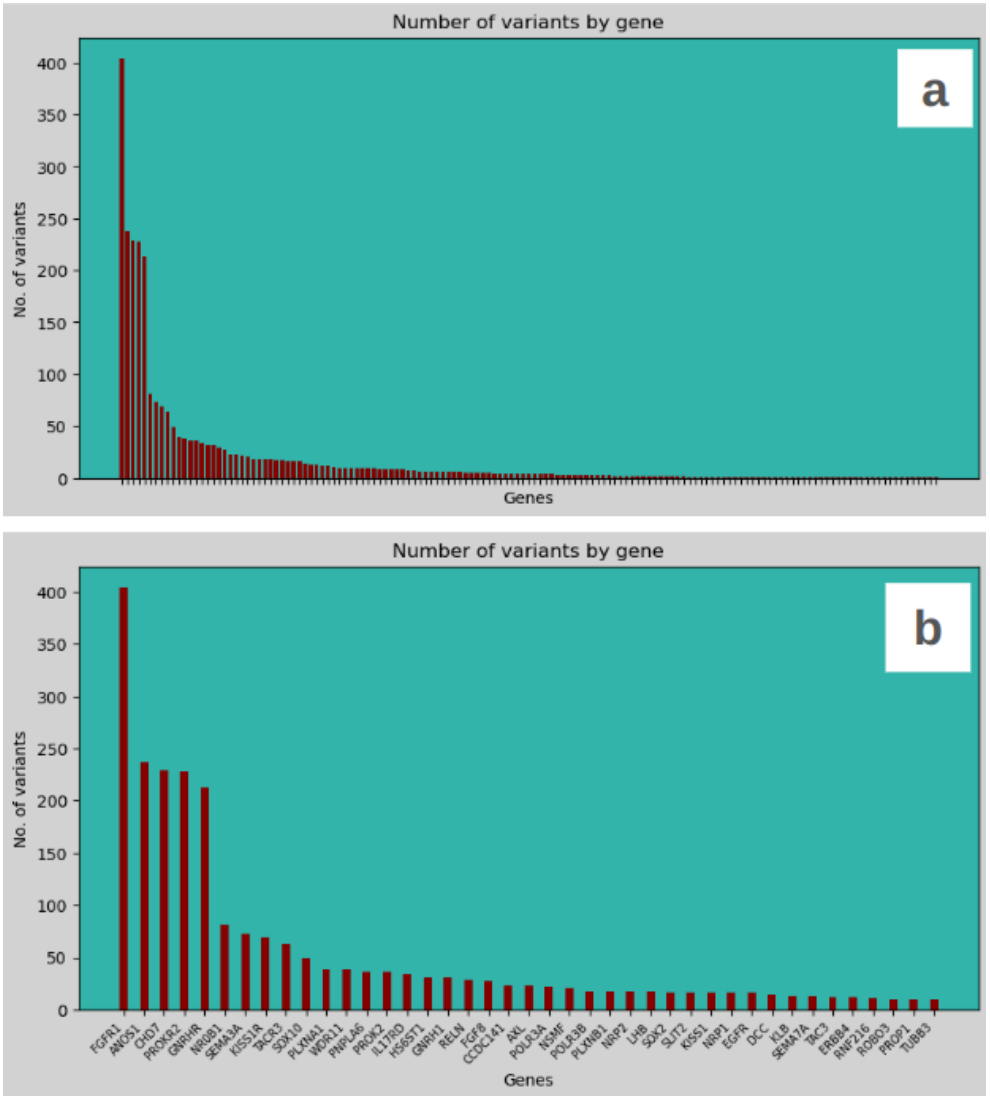

Figura 15. a) Gráfico de barras mostrando la cantidad de variantes localizadas en cada gen, ordenado de manera tal que los genes con mayor cantidad de variantes se posicionan a la izquierda. No se muestran los nombres de los genes en el eje X debido a la enorme cantidad de los mismos. b) Gráfico de barras mostrando la cantidad de variantes localizadas en cada gen para los 41 genes que cuentan con mayor cantidad de variantes.

Una manera alternativa de apreciar esto es observando el histograma de número de variantes por gen (Figura 16), el cual nos muestra que existe una gran cantidad de genes poseedores de un reducido número de variantes (más de 60 genes cuentan con apenas una sola variante reportada), mientras que existe una minoría de genes que cuentan con números de entre 40-420 variantes.

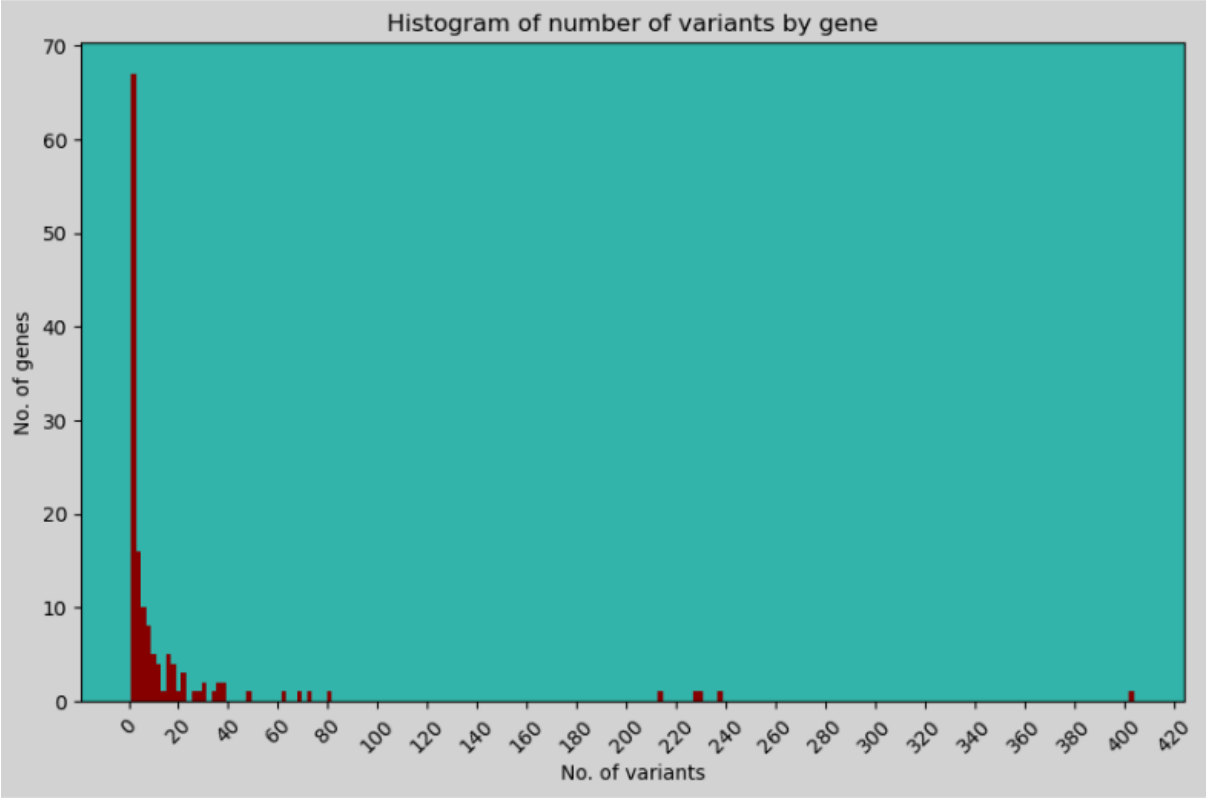

Figura 16. Histograma del número de variantes por gen, se aprecia una distribución con una enorme cantidad de genes dueños de un reducido número de variantes, mientras que pocos genes poseen un número de variantes arriba de 40.

El conjunto de las variantes estudiadas fue anotado con el pipeline desarrollado y utilizado por nuestro laboratorio, descrito en métodos. Algunos de los estudios más intuitivos e iniciales que realizamos fue el de saber si estas variantes ya habían sido previamente reportadas en una base de datos como lo es ClinVar, con cuál clasificación, y por otro lado observar cómo habían sido evaluadas por el software de interpretación automatizada de variantes según las recomendaciones de la ACMG, InterVar. Como podemos ver en la Figura 17a, el 60% de las variantes no se encontraba reportado en ClinVar, mientras que alrededor de un 15% de las mismas contaba con una entrada relacionada a una clasificación patogénica, y alrededor de un 8% con una clasificación benigna. Por otro lado, en la Figura 15b, vemos cómo aproximadamente un 30% de las variantes son clasificadas como P/LP según InterVar, mientras que aparece sólo un 5% con clasificaciones LB/B. Resulta interesante ver cómo InterVar, que toma en cuenta la clasificación de las variantes en ClinVar, adiciona evidencia para lograr más veredictos patogénicos (pasando de 15% a 30%), mientras que al realizarlo sobre las variantes benignas no logra el mismo resultado, sino que se modera a la hora de enviar a las variantes a clasificaciones benignas (pasando de 8% a 5%). De todas maneras, se sigue apreciando la mayoría de variantes adjudicadas en las categorías inciertas.

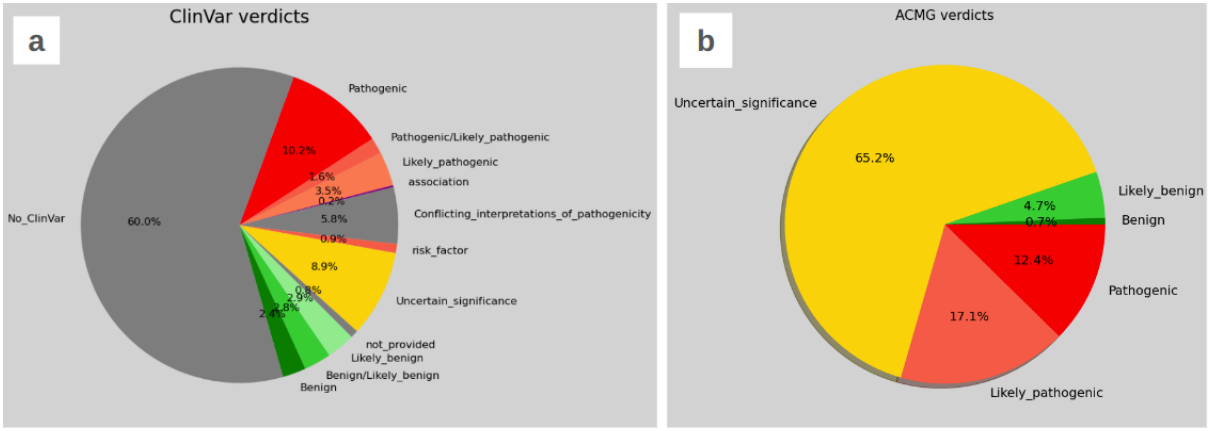

Figura 15. a) Gráfico de torta indicando estado de reporte en ClinVar de las variantes estudiadas. b) Gráfico de torta indicando clasificación de las variantes según recomendaciones de la ACMG implementadas por InterVar.

Un punto importante a considerar desde el inicio en este estudio es el de cómo una heurística, como lo es es sistema de clasificación de patogenicidad de variantes según la ACMG establecido en 2015 por Richards, derivó luego en un enfoque cuantitativo bayesiano. En el mismo, el objetivo es lograr transferir los diversos criterios desarrollados con un enfoque cualitativo, en una serie de factores numéricos que aumentan o disminuyen la probabilidad de base de una variante de ser patogénica (fijado como el 10% en función de la experiencia clínica). Esta técnica nos permite comenzar a pensar a las variantes, y fundamentalmente a su probabilidad de patogenicidad, como una variable ordinal, donde la misma puede tomar una serie de valores entre 0 y 1 (como toda probabilidad). Como podemos ver en la Figura 16, además de incluir los valores de probabilidad a posteriori de patogenicidad en el eje y, ordenamos a las variantes en función de su valor de frecuencia alélica poblacional en la base de datos gnomAD. Como era de esperarse, se puede observar una cierta tendencia de las variantes patogénicas a localizarse a frecuencias más bajas y las benignas, a frecuencias más altas. Una lectura fundamental adicional derivada de este gráfico viene dada por el hecho de que no todas las variantes de significado incierto (VUS) son lo mismo, algunas cuentan con mayor o menor cantidad de evidencia en favor de su patogenicidad.

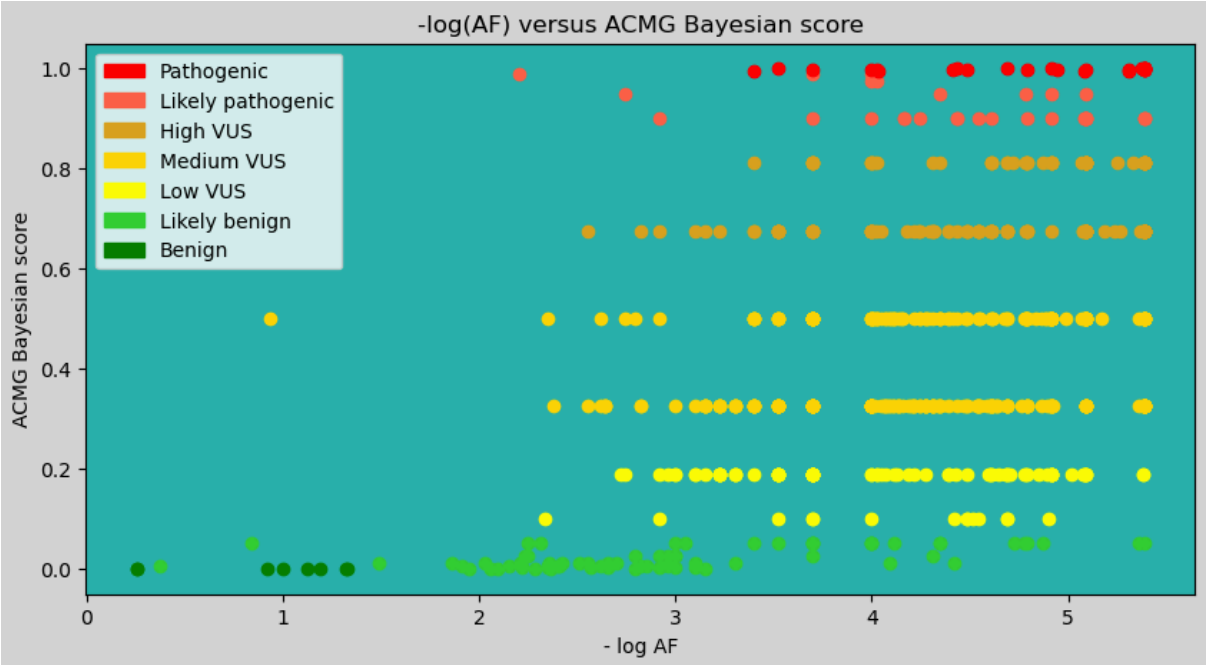

Figura 16: Dotplot de la probabilidad posterior de patogenicidad de las variantes incluídas en la revisión, en función del menos logaritmo de la frecuencia alélica poblacional de las mismas en exomas de gnomAD.

Determinación de un valor de corte frecuencia alélica poblacional para casos de CHH

Siempre que trabajamos en el análisis de casos clínicos estudiados por NGS, uno de los grandes desafíos con los que nos enfrentamos es el de las VUS, que como podemos ver constituyen una mayoría de las variantes, con nada menos que el 65% de representación. En los últimos años muchos esfuerzos se hicieron en mejorar y profundizar el abordaje de las mismas, como así también refinar la asignación de criterios, que en su momento se consideraba demasiado laxa, con el fin de reducir la asignación errónea de clasificaciones patogénicas o benignas a variantes que finalmente no lo son. En este contexto, contar con una estrategia de filtrado más potente reduce considerablemente el número de variantes, el tiempo dedicado al caso en particular, y por lo tanto el costo del estudio.

Uno de los filtros más potentes que poseemos para reducir el universo de variantes en estudio en un caso de genómica clínica, es la frecuencia alélica poblacional (AF) de las mismas. Por esta razón, una de las primeras estrategias de análisis que se busca refinar a la hora de mejorar los protocolos de priorización de grupos especializados como el nuestro es el de donde poner el valor de corte para la AF a fines de enriquecer el subconjunto remanente en variantes causales (aumento de la especificidad) mientras se mantiene a un mínimo la pérdida de sensibilidad. Complementariamente, este valor de corte puede ser utilizado a la hora de decidir entre aplicar etiquetas de baja (PM2) o alta (BS1) frecuencia. Para esto se utilizó el algoritmo pensado por Whiffin, de amplia difusión y aceptación en el rubro. El mismo realiza un uso inteligente de datos característicos de la patología (modelo de herencia, prevalencia, penetrancia) y del estudio de casos de la misma (máximo porcentaje de casos atribuibles a un solo gen y alelo causal más frecuente en el mismo) a fines de modelar una máxima frecuencia alélica poblacional creíble (MCPAF). En pocas palabras, nuestro objetivo será calcular una frecuencia alélica poblacional (la MCPAF) tal que cualquier otra variante cuya AF se encuentre por encima de la misma, su

responsabilidad como causal en forma mendeliana no será compatible con los datos sobre casos previos de la patología recabados al día de la fecha.

Una de las grandes decisiones a tomar, y que es un problema que los grupos asistenciales suelen tener, es el de cómo proceder al enfrentarse a un conjunto de genes causales donde conviven múltiples modelos de herencia. La solución más conservadora pensada por nuestro grupo fue la de considerar aquellos genes asociados a un modelo de herencia autosómico recesivo (AR), entendiendo que siempre alelos de los mismos pueden lograr una circulación ligeramente más elevada a nivel poblacional, comparado con aquellos provenientes de genes asociados a un modelo de herencia autosómico dominante (AD). De esta manera, a la hora de calcular una MCPAF estaremos logrando un valor de corte que ajustará bien para aquellos genes asociados a un modelo de herencia AR, mientras que a su vez también sirva para filtrar (sobradamente) variantes asociadas a modelos de herencia AD o ligados al X (XL).

Más allá de que la revisión sistemática llevada a cabo sólo incluye variantes documentadas como causales del cuadro clínico, es interesante observar en la Figura 17 el histograma de -log(AF) en gnomAD de las variantes incluídas. Para mayor claridad, hemos construido dos histogramas, uno con la totalidad de las variantes, y otro solo con las variantes que presentan una AF no-nula en gnomAD. En el primer histograma, a las variantes ausentes se les asignó un valor de AF correspondiente a la mínima AF registrada en la base de datos de variantes. Esto es imprescindible por dos motivos: por un lado, no es posible calcular el logaritmo de cero, y por otro lado, no incluir las variantes haría subrepresentarlas, lo cual cambiaría la distribución del histograma y no nos dejaría apreciar el panorama general de las variantes en estudio.

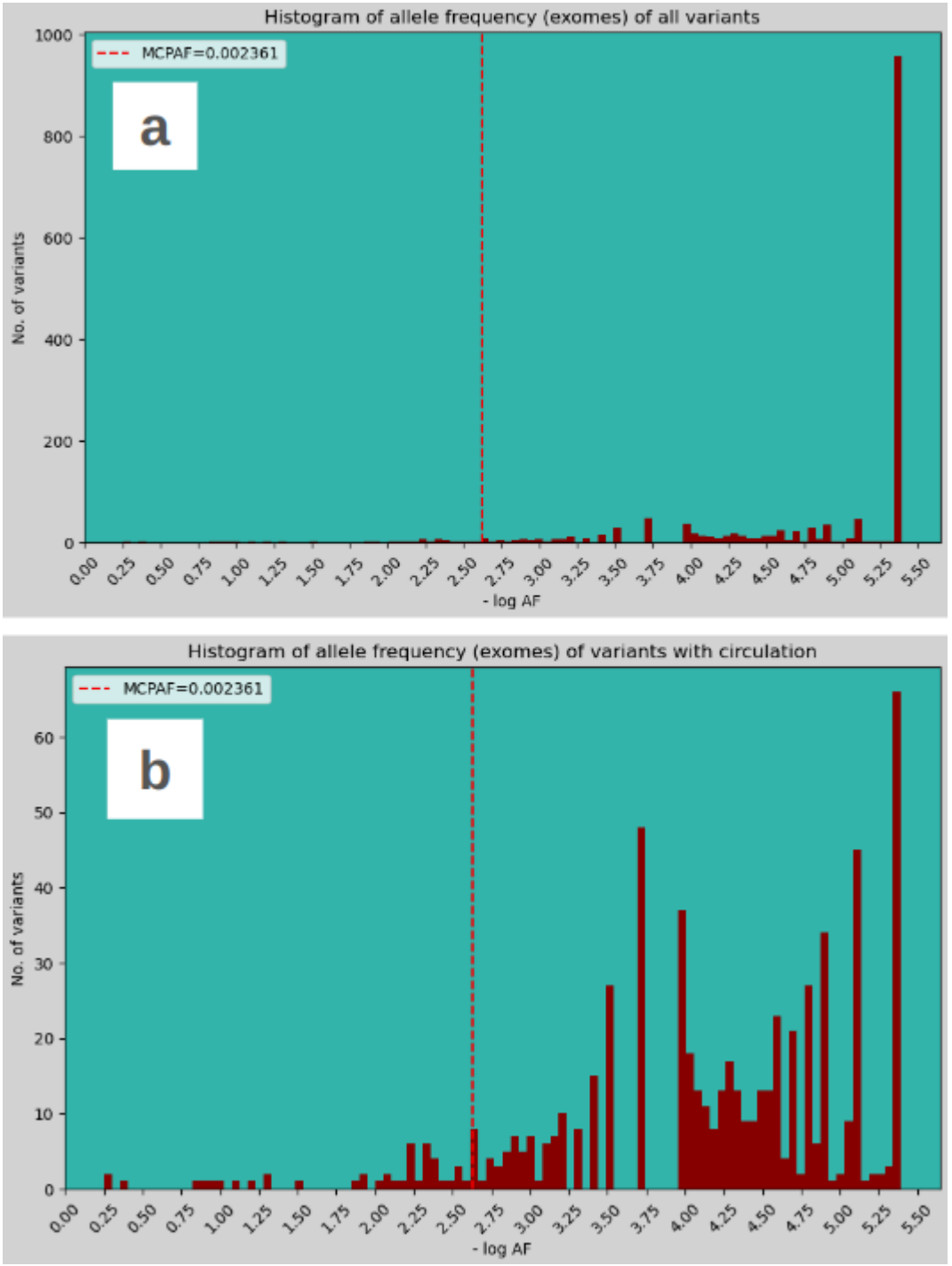

Figura 17. a) Histograma del número de variantes documentadas en función del menos logaritmo de la frecuencia alélica en exomas de gnomAD (-log AF) para todas las variantes incluídas. b) Histograma análogo pero considerando solamente aquellas variantes con AF no-nula en exomas de gnomAD. En ambas figuras se traza en rojo la recta que marca la MCPAF calculada por medio del algoritmo de Whiffin.

Para el cálculo de la MCPAF recurrimos a bibliografía a fines de recabar los datos de prevalencia máxima de CHH (1/4.000) y penetrancia (0,5). El gen más frecuentemente mutado fue PROKR2 (217 pacientes con variantes en PROKR2/1.939 pacientes) y la variante más recurrente en el mismo fue NM_144773.4: c.533G>C p.(Trp178Ser) (63 alelos

de la variante/243 alelos de PROKR2). Utilizando la fórmula planteada previamente, obtenemos una MCPAF de 0,002361.

Este valor de corte de MCPAF se muestra en los histogramas de la Figura 17 (ver la línea roja punteada), y si bien uno puede sentirse tentado de simplemente filtrar aquellas variantes con frecuencias poblacionales (o subpoblacionales) superiores a este umbral, es preferible adoptar un enfoque más conservador. Tomando en consideración la variabilidad en la determinación de las frecuencias, lo correcto es filtrar las variantes en función de una frecuencia alélica de filtrado (FAF) previamente calculada. Por un lado, en lugar de trabajar con AF de población general, se toma aquella AF de la subpoblación donde el alelo es más frecuente. Si bien esta estrategia podría parecer agresiva, se fundamenta en el hecho de que dado el impacto marginal que se requiere de una variante para transmitir una patología en forma mendeliana, en general su efecto se verá independientemente de que se produzca en una u otra subpoblación. Por otro lado, una vez determinada la subpoblación donde la variante es más frecuente, la FAF intenta ser sumamente conservadora en el hecho de prevenir que el valor de AF que se obtuvo para dicha subpoblación no sea un outlier por encima de la media. Es decir, modera la posibilidad de que al muestrear el consorcio de secuenciación la población en estudio, haya tomado por azar un inusualmente elevado número de individuos con la variante. Esto haría que su AF se vea sobreestimada, pudiendo ocasionar falsos negativos de variantes al filtrarlas por su AF. Para lograr esto, en términos técnico-estadísticos, lo que se hace es considerar la posibilidad de que el dato muestral obtenido en gnomAD (la AF que leemos sobre el browser) se ubique en el valor correspondiente al 95% del área bajo la curva de la distribución muestral (tipo Poisson) para la frecuencia de la variante. La FAF lo que hace es estimar el valor más probable para esa distribución, que siempre conducirá a un valor de AF por debajo del original. En función de este valor de FAF, al que se accede fácilmente desde la base de datos de gnomAD, podemos filtrar las variantes o realizar otras tareas.

Podemos observar en la figura 18 que al confeccionar un histograma con estos valores de FAF la distribución se mueve hacia la izquierda, posicionando una cantidad mucho mayor de variantes en la zona de frecuencia alélica por encima de la esperada para el desorden, y por lo tanto fortaleciendo la evidencia sobre su posible benignidad.

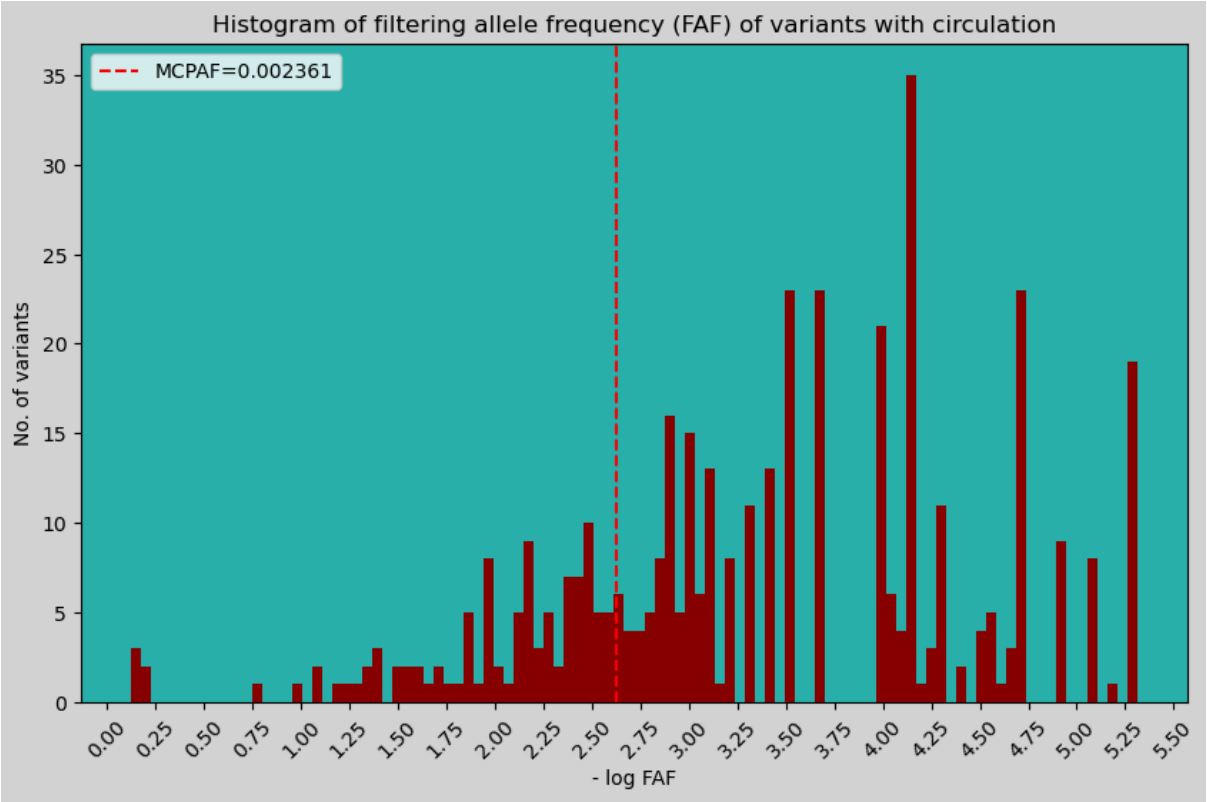

Figura 18. Histograma del número de variantes en función de menos logaritmo de la frecuencia alélica de filtrado de gnomAD (-log FAF) para todas las variantes con FAF no-nula. Se traza en rojo la recta que marca la MCPAF calculada por medio del algoritmo de Whiffin.

Una métrica que podría servirnos para apreciar cómo se van distribuyendo las variantes relevadas entre los grupos de clasificación “Patogénica/Benigna” y “FAF mayor a la MCPAF/FAF menor o igual a la MCPAF” es la de confeccionar una tabla de contingencia y observar cómo va variando el OR, para luego realizar un test de Fisher para verificar su significancia a partir del cálculo de su valor p. Para esto primero debemos armar la tabla de contingencia. Si la misma se confecciona a partir de las anotaciones crudas de InterVar, y simplemente tomamos aquellas variantes clasificadas como P/LP/LB/B, y luego practicamos el Test de Fisher, el resultado es el siguiente:

	FAF > MCPAF	FAF ≤ MCPAF
Variantes benignas	10	0
Variantes patogénicas	0	354

Valor de Odds Ratio: inf Valor p del test de Fisher: 1,01e-19

Como podemos ver, en este caso la separación sería perfecta. Sin embargo, resulta muchísimo más interesante observar qué pasaría si incorporamos a las variantes de significado incierto (VUS) en el cálculo. Para esto, podemos valernos del pasaje del conjunto de etiquetas asignado a cada variante, hacia su interpretación probabilística generando un score bayesiano. Si trazamos un valor de corte de probabilidad en 0,5, y establecemos que aquellas variantes con una probabilidad de patogenicidad que supere el

mismo sean consideradas en este caso las patogénicas, la tabla y sus resultados serán los siguientes:

	FAF > MCPAF	FAF ≤ MCPAF
Post_P ≤ 0.5	102	510
Post_P > 0.5	5	901

Valor de Odds Ratio: 36,04 Valor p del test de Fisher: 2,87e-36

Como podemos observar, las variantes que cuentan con suficiente evidencia para cruzar el valor de corte de probabilidad de patogenicidad poseen en su mayoría una FAF por debajo del valor de corte establecido por la MCPAF, salvo contadas excepciones. En cuanto a las variantes que poseen menores evidencias de patogenicidad, observamos que muchas poseen una FAF por encima de la MCPAF, lo que las posiciona a priori como falsos positivos reportados en bibliografía. Similar a lo que ocurre en el panorama genómico, muchas variantes con escasa evidencia de patogenicidad son poco frecuentes, y deberán valerse de otros criterios para poder migrar tanto hacia clasificaciones patogénicas como benignas.

Utilización de la evidencia de casos previos de variantes en el ajuste de su interpretación

Un dato extremadamente interesante, y que precisamente constituye uno de los grandes objetivos de la revisión sistemática, es observar qué variantes fueron evidenciadas repetidamente en diversos individuos (y que dichos individuos no se encuentren repetidos en diversos trabajos, habitual de encontrar en el día a día, pues haría que se cuente repetidamente la misma evidencia).

Las últimas recomendaciones emitidas por ACMG y grupos especializados de expertos en lo respectivo a la aplicación de etiquetas referidas a pacientes reportados previamente, apuntan en primer lugar a que debemos diferenciar entre variantes localizadas en genes asociados a los diferentes modelos de herencia. Por lo mismo, el criterio PS4 tal como lo conocemos se aplicará en aquellos casos donde el modelo de herencia asociado al gen sea AD o XL. Se solicita que la variante presente una FAF inferior a la MCPAF. Para contar casos individuales de aparición de la variante, se comprueba que su cigosidad en el paciente masculino sea heterocigota o hemicigota o, en el caso de pacientes femeninos, homocigota en caso que la variante se localice en el cromosoma X o heterocigota en los demás. Si no contamos con el dato de sexo de algún paciente, solo consideramos variantes presentes en autosomas en estado heterocigota. Las variantes con mayor cantidad de casos se presentan en la Figura 19.

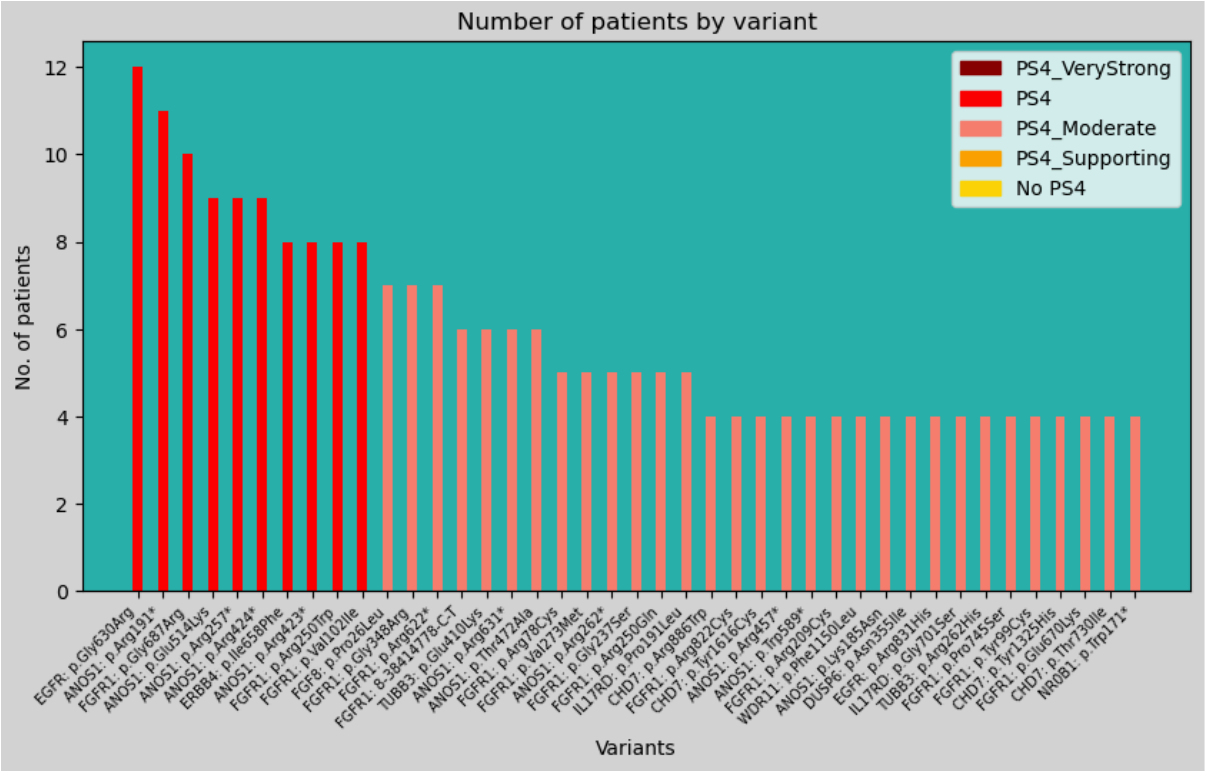

Figura 19. Gráfico de barras presentando de mayor a menor, las variantes localizadas en genes asociados a patrones de herencia AD o XL con mayor cantidad de apariciones en pacientes. Los colores bordó, rojo, rosa o naranja se adjudican en función del número de pacientes poseedores de la variante (mayor o igual a 16, 8, 4 y 2 respectivamente).

En la Figura 19 se observa una coloración en función del número de pacientes poseedores de la variante en cuestión. Estos colores se asignan en función de valores de corte para el número de pacientes, siendo estos 16, 8, 4 y 2 pacientes o más. Estas categorías se asocian a modificaciones en el grado de evidencia que se termina adjudicando al criterio. Estos valores de corte fueron elegidos en función de mantener un enfoque conservador (exigiendo al menos 2 pacientes poseedores de la variante para empezar a aplicar el criterio y asignar la etiqueta correspondiente) mientras que se hace valer la astringencia en la especificidad fenotípica de los individuos para entrar en la revisión y se permite llegar al máximo nivel de evidencia en caso de encontrarse documentada en al menos 16 pacientes. En el medio, se respeta la progresión exponencial de 2. Como podemos ver, ninguna variante logra un grado de evidencia Very Strong, sin embargo muchas de ellas conservan el grado original Strong, y las demás se ubican en las categorías rebajadas PS4_Moderate o PS4_Supporting.

Para las variantes localizadas en genes asociados a modelo de herencia AR, el conteo se ha visto modificado en función de una cantidad de variables a considerar, incluyendo la confirmación de la baja frecuencia de la propia variante, la cigosidad de la variante (teniendo en consideración la posibilidad de una homocigosis o una heterocigosis compuesta para el caso en estudio), la confirmación o presunción de encontrarse la otra variante en trans y la clasificación de la variante acompañante. Tomamos la decisión de comprobar que la propia variante posea una FAF por debajo de la MCPAF, y asumimos, dado el grado de curación de los trabajos pesquisados, que las variantes compatibles con casos de heterocigosis compuesta se encontraban confirmadas en trans.

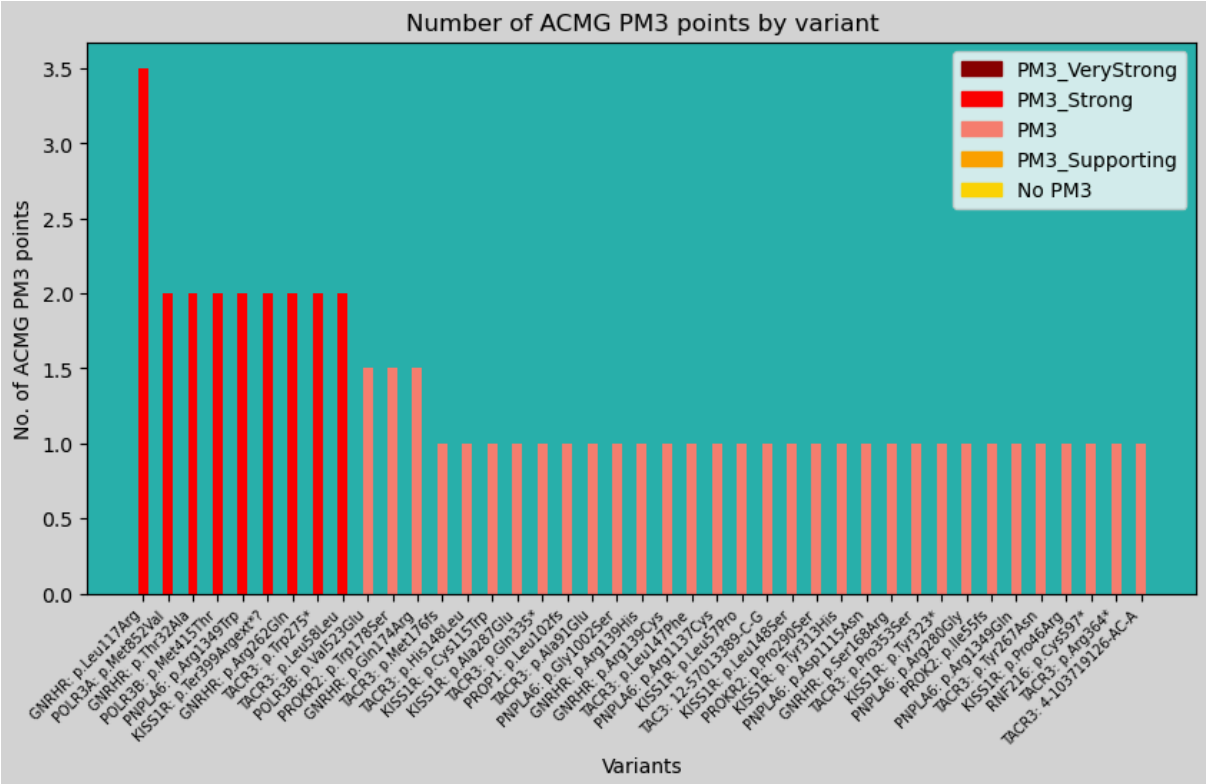

Figura 20. Gráfico de barras presentando de mayor a menor, las variantes localizadas en genes asociados a patrón de herencia AR con mayor cantidad de apariciones en pacientes. Los colores bordó, rojo, rosa o naranja se adjudican en función del puntaje obtenido según recomendaciones de la ACMG sobre el criterio PM3 (mayor o igual a 4, 2, 1 y 0.5 respectivamente).

Optimización del uso de predictores bioinformáticos en interpretación de variantes

En el diagnóstico molecular de EPoFs, la clasificación de la patogenicidad de variantes ha avanzado con la evaluación de predictores bioinformáticos. Estos predictores fueron combinados para mejorar su precisión, dando lugar a los metapredictores. Entre ellos, REVEL se destacó por su alta reproducibilidad en variantes missense, convirtiéndose en una herramienta ampliamente aceptada. Pejaver et al. calibraron estos predictores, permitiendo ajustar el nivel de evidencia del criterio PP3/BP4. La Figura 21 muestra un histograma con los valores de predicción de REVEL para las variantes de la revisión sistemática, cuyo score varía de 0 a 1, indicando 1 la máxima certeza de patogenicidad.

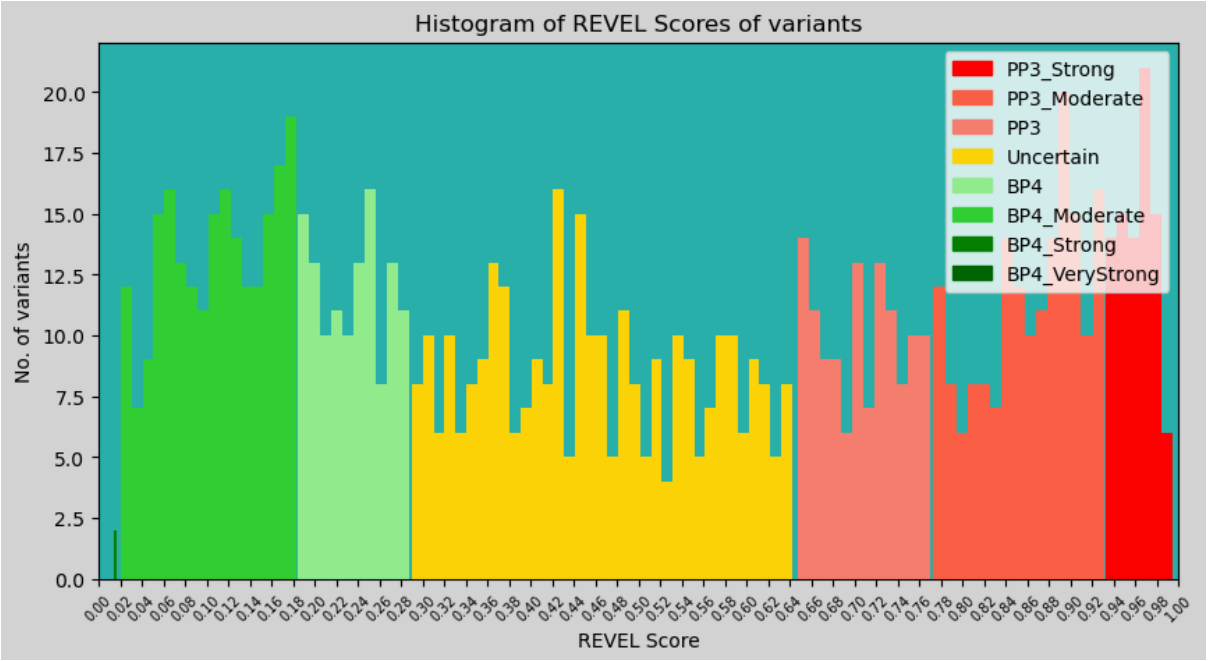

Figura 21. Histograma de las predicciones efectuadas por REVEL para las variantes missense incluídas en la revisión sistemática. Los colores indican los rangos de predicciones de REVEL donde se producen los ajustes de los criterios PP3/BP4.

Como podemos ver, los valores de las predicciones se encuentran formando una distribución bastante homogénea a lo largo del rango 0-1. Sin embargo, la calibración del veredicto de REVEL y de otros predictores nos permitió entender que no todas las predicciones significan lo mismo, más aún, en función de su valor numérico podemos ajustar el nivel de evidencia asociado, llegando a grados muy por encima de los tradicionales Supporting que implican PP3 y BP4. Estos rangos se encuentran indicados en la Figura 21, y en la Figura 22 podemos observar los porcentajes de variantes recayendo en los mismos. Vemos que prácticamente el 70% de las variantes cuenta con algún tipo de predicción no-incierta sobre su efecto molecular, como así también podemos observar que el 45% cuenta con predicciones por encima del nivel de evidencia Supporting, lo cual puede sernos de gran utilidad a la hora de interpretar mejor a las mismas, en especial las desafiantes VUS.

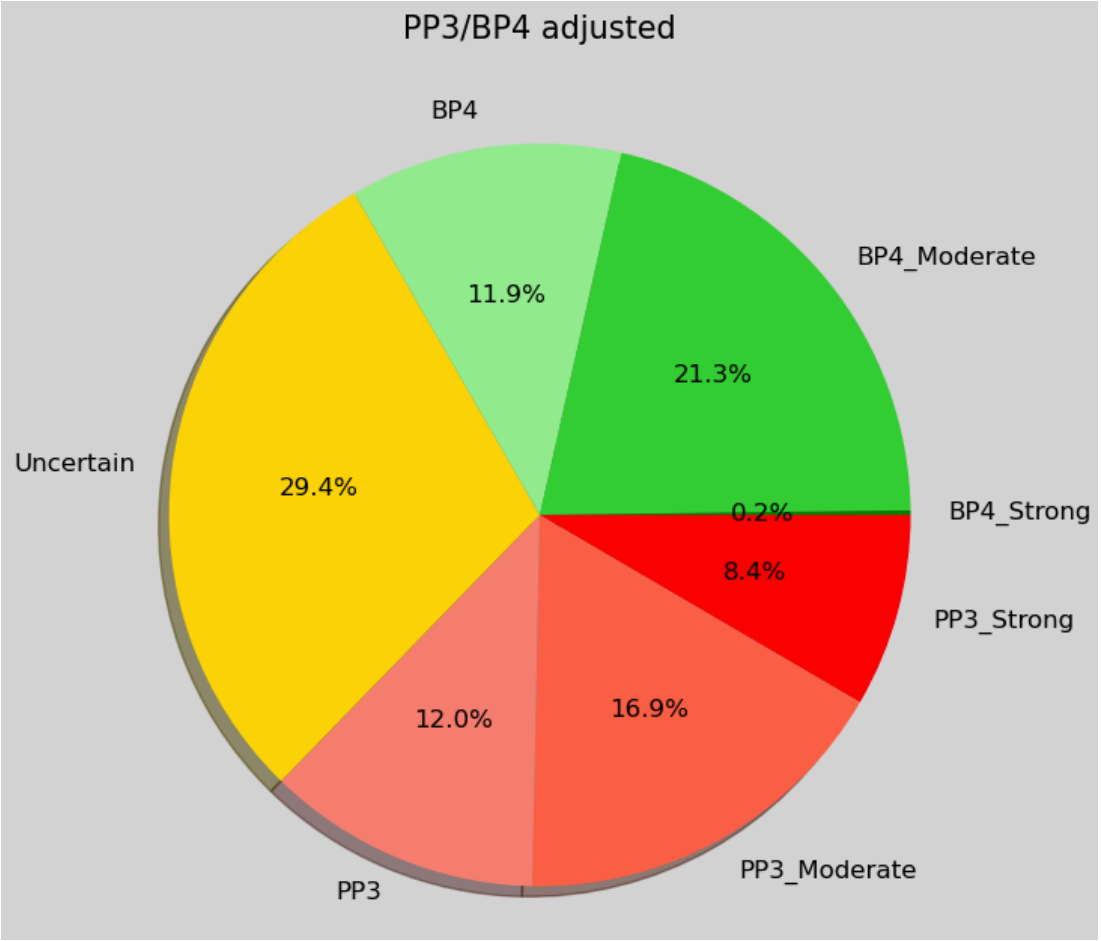

Figura 22. Gráfico de torta indicando la distribución porcentual de los diferentes valores de las predicciones de REVEL sobre las variantes missense de la revisión sistemática.

Evidencia aportada por entradas de las variantes en bases de datos

Una fuente de evidencia que fue pensada originalmente por los autores de las recomendaciones ACMG, fue la de considerar el reporte previo de la variante en alguna base de datos de variantes de interés clínico, aún cuando sus fundamentos no fueran accesibles. Con el pasar de los años, con el advenimiento del enfoque conservador a la hora de la interpretación, esta fuente de evidencia fue de las primeras en ser desestimada. InterVar utilizaba bases de datos como ClinVar para introducir PP5 y BP6, las etiquetas asociadas a reportes patogénicos o benignas de las variantes, respectivamente. Como vemos en la Figura 23, alrededor del 20% de las variantes contaba con un aporte de evidencia de este tipo.

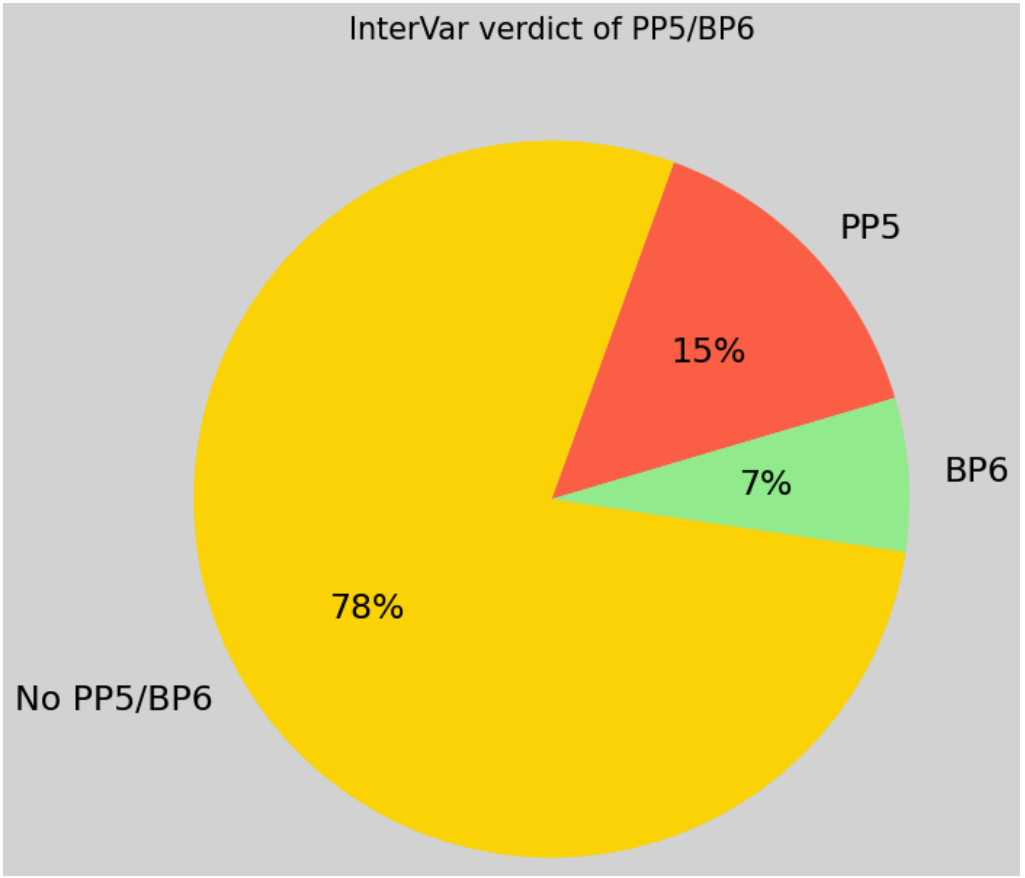

Figura 23. Gráfico de torta donde se muestra los porcentajes de variantes a las cuales InterVar le asigna PP5, BP6 o ninguna de las dos.

Refinado automático de los criterios ACMG para optimizar la priorización e interpretación de variantes en casos de CHH

Como hemos visto, el estudio de la casuística de casos de la patología en estudio nos permite desarrollar herramientas extremadamente útiles para la resolución futura de casos. Sin embargo, son muchas otras las aplicaciones que podemos dar a toda esta información, como podría ser la anotación automática de variantes por un pipeline bioinformático. En nuestro caso, basados en todos los estudios de variables inherentes a las variantes explicados anteriormente, realizamos modificaciones al anotado de las mismas por el tradicional software bioinformático InterVar, generando una versión más precisa del mismo para las variantes de esta patología. El mismo fue concebido como un algoritmo pensado para asignar automáticamente los distintos criterios de evidencia a las variantes, pero donde nosotros podemos intervenir en el mismo para tomar dos decisiones fundamentales, como ser: 1) decidir asignar o no un determinado criterio a una variante, y 2) decidir el grado que deseamos asignar a un determinado criterio aplicado a una variante. Como podemos ver en la Figura 24, estas dos intervenciones traen aparejadas cambios en la clasificación original de las variantes. Por otro lado, se aprecia que la clasificación VUS se ve fraccionada en tres grupos dependiendo de la probabilidad a posteriori de patogenicidad, tal como se definió cuando se pasó al enfoque bayesiano para la clasificación, y luego se comenzó a hablar de las High VUS, Medium VUS y Low VUS.

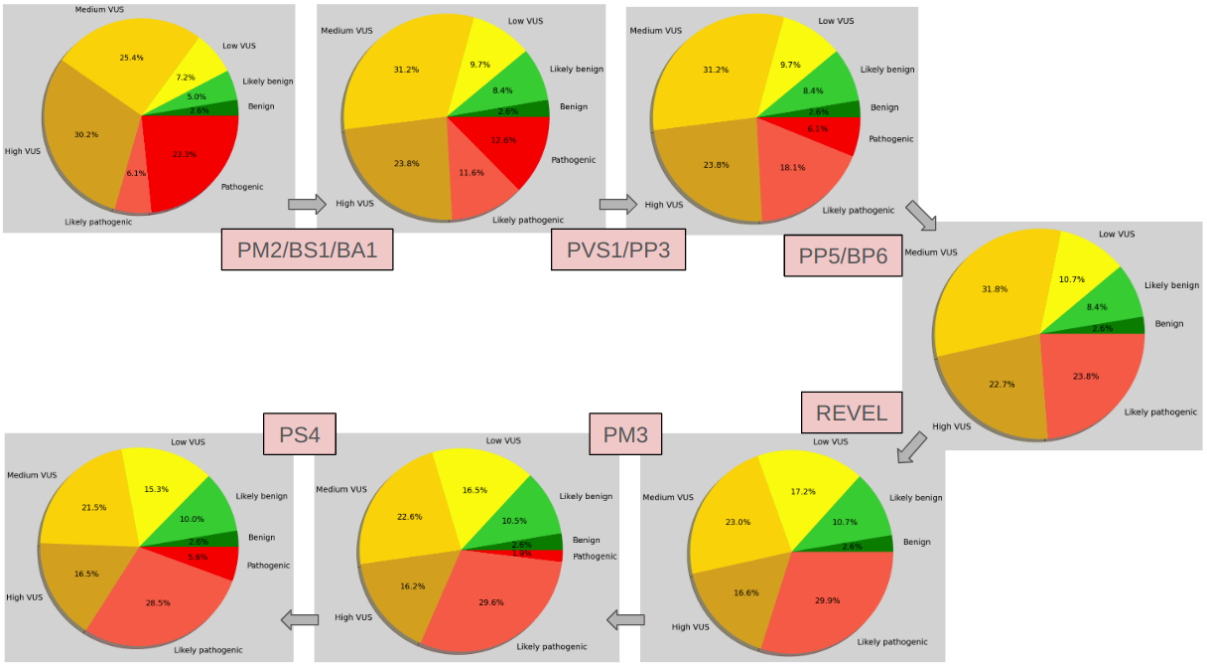

Figura 24. Progresión de cómo se ven modificados los porcentajes de variantes correspondientes a cada categoría según Richards 2015, como así también las subcategorías dentro de las VUS.

Dentro de los resultados más significativos podemos comentar que el conjunto de variantes pertenecientes a categorías P/LP/LB/B pasa, con el mejorado esquema de clasificación, de un 37% a un 47%, lo cual nos indica que contamos con una mayor cantidad de variantes con una buena confianza de estar siendo correctamente clasificadas. Por otro lado, las VUS como un todo disminuyen de un 63% a un 53%, y resulta interesante apreciar que del 100% de las mismas, las Medium VUS, aquellas que están en una situación de particular incertidumbre, pasan de constituir un 25,4% a un 21,5 %. Estos movimientos en la clasificación de las variantes nos permitirań luego realizar filtrados más potentes, ordenando las variantes de manera más efectiva en función de su probabilidad de patogenicidad, y ahorrar tiempo y esfuerzo, no solo para encontrar un diagnóstico molecular, sino también para investigar y asignar la evidencia de criterios como los referidos a casos reportados previos (PM3, PS4).

El horizonte genético del CHH: hacia una mejor elaboración de paneles de estudio

Más allá de la naturaleza de las variantes genéticas halladas, otro gran interrogante abierto al realizar la revisión sistemática era el de apreciar la presencia de muchas de las mismas localizadas sobre genes de dudosa conexión biológica con la fisiopatogenia y fenotipo asociados, o sea CHH. La misma pregunta se plantea cotidianamente en los grupos especializados cuando se debe responder si una variante compatible con ser causal de una alteración molecular (es decir, con etiquetas moleculares de carácter patogénicas) hallada en un paciente, se encuentra en un gen con indicios suficientes para explicar la relación genotipo-fenotipo, y por lo tanto ser considerada diagnóstica. Para investigar aún más sobre las relaciones existentes entre los genes donde se encontraron variantes en la revisión, se procedió a realizar un estudio de redes funcionales de genes utilizando el software Cytoscape, principalmente stringApp, librería que posee montada toda la información de la base de datos STRING. En la Figura 14a podemos observar como se

ordena la red funcional para los 143 genes evaluados en esta la revisión (en esta red las aristas uniendo los nodos puede constituir cualquier tipo de evidencia, desde contactos proteína-proteína evidenciados por experimentos, hasta la co-aparición de los dos nombres de los genes en una única publicación). En la Figura 25b solo se muestran representadas las interacciones proteína-proteína, pero los nodos se encuentran en exactamente el mismo lugar que en la Figura 25a. En ambas figuras observamos algunos nodos más grandes, correspondientes a aquellos genes que la base de datos STRING ya reporta una asociación a hipogonadismo hipogonadotrófico (responde a una entrada de enfermedad en STRING).

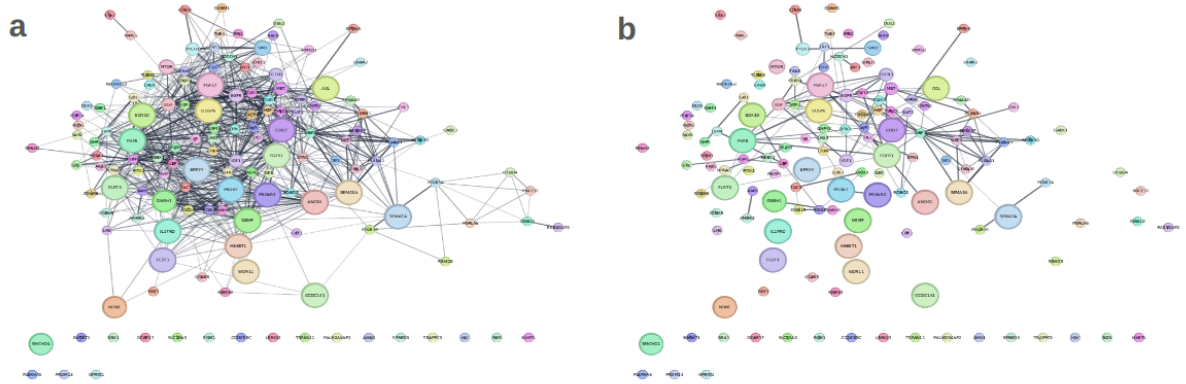

Figura 25. a) Red mostrando las interacciones funcionales existentes entre todos los genes relevados en la revisión sistemática. El tamaño de los nodos posee un tamaño estándar a menos que la base de datos STRING reporte una asociación previa con hipogonadismo hipogonadotrófico. Como esta asociación se realiza en forma de score, el tamaño del nodo es proporcional al valor del mismo. El grosor de las aristas refleja el grado de confianza que se asigna en STRING a la interacción. b) Misma red de genes pero donde las aristas representan exclusivamente interacciones proteína-proteína.

Como podemos ver, al considerar el conjunto de evidencia funcional, una gran mayoría de los nodos de mayor tamaño se encuentran muy bien integrados a la red, mientras que cuando solo dejamos los contactos proteína-proteína, muchos de ellos se desconectan. Esto podemos verlo mejor en la Figura 26 donde se produjo simplemente un reordenamiento espacial de los nodos para mejorar la visualización.

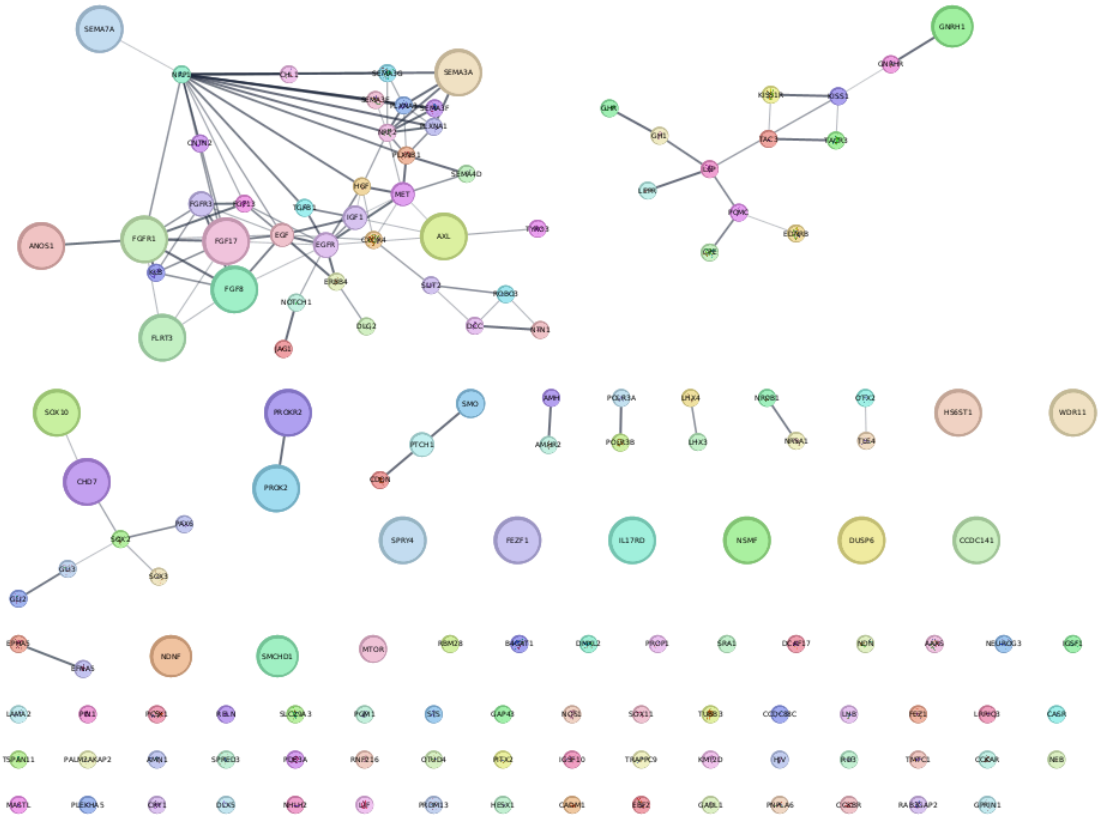

Figura 26. Red de genes reordenada donde las aristas representan interacciones proteína-proteína.

Para seguir ilustrando este panorama, podemos aplicar el algoritmo de clustering de Markov (MCL) que nos permite hallar clusters naturales dentro de las redes de genes que planteamos previamente, tanto en la funcional como en la de interacciones proteína-proteína. Ambas podemos visualizarlas en la Figura 27.

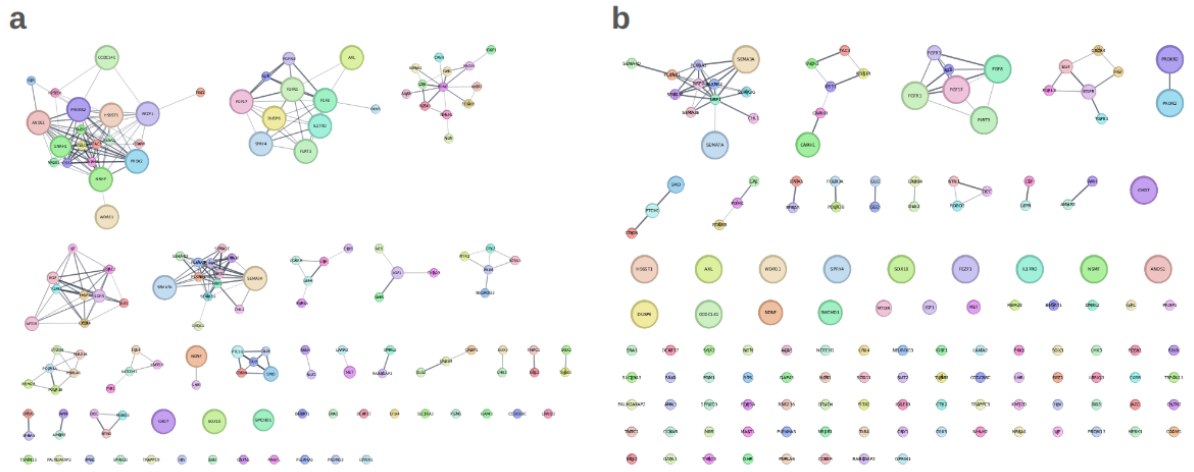

Figura 27. a) Red funcional de genes tras aplicar el algoritmo MCL. b) Red de interacciones proteína-proteína correspondientes a sus genes tras aplicar el algoritmo MCL.

Como podemos observar, en las distintas visualizaciones siempre queda una gran cantidad de nodos por fuera de la red principal, o los clusters naturales que el algoritmo encuentra. Por lo tanto, pensando primeramente en la biología y fisiopatología subyacente

procedimos a realizar un análisis de enriquecimiento de términos en este set de genes (GSEA) con la herramienta de Cytoscape. Una vez obtenidos los resultados, nos quedamos con las categorías de FDR más bajo y pertinentes con la fisiopatología del hipogonadismo hipogonadotrófico de acuerdo con las bases de datos DISEASES, Monarch Phenotype y GO Biological Process.

Tabla 1. Bases de datos y categorías seleccionadas para el análisis de enriquecimiento de términos.

Base de datos	Términos	FDR
Diseases	Endocrine system disease	1.63e-46
	Hypogonadotropic hypogonadism	2.83e-38
	Kallmann syndrome	4.60e-35
	Gonadal disease	5.09e-33
	Genetic disease	3.84e-24
	Pituitary gland disease	5.17e-16
	Hypopituitarism	3.14e-15
	Monogenic disease	3.07e-13
	Disorders of sexual development	1.63e-05
	Gonadal dysgenesis	3.00e-04
Monarch Phenotype	Hypogonadotropic hypogonadism	2.86e-68
	Delayed puberty	3.87e-41
	Puberty and gonadal disorders	4.04e-54
	Abnormality of the endocrine system	1.7e-52
	Decreased testicular size	1.85e-49
	Abnormal circulating hormone concentration	2.00e-49
	Abnormality of reproductive system physiology	5.02e-48
	Abnormality of the hypothalamus-pituitary axis	9.67e-48
	Micropenis	4.13e-31
	Anosmia	1.60e-42
GO Biological Process	Neurogenesis	6.87e-38
	Axon guidance	4.85e-36
	Generation of neurons	2.72e-35
	Nervous system development	5.87e-34
	Forebrain development	1.09e-33

	Axonogenesis	2.17e-16
	Cell differentiation	3.52e-31
	Chemotaxis	1.67e-30
	Cell migration	2.2e-22
	Neuron migration	2.74e-18

Una vez que contamos con los datos del GSEA, procedimos a contar las apariciones de los genes en las diferentes categorías, a fines de saber en qué porcentaje de las mismas se encontraban. Esta información podemos verla representada en el histograma de la Figura 28.

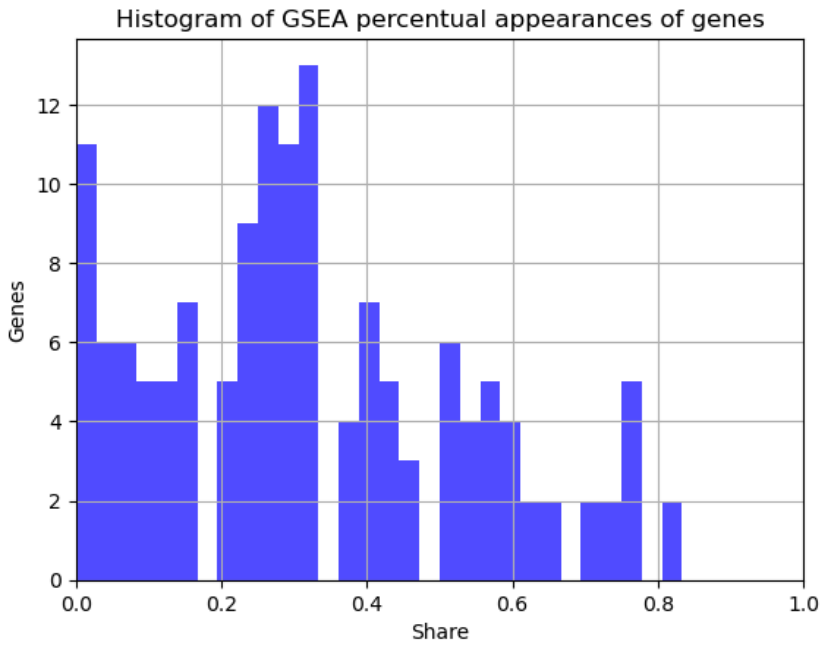

Figura 28. Histograma de los porcentajes de pertenencia a las diferentes categorías seleccionadas para los genes incluídos en la revisión sistemática.

Decidimos tomar un valor de corte de 0,2 para definir a un subconjunto de 40 genes que presentaban a priori una asociación escasa con la fisiopatogenia que estudiamos, mientras que los otros 103 genes tenían a priori según GSEA una alta asociación con CHH. Luego, se procedió a analizar si estos grupos de genes presentaban un enriquecimiento en variantes benignas y/o patogénicas. Para separar las variantes se consideró como benignas aquellas con una probabilidad a posteriori de patogenicidad menor o igual a 0,188, mientras que se consideró variante patogénica a aquella con probabilidad a posteriori de patogenicidad mayor o igual a 0,675 según el enfoque bayesiano de ACMG. Los resultados se presentan a continuación:

	GSEA % ≥ 0,2	GSEA % < 0,2
Post_P ≥ 0,675	748	20
Post_P ≤ 0,188	37	357

Valor de Odds Ratio: 3,88 Valor p del test de Fisher: 1,08e-06

Como podemos observar, existe un claro enriquecimiento en variantes con abundante evidencia de patogenicidad y cuyos genes se encuentran relacionados a los términos de búsqueda sobre los cuales realizamos el GSEA. Y viceversa, aquellos que poseen menores evidencias de patogenicidad, tienden a encontrarse menos ligados a los mismos. Dicho de otro modo, el análisis de enriquecimiento (GSEA) permite seleccionar genes que poseen una notoria probabilidad de poseer variantes patogénicas asociadas a CHH.

El hecho de haber construido una red de genes, permite también monitorear alguna variable topológica de la misma como el grado de los nodos, es decir, la cantidad de conexiones que los mismos presentan con sus pares. En la Figura 29, podemos observar el histograma general del grado de los nodos-gen que presenta en forma superpuesta las distribuciones correspondientes al grupo de genes que presentaban valores superiores o inferiores al valor de corte de 0,2 para el porcentaje de apariciones en las categorías del GSEA.

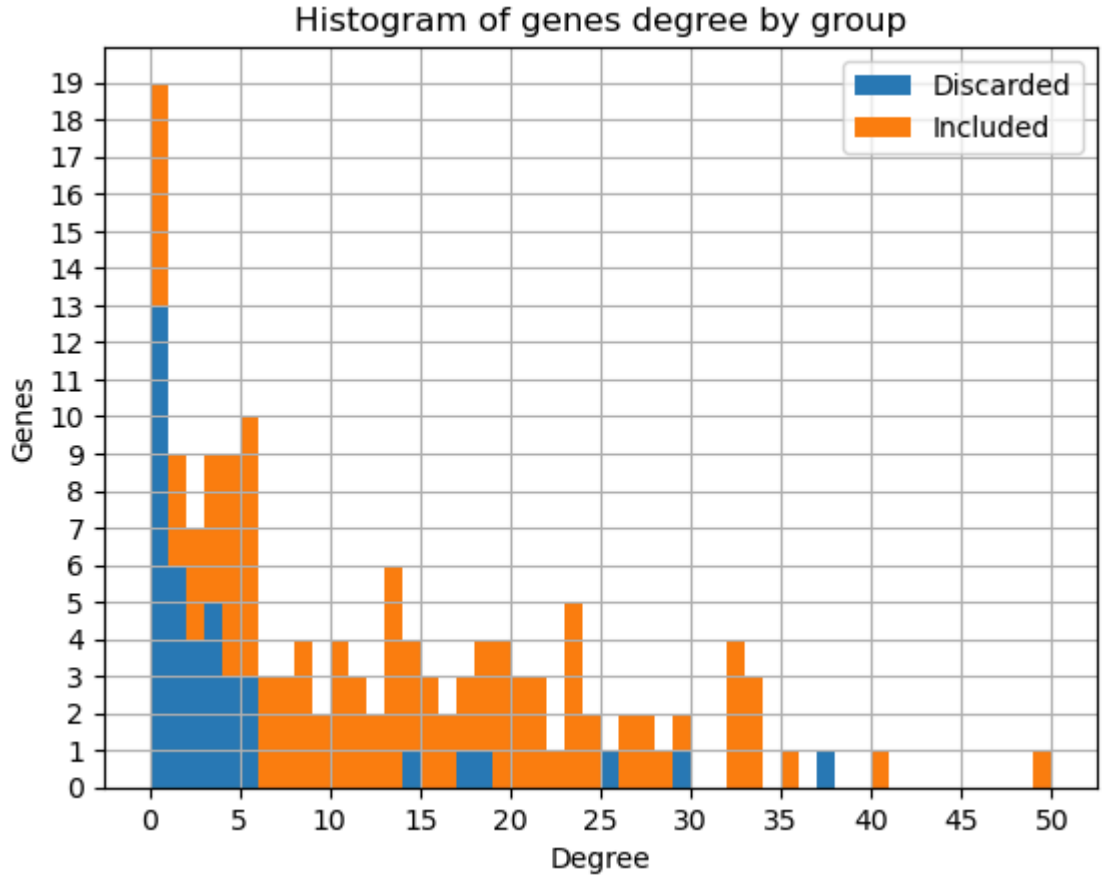

Figura 29. Histograma del grado de los nodos, diferenciándose la población de los mismos correspondientes a genes que obtuvieron un porcentaje de apariciones en categorías del GSEA ≥0,2 (Included) o <0,2 (Discarded).

Observando este histograma, podríamos poner un valor de corte para el grado de 5. Para definir aquellos genes candidatos asociados a CHH. Con esta información podremos definir un primer panel conformado por genes con porcentaje de apariciones en las

categorías del GSEA mayor o igual a 0,2 y grado mayor o igual a 5, constituído por 74 genes:

ANOS1, PROKR2, FGFR1, PROK2, SEMA3A, CHD7, LEPR, KISS1R, IL17RD, AXL, FGF8, FGF17, NR0B1, GNRHR, NRP2, GNRH1, NOTCH1, TACR3, KISS1, SOX10, PROP1, SOX2, POLR3A, SLIT2, DCC, SEMA3E, HS6ST1, NSMF, TUBB3, PLXNA1, SPRY4, NRP1, TAC3, HESX1, SEMA3F, PLXNA3, PLXNB1, DUSP6, PCSK1, FEZF1, SOX3, LHB, SEMA7A, SEMA3G, PTCH1, LHX3, LEP, FLRT3, OTX2, ERBB4, CNTN2, EFNA5, IGF1, PAX6, SEMA4D, EDNRB, GH1, GHR, MET, GLI2, LHX4, PITX2, POMC, CXCR4, GAP43, FGF13, GLI3, JAG1, SMO, NR5A1, NTN1, AMH, AMHR2, TGFB1

Luego podemos pensar un panel donde se incluye al anterior, pero se agregan aquellos genes con solo una de las dos características presentes (35 genes), como buenas hipótesis:

ANOS1, PROKR2, FGFR1, PROK2, SEMA3A, CHD7, LEPR, KISS1R, IL17RD, AXL, FGF8, FGF17, NR0B1, LAMA2, GNRHR, NRP2, GNRH1, NOTCH1, TACR3, KISS1, SLC29A3, RNF216, CCDC141, SOX10, PROP1, SOX2, POLR3A, EGF, SLIT2, DCAF17, DMXL2, CDON, RELN, PNPLA6, TRAPPC9, DCC, POLR3B, SEMA3E, HS6ST1, NSMF, TUBB3, SMCHD1, NDNF, PLXNA1, SPRY4, NRP1, WDR11, TAC3, HESX1, NDN, SEMA3F, PLXNA3, PLXNB1, DUSP6, PCSK1, FEZF1, SOX3, LHB, SEMA7A, NHLH2, SEMA3G, PTCH1, LHX3, CHL1, LEP, FLRT3, OTX2, RBM28, MTOR, EGFR, B4GAT1, ROBO3, DLX5, HGF, ERBB4, CNTN2, FEZ1, EFNA5, IGF1, PAX6, SEMA4D, EDNRB, LIF, TYRO3, GH1, GHR, MET, EPHA5, CCKAR, SOX11, GLI2, LHX4, PITX2, FGFR3, POMC, CXCR4, GAP43, FGF13, GLI3, JAG1, SMO, NR5A1, NTN1, AMH, AMHR2, RAB3GAP2, TGFB1, HJV, NEUROG3

Por último, podemos conformar un grupo de genes de grado bajo, siempre y cuando los mismos no presenten una asociación curada por STRING a hipogonadismo hipogonadotrófico. Este grupo, de 49 genes, comparte con el grupo de genes que obtuvieron un porcentaje de apariciones en categorías del GSEA < 0,2 (40 genes) un total de 31 miembros, es decir, un 77,5% de los mismos. Estos son:

PGM1, IGSF1, CASR, AAAS, NEB, GPRIN1, PIN1, PRDM13, PDE3A, CCDC88C, SRA1, OTUD4, CADM1, STS, CRY1, TMTC1, RD3, NOS1, IGSF10, EBF2, TSPAN11, MASTL, CCKBR, GADL1, SPRED3, DLG2, TLE4, LRRIQ3, PALM2AKAP2, PLEKHA5, AMN1

Dado que estos genes presentan un grado bajo, como así también una escasa asociación al fenotipo en estudio y procesos biológicos asociados a la fisiopatogenia del mismo, resulta un tanto polémico incluirlos en paneles de genes de primera instancia para investigar casos de CHH. Por lo tanto creemos que deben ser relegados para su inclusión en análisis de casos a una segunda instancia, y sólo en caso de que no se hayan encontrado variantes relevantes en los genes candidatos enunciados previamente. La Figura 30 ilustra este panorama.

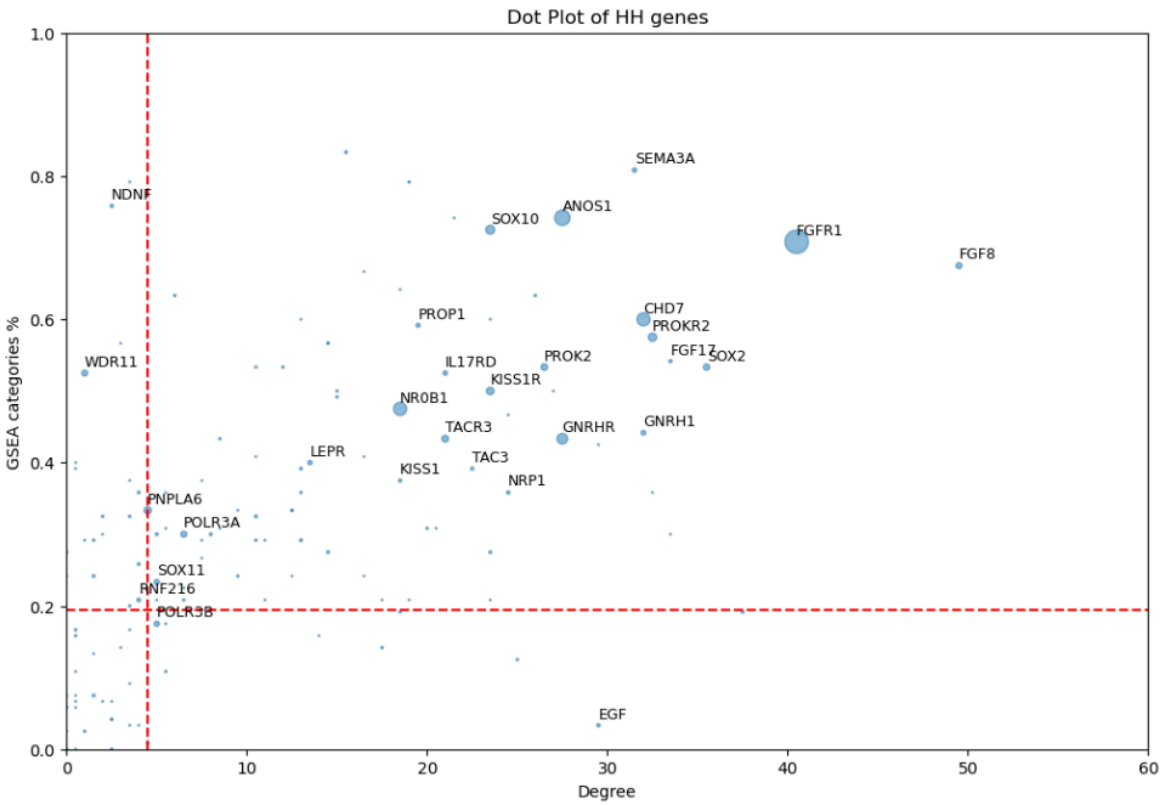

Figura 30. Dot Plot donde se grafican los genes participantes de esta revisión, graficando el porcentaje de categorías del GSEA que integran en función de su grado en la red. Las líneas vertical y horizontal corresponden a los valores de corte que mencionamos anteriormente en el texto de 5 y 0,2, respectivamente. El tamaño de los puntos es proporcional a la cantidad de variantes con probabilidad a posteriori de patogenicidad igual o superior a 0,675 según el enfoque bayesiano de ACMG. Solo se muestra el nombre de aquellos genes que cuentan con al menos 3 variantes de este tipo.

Oligogenicidad en un mundo que busca variantes mendelianas

Si bien los análisis presentados hasta ahora pensamos serán de ayuda para resolver futuros casos de genómica clínica, específicamente a la hora de detectar variantes de alto impacto en genes donde los patrones de herencia son del clásico tipo mendeliano, no podemos desentendernos de los avances que se han hecho en los últimos años en materia de oligogenicidad. Entendemos este concepto como el escenario en el cual variantes localizadas en diferentes genes, con interacciones epistáticas entre ellos, suman sus contribuciones marginales para alterar un proceso fisiológico normal y contribuir al desarrollo de un fenotipo clínico.

La hipótesis de la oligogenicidad como causa de hipognadismo hipogonadotrófico congénito data de varios años y encuentra sustento en diversas publicaciones. Por lo tanto, para empezar nuestro estudio, planteamos en la Figura 31 un heatmap donde en los ejes se posicionaron aquellos genes que cuentan al menos con 10 variantes en este estudio, analizando en cada caso la co-ocurrencia de variantes.

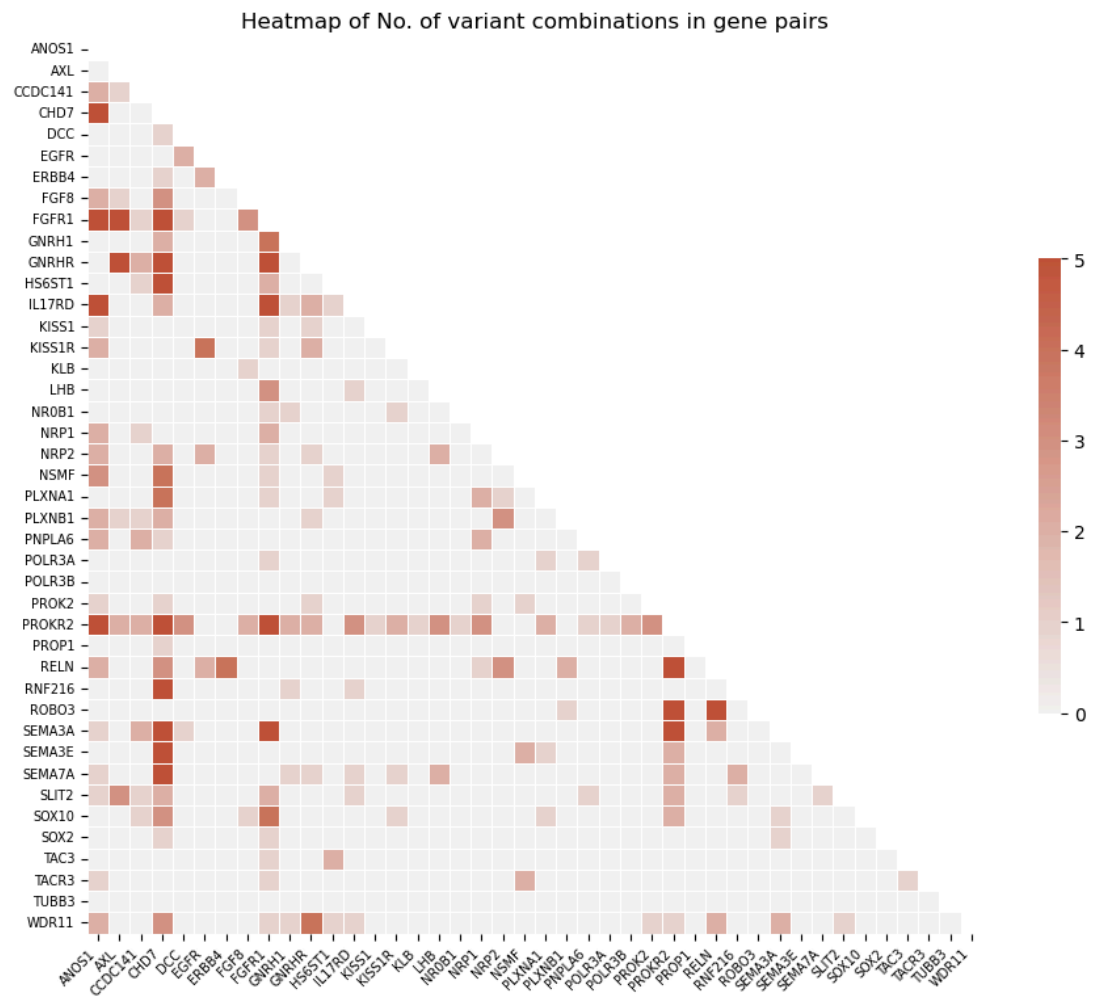

Figura 31. Heatmap donde se observan en sus ejes aquellos genes que cuentan con al menos 10 variantes en esta revisión sistemática. La intensidad de color rojo es proporcional al número de combinaciones (co-ocurrencias) reportadas en pacientes para los dos genes interseccionados en cada celda.

No pasará desapercibida la presencia de genes con una cantidad apreciable de combinaciones, visualmente representado por una gran presencia de celdas de color rojo intenso por fila. Estas filas corresponden a genes como PROKR2, CHD7, FGFR1 o ANOS1. No deja de llamar la atención que son los genes mejor estudiados y que se conocen por su asociación con CHH hace muchos años. Por eso mismo, inmediatamente se realizó un test de correlación de Spearman para enfrentar la presencia de combinaciones en genes contra la presencia de variantes reportadas en general en los mismos. Se halló un valor de coeficiente de correlación de Spearman de 0,76 con un valor p de 1,90e-28. Entendemos que existe una correlación positiva entre el hallazgo de combinaciones y el hecho de que ciertos genes hayan sido más estudiados. El número de combinaciones totales halladas por gen se muestra en la Figura 32a, mientras que los 41 genes con mayor cantidad de asociaciones se muestran en la Figura 32b.

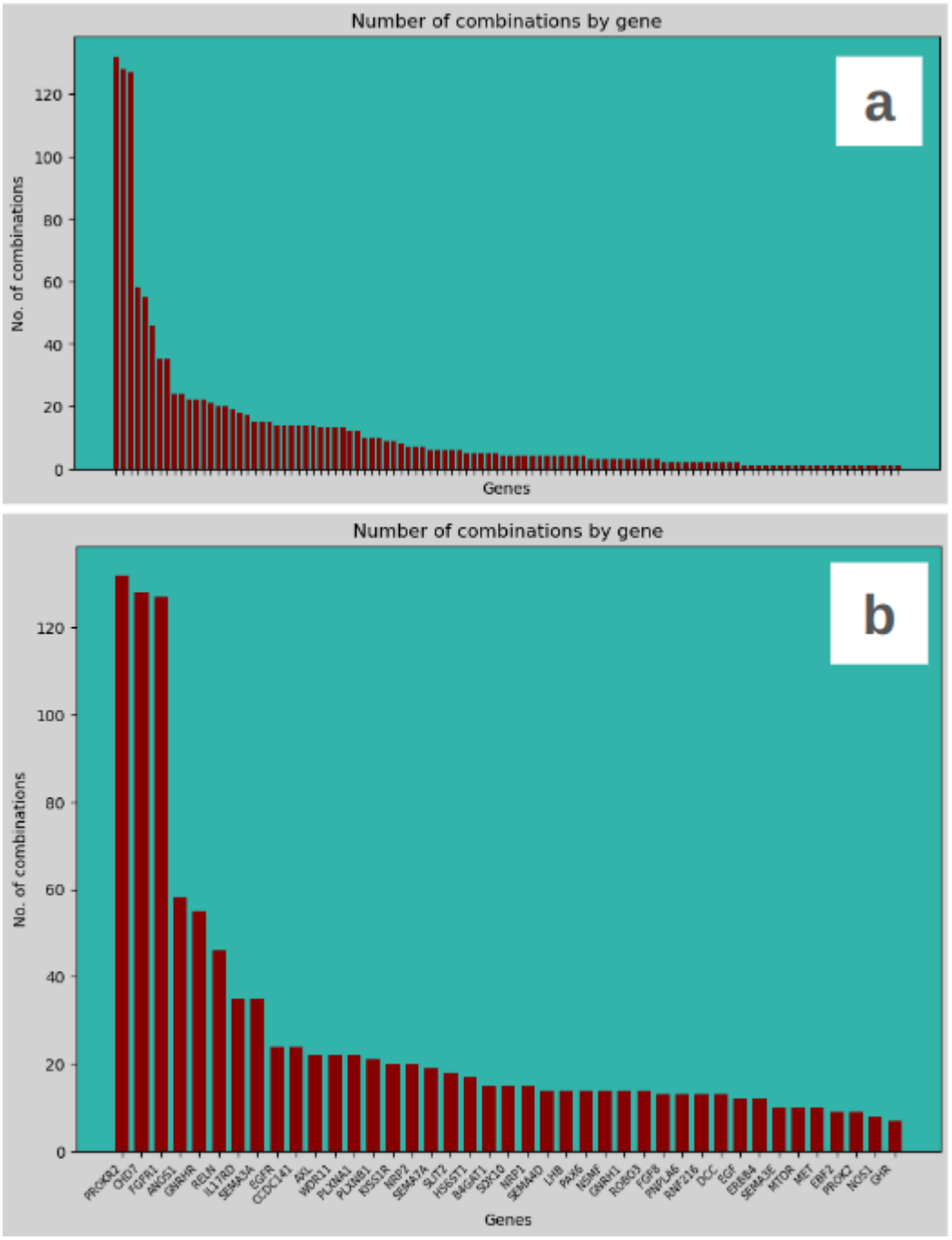

Figura 32. a) Gráfico de barras del número de combinaciones documentadas para todos los genes de la revisión. b) Gráfico de barras del número de combinaciones documentadas para aquellos 41 genes con mayor número de las mismas.

En este contexto, hicimos uso del algoritmo de machine learning basado en random forest llamado ORVAL que busca evaluar potenciales efectos oligogénicos, al otorgar un score de patogenicidad combinado para pares de variantes (VarCoPP). En función del mismo se construyó el histograma de la Figura 33, donde podemos ver la ocurrencia de combinaciones en función de su respectivo score.

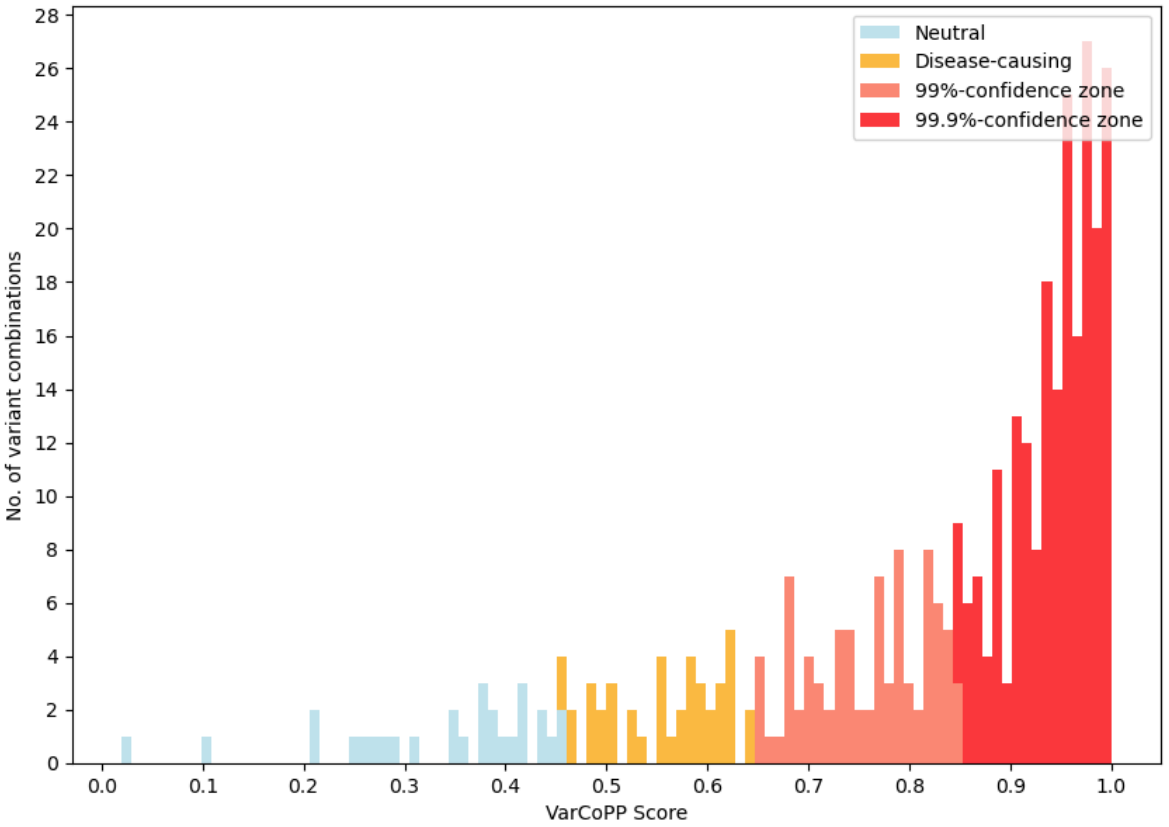

Figura 33. Histograma del score de patogenicidad de combinación de variantes VarCoPP para las combinaciones relevadas en el contexto de la revisión sistemática.

Como podemos ver, existe un enriquecimiento apreciable en combinaciones que superan al menos el umbral correspondiente a la causalidad de enfermedad, independientemente de la confianza de la predicción. Todas estas combinaciones sin embargo pertenecen a diferentes modelos de herencia oligogénica según lo definido previamente en literatura, y que recordemos corresponden a: Verdaderos Digénicos (True Digenic), Monogénico más modificador (Monogenic + Modifier) y Diagnóstico Molecular Dual (Dual Molecular Diagnosis).

En la Figura 34 podemos observar como el histograma de la Figura 32 en verdad puede ser interpretado en sus componentes según modelo de herencia.

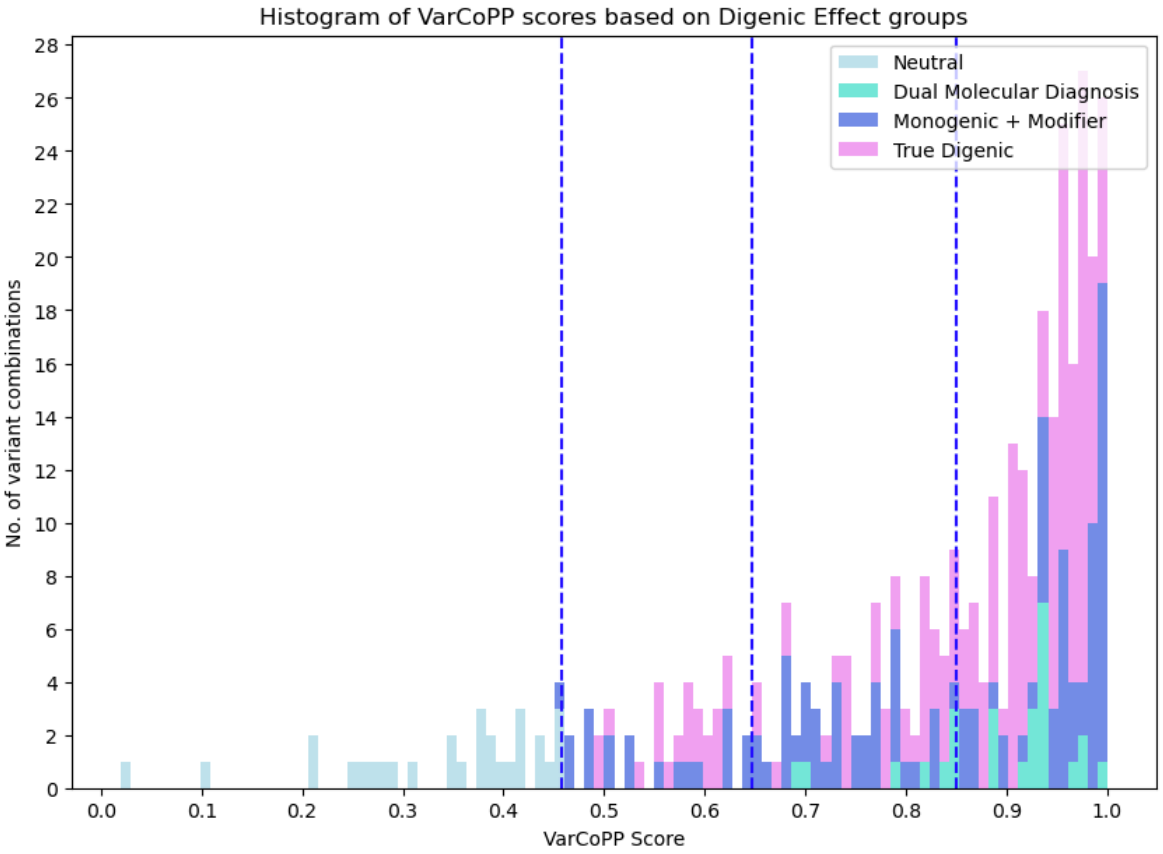

Figura 34. Histograma del score de patogenicidad de combinación de variantes VarCoPP para las combinaciones relevadas en el contexto de la revisión sistemática, discriminado por el modelo de herencia correspondiente a las mismas. Las líneas azules punteadas corresponden a los valores de corte para las categorías de confianza en la predicción de patogenicidad como se apreciaba en la Figura 33.

Logra apreciarse en la Figura 34 una mayoría de combinaciones pertenecientes al modelo de herencia Verdadero Digénico (True Digenic), lo cual puede apreciarse más claramente en la Figura 35b. A su vez, se presenta también en la Figura 35a los porcentajes de combinaciones pertenecientes a las distintas clasificaciones según el score de VarCoPP.

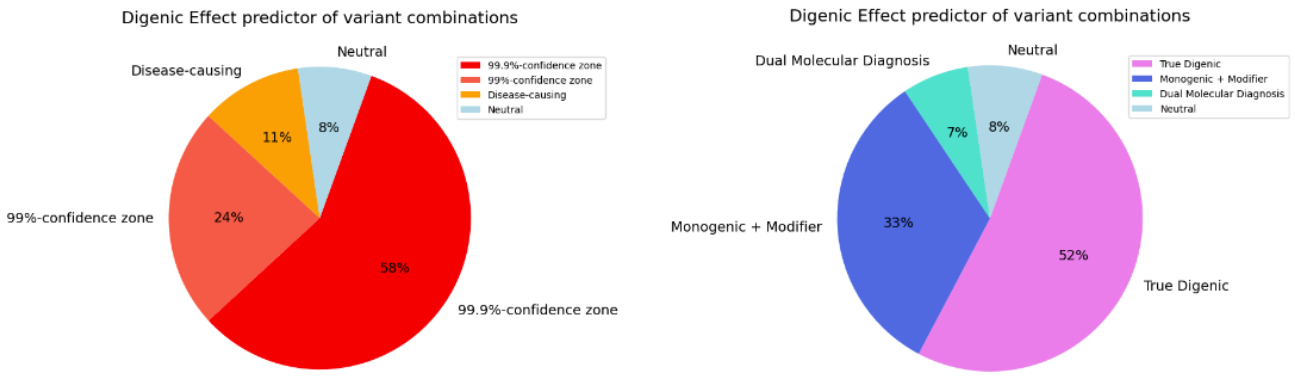

Figura 35. a) Gráfico de torta con los porcentajes de combinaciones correspondientes a su categoría según score de VarCoPP. b) Gráfico de torta con los porcentajes de modelos de herencia asociados a las combinaciones en caso de contar con un score de VarCoPP compatible con patogenicidad, de lo contrario solo aparece representado el porcentaje de combinaciones neutrales en celeste.

También resulta interesante, en este caso, observar cuáles son las contribuciones que los diferentes genes realizan a las diferentes categorías de combinaciones, particularmente las relativas a los distintos modelos de herencia. Esto se encuentra presentado en la Figura 36.

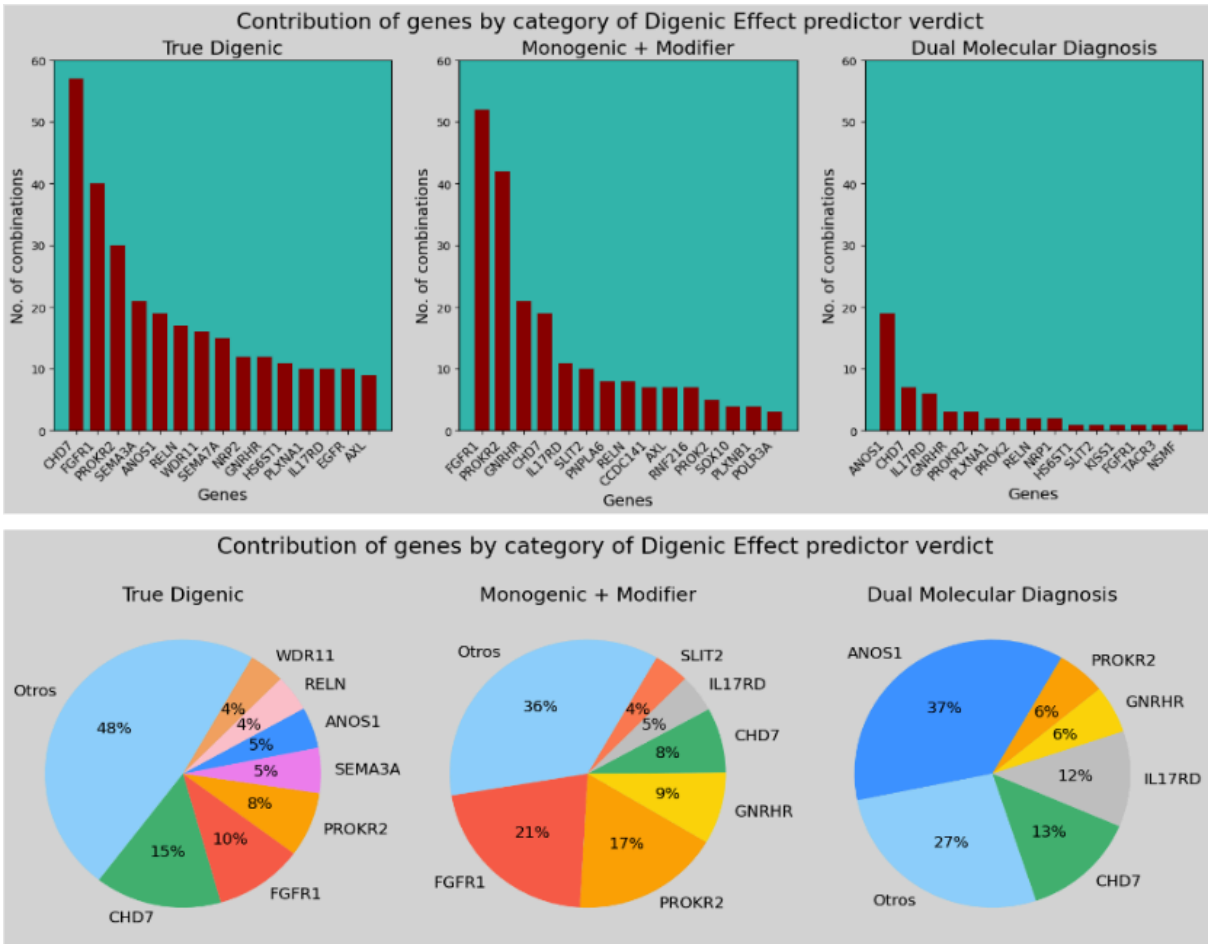

Figura 36. En la parte superior de la imagen se observan sendos gráficos de barras discriminados por modelo de herencia representando cuáles genes contribuyen con mayor número de apariciones en combinaciones con veredicto patogénico según VarCoPP. En consonancia, en la parte inferior de la imagen aparecen sus respectivos gráficos de torta con los porcentajes de apariciones en combinaciones de estos genes.

En general, observamos que existía una relación entre reporte de variantes en genes y reportes de sus combinaciones, aunque no eran mayoría los casos que hablaban explícitamente de oligogenicidad.

Una hipótesis que surgió durante el análisis, era la de si por casualidad no pudiera revelarse un patrón por ejemplo desde el agrupamiento de genes en los clusters expuestos en la Figura 27. Para eso, definimos un conjunto de 26 clusters y confeccionamos un heatmap para apreciar, en la Figura 37, cómo se daban las combinaciones en función de los clusters a los cuales pertenecían los genes participantes de las mismas.

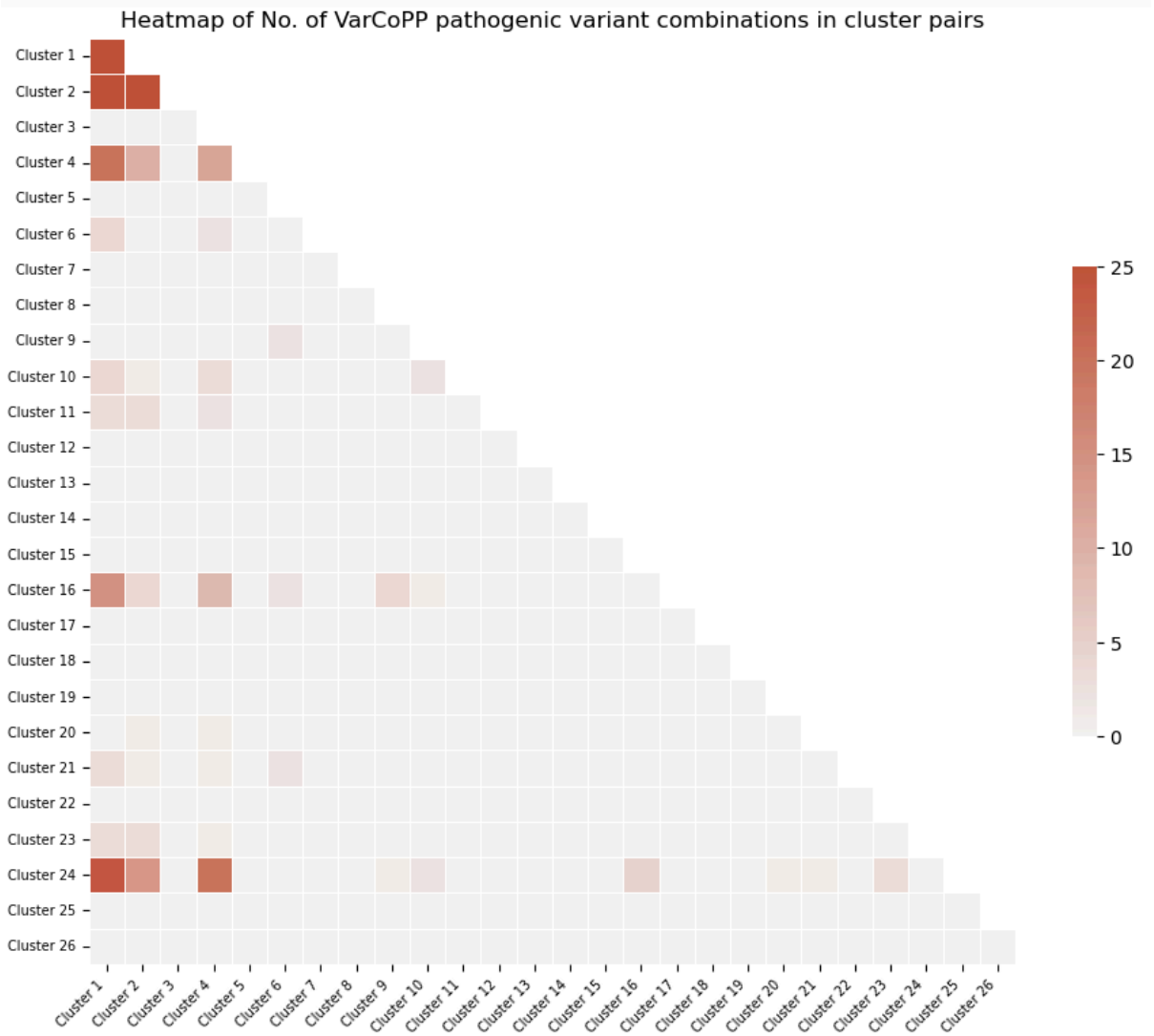

Figura 37. Heatmap donde se observa el número de combinaciones entre clusters integrados por los genes tal y como se observan en la Figura 27. La intensidad de color rojo es proporcional al número de combinaciones reportadas en pacientes para los dos clusters interseccionados en cada celda.

De nuevo nos encontramos con un patrón similar donde clusters con genes con mayor cantidad de apariciones presentaban mayor cantidad de combinaciones. Se realizó un test de Spearman para evaluar la correlación existente entre el reporte de genes en determinados clusters y el reporte de combinaciones. Este entregó un coeficiente de correlación de Spearman de 0.88 con un valor p de 7,02e-09.

Por otro lado, y para finalizar, queríamos dejar plasmado que el patrón de combinaciones entre genes no cambia apreciablemente cuando comparamos contra el heatmap incluyendo solo combinaciones patogénicas según VarCoPP (Figura 38).

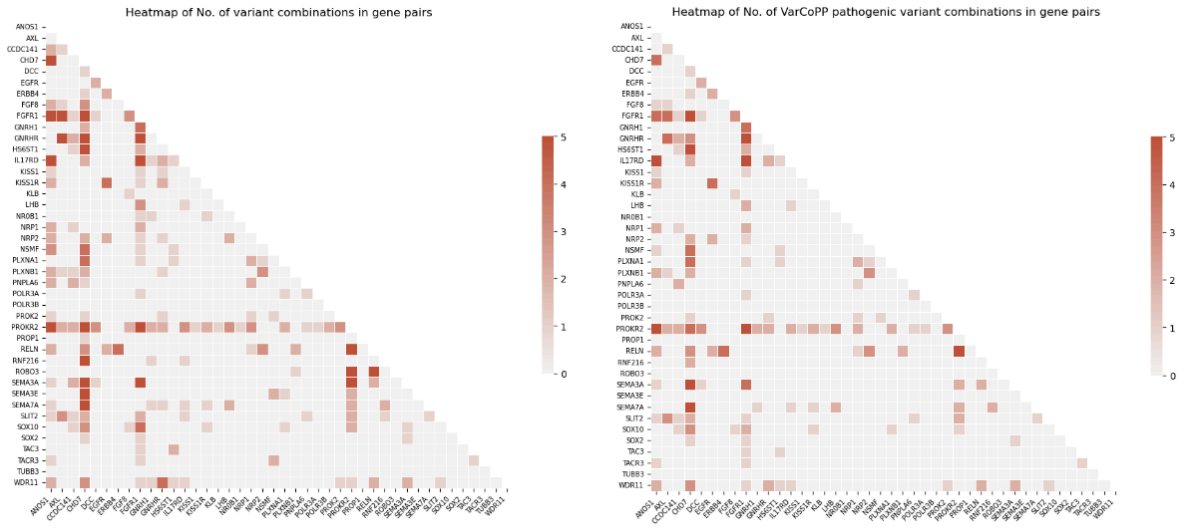

Figura 38. Comparación entre heatmaps generados para las combinaciones entre genes reportadas en la revisión sistemática. A la izquierda vemos el heatmap ya planteado en la Figura 31, mientras que a la derecha se muestra el histograma análogo pero incluyendo sólo combinaciones scoreadas como patogénicas por VarCoPP.

Además de los análisis propios de la revisión que planteamos previamente, también pensamos en desarrollar algunas herramientas a partir de nuestros estudios en lo respectivo a arquitectura y estructura de las proteínas involucradas. Por ejemplo, desarrollamos scripts en lenguaje R a fines de plasmar gráficamente la localización de variantes missense en un esquema unidimensional (es decir, a nivel secuencia) como podemos ver en la Figura 39a. A su vez, dado que la información de la revisión sistemática es solo una parte del saber general, y dado que es una inquietud habitual del analista saber sobre la presencia de variantes en bases de datos de variantes de interés clínico como ClinVar, o de frecuencia alélica poblacional como gnomAD, hicimos lo correspondiente en la Figura 39b. No es una sorpresa que el esquema nutrido con información de la revisión sistemática se encuentre enriquecido en variantes missense patogénicas mientras que la correspondiente a ClinVar/gnomAD carga con una cantidad superior de variantes missense benignas.

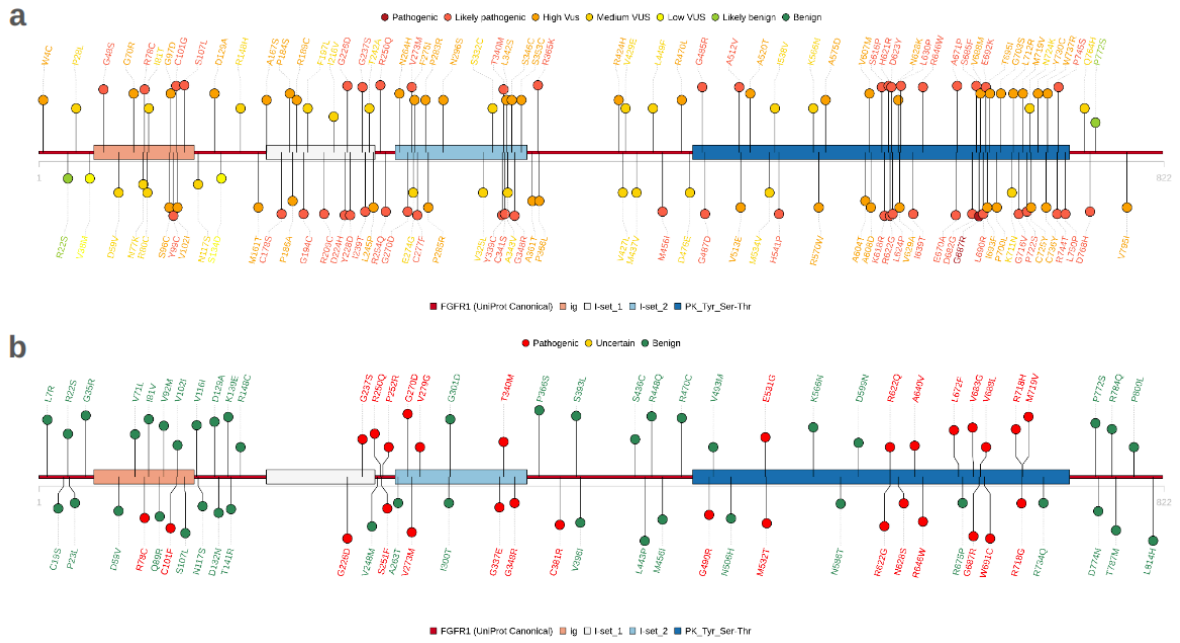

Figura 39. a) Representación unidimensional del receptor codificado por el gen *FGFR1* (tomado a modo de ejemplo). Las cajas de colores representan los diferentes dominios localizados a partir de Pfam. Las variantes se encuentran clasificadas en las diferentes categorías de patogenicidad en función de su probabilidad a posteriori de patogenicidad tras haber realizado toda la serie de ajustes según recomendaciones de ClinGen. b) Misma representación pero utilizando todas aquellas variantes relevadas en ClinVar y gnomAD.

Si bien las dos visualizaciones lineales muchas veces nos ayudan a comprender problemas relativos a la susceptibilidad de determinadas regiones de las proteínas a la variabilidad, como bien sabemos es una de las representaciones más distantes para modelar un problema a todas luces tridimensional. Por eso mismo, elaboramos representaciones 3D donde localizamos las variantes en cuestión, como se muestra en la Figura 40.

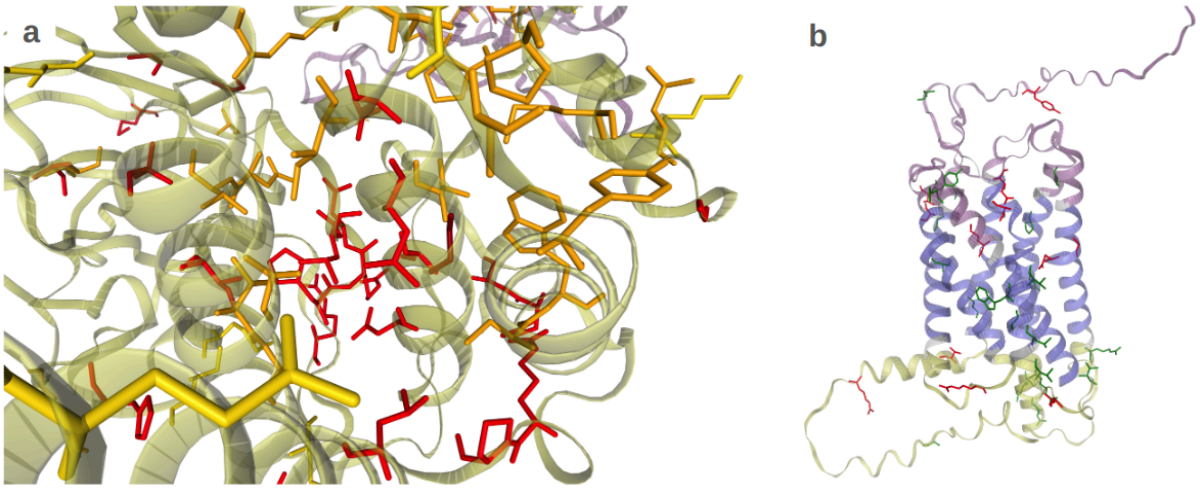

Figura 40. a) Representación tridimensional de parte del dominio kinasa de FGFR1, en las posiciones donde se localizan variantes relevadas en la revisión se hace representación licorice, el color de los aminoácidos sigue el código de la Figura 39 (P en rojo oscuro, LP en rojo, High VUS en naranja, Medium VUS en dorado, Low VUS en amarillo, LB en verde claro, B en verde oscuro). b)

Representación análoga de la proteína PROKR2, se colorea en violeta los segmentos extracelulares, en azul los segmentos transmembrana y en amarillo los segmentos intracelulares. En rojo se muestran las variantes patogénicas de ClinVar, en verde las variantes benignas de ClinVar y gnomAD.

Si bien hasta este momento nos dedicamos casi enteramente a hablar de conceptos más moleculares de las variantes, en general, al estudiar a un paciente nuevo, nos resulta de interés observar si la misma variante (o variantes cercanas) a una que hipotetizamos como causal ha generado un fenotipo que solapa con el de nuestro paciente. Por ejemplo, en lo que respecto a hipogonadismo hipogonadotrófico, una característica fenotípica comúnmente estudiada es la presencia de alteraciones en el olfato (total o parcial, anosmia o hiposmia respectivamente). Por eso mismo, contar con mapas fenotípicos de las variantes nos será de utilidad, como podemos ver en la Figura 41.

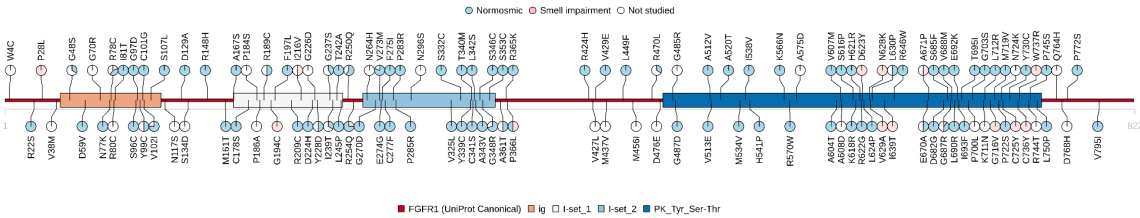

Figura 41. Representación de las variantes relevadas durante la revisión para el gen FGFR1 a nivel de la proteína codificada, los círculos acompañando son gráficos de torta indicando los porcentajes de pacientes que presentan fenotipos normósmicos, anósmicos o hipósmicos.

El valor agregado que nos aportan estas herramientas viene dado por una diferencia sutil: nosotros podríamos perfectamente haber generado visualizaciones análogas basándonos en bases de datos como ClinVar, pero en este caso, nosotros contamos con la curación previa de nuestro equipo clínico de las variantes. Esto refina la hipótesis de trabajo “Existen variantes patogénicas cercanas en secuencia o estructura a mi nueva variante en estudio?” a “Existen variantes con una probabilidad de patogenicidad a posteriori compatible con ser causal de un cuadro de CHH cercanas en secuencia o estructura a mi nueva variante en estudio?”. Recordemos que las asociaciones genotipo-fenotipo no son bidireccionales en todos los casos, y contar con una preselección previa de pacientes con fenotipo compatible con CHH representa una enorme ventaja.

Conclusiones

Este capítulo ofrece diversas reflexiones en lo respectivo al retorno que un esfuerzo de considerable magnitud como realizar una revisión sistemática de variantes genéticas asociadas a una patología puede tener a la hora de implementar procedimientos estándar operativos en un servicio hospitalario especializado en la misma. En primer lugar, observamos cómo conocer la casuística previa de pacientes reportados nos permitió calcular valores de corte de frecuencia alélica poblacional de variantes, lo cual encuentra su mayor utilidad a la hora de realizar tareas de priorización de variantes. Este enfoque facilita el descarte de gran cantidad de las mismas y optimizar el tiempo de análisis. Por otro lado, esto también tiene implicancias a la hora de optimizar la interpretación de las variantes, donde podemos ver cómo un manejo de la base de datos nos permitió ajustar el nivel de evidencia de criterios referidos a pacientes previamente reportados con la patología en forma automatizada según las últimas recomendaciones de ClinGen.

En segundo lugar, dentro del sector académico, la evidencia científica ha sido históricamente valorada en función del año de publicación y de la rigurosidad metodológica con la que fue obtenida. Es frecuente encontrarse con variantes reportadas en genes de incierta relación con la etiopatogenia de la enfermedad, resultando en meras hipótesis de trabajo más que en variantes con relevancia clínica comprobada. Mediante el uso de herramientas de análisis de redes génicas y su topología, logramos confeccionar paneles de alta y moderada confianza de asociación con CHH, además de excluir un subconjunto de los mismos cuya evidencia acumulada es insuficiente.

La oligogenicidad ha sido un concepto recurrente en la literatura sobre CHH durante los últimos años, y numerosos estudios han intentado identificar las variantes implicadas, con resultados variables. Un aspecto relevante es que, en muchos casos, la metodología empleada ha seguido un enfoque mendeliano estricto o, por el contrario, ha resultado excesivamente flexible en la priorización de variantes. Para abordar de manera rigurosa el estudio de la oligogenicidad, es fundamental profundizar en la naturaleza de los genes implicados y en las interacciones epistáticas entre ellos, recurriendo a algoritmos que permitan capturar la complejidad del problema de manera integradora. En este sentido, la comunidad científica debe estar dispuesta a considerar estrategias alternativas cuando el modelo mendeliano no permite identificar variantes potencialmente causales de la patología.

Finalmente, resulta cada vez más evidente la necesidad de contar con especialistas en análisis de grandes volúmenes de datos genómicos y en la interpretación de estructuras cristalográficas de proteínas y ADN dentro de los equipos encargados de la evaluación de variantes genéticas. Disponer de información altamente curada a partir de una revisión sistemática de variantes proporciona un insumo valioso para este tipo de análisis, especialmente en la caracterización de casos más complejos que requieren un mayor nivel de detalle.

Sección 2: Casos clínicos.

En esta sección procederemos a presentar los resultados correspondientes a la cohorte de pacientes del Hospital de Niños Ricardo Gutiérrez con diagnóstico clínico de hipogonadismo hipogonadotrófico y que fueron seleccionados para estudio genético debido a que presentaban algunas de las características fenotípicas orientativas de un cuadro de CHH.

Sobre un total de 39 pacientes, se encontró al menos una variante explicativa del cuadro del paciente en 18 de los mismos, correspondiendo a un 46% (Figura 8a), porcentaje que se encuentra en línea con las experiencias provistas por otros laboratorios similares. En estos pacientes se registró un total de 22 alelos afectados, debido a la presencia de un paciente homocigota para una variante en *GNRHR*, un paciente que presentaba tanto una variante homocigota en *CDH7* como una heterocigosis compuesta en *HESX1*, y a otro paciente portando una combinación oligogénica de variantes en los genes *PROK2-PROKR2*. Los tipos de variantes halladas pertenecen en general a cambios de un único nucleótido o pequeños InDels en regiones codificantes, a excepción de una minoría de casos con variantes intrónicas profundas o grandes deleciones (copy number variants, CNV).

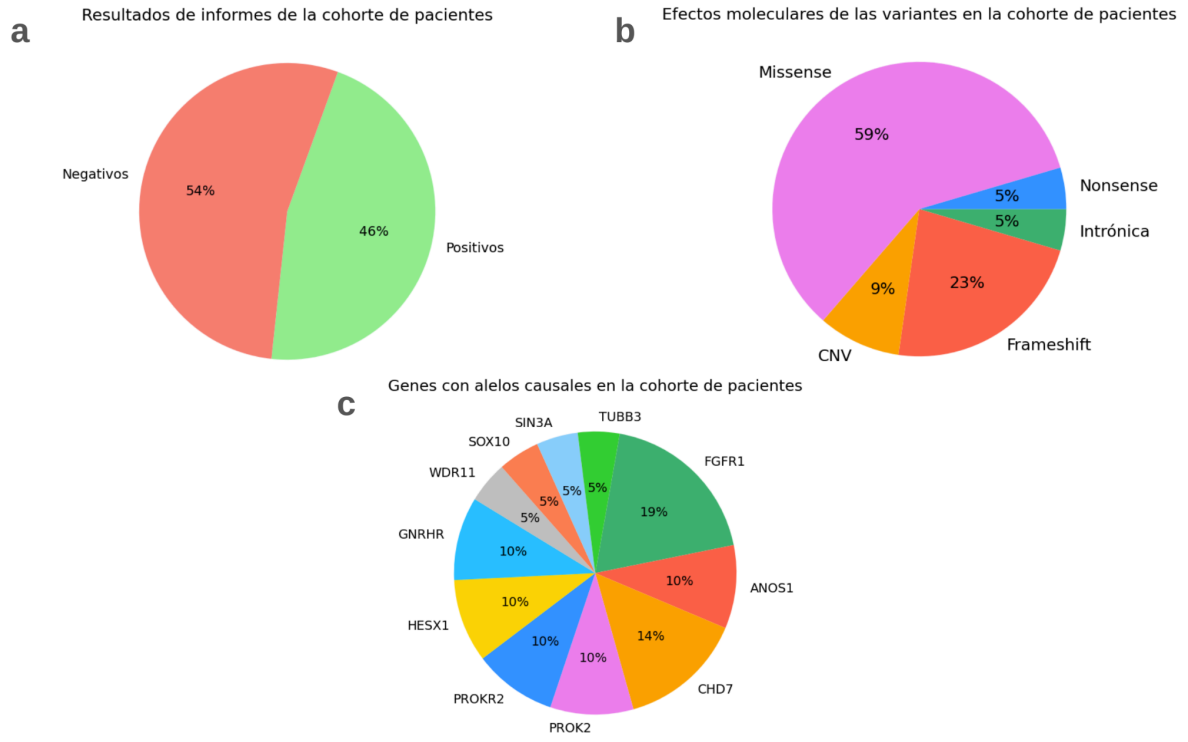

Figura 8: a) Porcentaje de casos donde se lograron resultados compatibles con la definición de informe positivo o negativo. b) Tipos y porcentajes de variantes halladas en la cohorte de pacientes. c) Genes y porcentajes correspondientes a aquellos donde se detectaron alelos causales.

Si bien siempre resulta de sumo interés el estudio general de la cohorte, en este capítulo nos enfocaremos en el estudio del paciente en forma individual, que es donde se encuentra la verdadera riqueza del análisis y donde se ponen de relevancia algunos de los obstáculos más interesantes con los que nos hemos encontrado. En consecuencia, nos dedicaremos a hacer el análisis exhaustivo de cinco pacientes elegidos en función de las

características particulares que mostraban las variantes detectadas en los alelos causales afectados.

Casos de estudio representativos de la cohorte de pacientes

Paciente 1.

El primero de los casos que tomamos como ejemplo interesante de los alcances de los análisis llevados a cabo en este trabajo es el de un paciente pediátrico que se presentó por primera vez en nuestro servicio a la edad de 7 años. Si bien no contaba con antecedentes familiares de relevancia, sus antecedentes personales de criptorquidia bilateral refractaria a tratamiento con hCG, finalmente resuelta por orquidopexia, sumado a perfiles hormonales compatibles con hipogonadismo hipogonadotrófico ya por la edad de 12 años y una ausencia de visualización de bulbos olfatorios en estudios de imágenes, terminaron derivando en la indicación de realización de un estudio genético.

En el mismo, se priorizó una variante NM_023110.3(FGFR1):c.1862A>C (p.His621Pro) en heterocigosis localizada en el gen que codifica para el receptor del factor de crecimiento de fibroblastos de tipo 1. Dicha variante se encontraba ausente en una base de datos de frecuencia alélica poblacional como lo es gnomAD (PM2_Supporting). La localización en la arquitectura de dominios proteicos de esta variante la ubicaba en el dominio Protein tyrosine and serine/threonine kinase de Pfam (PfamID: PF07714.20), el cual es fundamental a la correcta señalización mediada por el receptor. Debemos considerar la naturaleza del cambio missense en cuestión, donde la inserción de una prolina, con todas sus características estructurales particulares dada su constitución cerrada de anillo causante de giros bruscos del backbone, sumado a la ausencia de una histidina (dado su pKa, hablamos de un aminoácido cargado positivamente alrededor del 50% del tiempo al pH celular) en el contexto de un dominio que naturalmente tiene necesidad de estabilizar cargas negativas de los fosfatos del ATP, nos da motivos para pensar en un posible efecto deletéreo en la estructura y por ende en la función (PM1). Esto va en consonancia con el hecho de que la variante missense contaba con una predicción altamente deletérea de un metapredictor bioinformático de amplio uso como lo es REVEL (PP3_Strong), sumado a la altísima conservación que presentaba la histidina en el AMS que define al dominio Pfam correspondiente (Figura 9b y 9c). Su localización al interior del dominio kinasa hacía pensar en un rol directo en la catálisis enzimática, por lo que uno de los recursos a los cuales se fue a consultar fue al Catalytic Site Atlas (CSA), base de datos especializada específicamente en el rol que determinados residuos de aminoácido poseen en la catálisis. Para nuestra sorpresa la histidina en cuestión no estaba documentada como participante de la catálisis, no obstante lo cual, nos encontramos que otros cuatro residuos se encontraban extremadamente vecinos tridimensionalmente (cercanos en el espacio) y sí participaban: D623, R627, N628 y D641 (Figura 9a). Dado que la inserción de una prolina al interior del sitio activo del dominio kinasa por seguro trae como consecuencia una disposición tridimensional anómala entre estos residuos, esto apoya plenamente nuestra hipótesis de esta variante como causa del fenotipo presentado por el paciente.

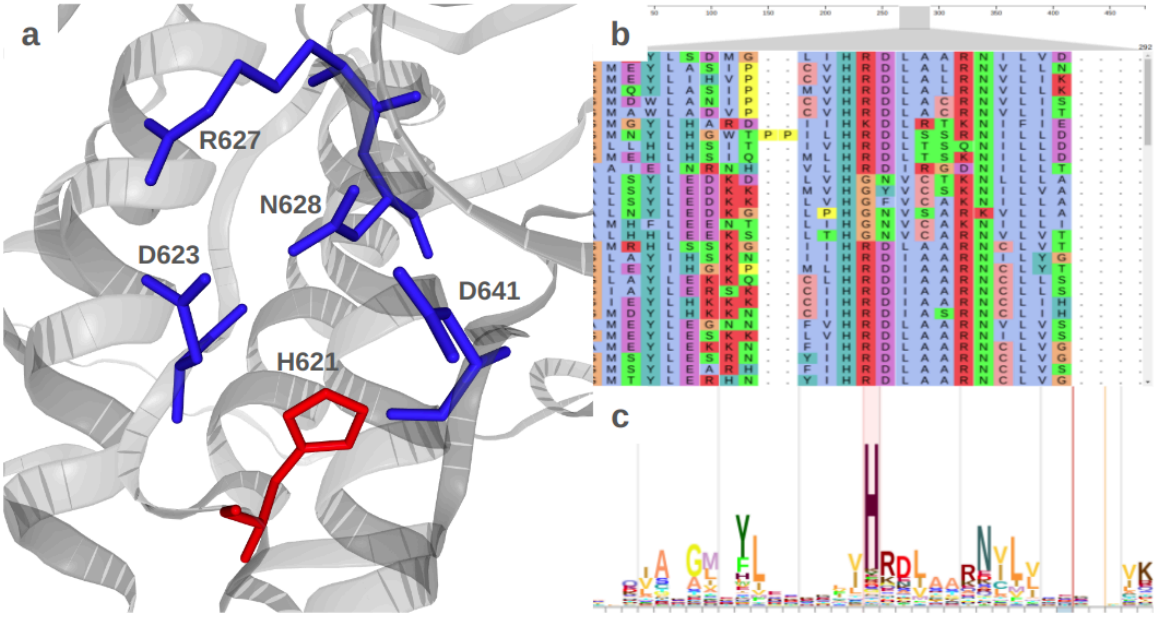

Figura 9. a) Visualización 3D del sitio actiVo del dominio kinasa de FGFR1. En azul, residuos documentados en CSA como relevantes a la catálisis enzimática. En rojo figura la histidina afectada por la variante. b) AMS seed sobre el cual se definió el dominio kinasa de Pfam, en el centro se observa la posición correspondiente a la histidina en color cian. c) HMMlogo de Pfam donde se evidencia la gran conservación de la histidina.

Dentro de los objetivos de esta tesis siempre se encontró muy presente el realizar predicciones en lo respectivo a los efectos sobre la estabilidad estructural que una determinada variante pudiera ocasionar. En este sentido, la utilización de software como FoldX para realizar cálculo termodinámico y evaluar dicha estabilidad por medio de estimar un cambio en la energía libre de Gibbs para la reacción de plegamiento del polipéptido es de gran ayuda. Para esto, nos dirigimos a la base de datos del Protein Data Bank (PDB) a encontrar estructuras cristalográficas, pero no solo del receptor codificado por el gen FGFR1, sino de toda proteína que contara con un dominio kinasa tal como define Pfam para nuestra proteína de interés, y realizamos el modelo del cambio missense y la posterior evaluación del impacto en la estabilidad. Este cálculo realizado sobre 10 estructuras cristalográficas que presentaban al menos un 70% de cobertura del dominio kinasa y por lo menos un 60% de identidad de secuencia con el dominio correspondiente de FGFR1, arrojó un valor de ΔΔG de 1,93 ± 1,49 kCal/mol con una mediana de 2,15 kCal/mol. Basándonos en valores de corte puestos por nuestro grupo, nuestros resultados se muestran altamente compatibles con la hipótesis de que nuestra variante impacta directamente en la estabilidad estructural del dominio kinasa del receptor, y que por ende generaría una proteína inestable, lo que derivaría en las consecuentes pérdida de función y etiopatogenia observadas.

Por otro lado, resulta interesante mencionar que al inspeccionar la base de datos gnomAD, donde ya mencionamos que la variante no se encontraba reportada, nos encontramos con que, en este receptor, el conocido Z-score de variantes missense indica que en la proteína existe un desbalance significativo en la relación entre las variantes observadas y esperadas de este tipo. El mismo entrega un valor de 4,28, muy por encima del valor de corte de 3,09 propuesto por ClinGen en sus recomendaciones (esto permite asignar la etiqueta PP2).

Por último, es de destacar que se ha reportado una variante missense diferente en la misma posición (p.H621R) en un paciente con Sindrome de Kallmann, la cual podemos clasificar conservadoramente como Probablemente Patogénica según ACMG (lo que nos permite agregar a nuestra variante la etiqueta PM5). La variante fue corroborada por Sanger en estado heterocigosis en el paciente, mientras que por la misma técnica se observó su ausencia en los padres, todo en el contexto de un fenotipo específico para el gen (PM6).

El conjunto de toda esta evidencia nos permite aplicar la clasificación ACMG de variante Probablemente Patogénica (con una probabilidad de patogenicidad a posteriori (Post_P) de 0,988). Cabe notar que en el cálculo de esta probabilidad a posteriori solamente se consideró un aporte equivalente a Strong entre las etiquetas PM1, PP2 y PP3 pues se encuentran inherentemente enlazadas por una cantidad de supuestos comunes.

Caso 2.

El segundo de los casos que abordaremos en profundidad es el de un paciente pediátrico que desde los primeros meses de vida ya exhibía un tamaño de pene varios desvíos estándar por debajo de lo esperado para su edad. Al realizar dosajes hormonales durante su mini-puberty (3 meses), sus valores disminuídos de LH, T y AMH resultaban compatibles con un CHH. Fue tratado con testosterona durante varios períodos de su vida con buena respuesta, pero llegando a sus 15 años se apreció que su peso, talla y velocidad de crecimiento no eran los adecuados. Se indicó la realización de una prueba de infusión de GnRH donde se observó que, si bien su nivel basal y respuesta de FSH eran normales, no lo era así su respuesta de LH. Se realizó el estudio de imágenes de SNC el cual arrojó que existía una marcada disminución de volumen del bulbo y cintilla olfatoria en ambos lados. El paciente refería que tanto él, como su padre, sólo percibían olores fuertes, compatible con hiposmia. En función de esta información clínica, se indicó la realización del estudio genético del paciente el cual arrojó una variante NM_018117.12:c.163dup en el gen WDR11 en estado de heterocigosis. La misma no se encontraba reportada en bases de datos de frecuencia alélica poblacional como gnomAD (PM2_Supporting). Esta duplicación de una base nucleotídica conlleva una pérdida del marco de lectura habitual, presuponiendo una pérdida de función en este gen, que presenta un score pHaplo de 0,76, compatible pero no definitorio en cuanto a su sensibilidad a la pérdida de función. No obstante, la presencia de ensayos funcionales y pacientes reportados previamente con variantes similares permiten establecer con confianza la pérdida de función como mecanismo de enfermedad. Siguiendo las recomendaciones de ClinGen, dado que esta variante frameshift se localiza en el exón 2/29 y no escapa al conocido mecanismo de ARNm-NMD (Figura 10) según lo descrito en literatura, podemos plantear firmemente la hipótesis de que el producto generado por este gen se someta a este mecanismo de control del ARNm (PVS1). Finalmente, se comprobó la presencia de la variante en el paciente por medio de la técnica de Sanger, como así también en su padre, no siendo así en su madre donde se encontraba ausente.

El conjunto de toda esta evidencia según los criterios de ACMG nos lleva a una probabilidad a posteriori de patogenicidad de 0,988, la cual nos permite asignar una clasificación de la variante de Probablemente Patogénica. Resulta interesante notar que la combinación de evidencia PP + PVS1 no nos hubiera permitido llegar a la misma clasificación en el esquema de Richards 2015.

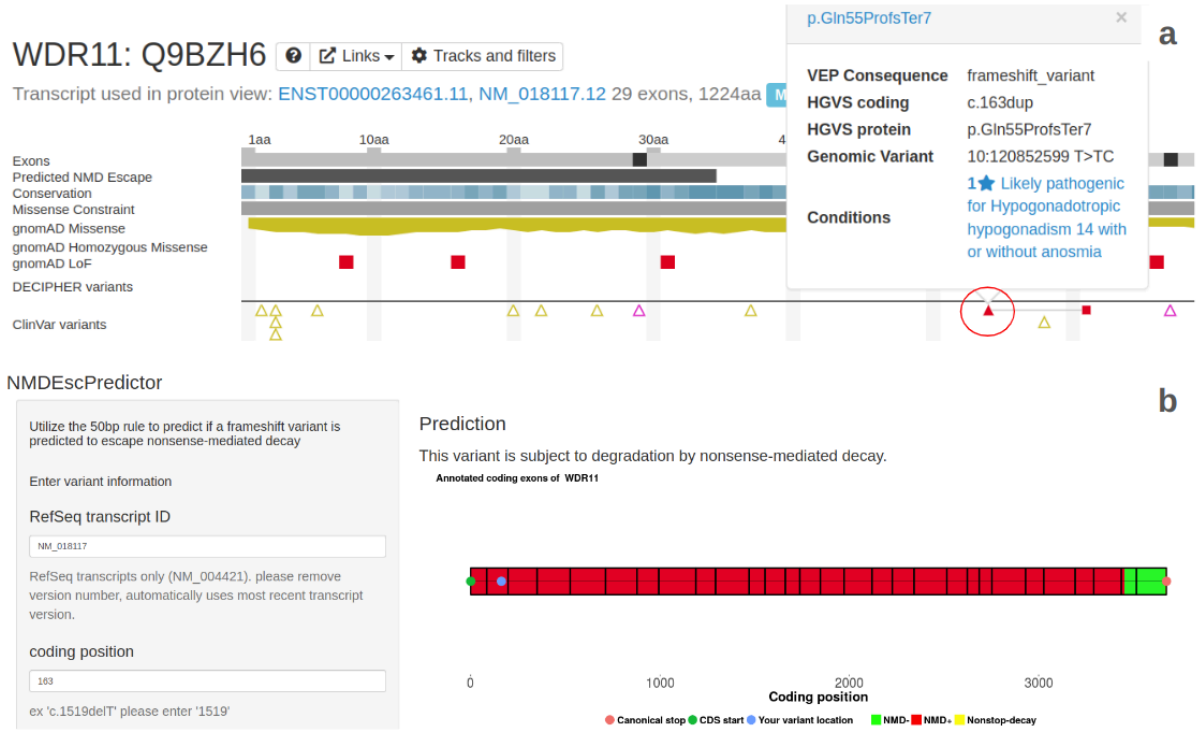

Figura 10: a) Visualización de la localización de la variante (triángulo rojo, encerrada en un círculo del mismo color), apreciándose que se encuentra por fuera de la zona de escape de NMD. b) Visualización de la variante en el exón 2/29 del gen WDR11 y confirmación de la predicción de NMD por el predictor NMDEscPredictor.

Caso 3.

El tercero de los casos que presentamos corresponde a un paciente que consulta por primera en el Servicio de Endocrinología del Hospital Gutiérrez a la edad de 15 años y 6 meses, derivado desde otro servicio para realizar un abordaje multidisciplinario en base a un diagnóstico clínico de hipogonadismo hipogonadotrófico. Sus antecedentes neonatológicos indicaban que había experimentado una taquipnea sin requerimientos de O2, que al ser investigado por electrocardiograma develó una coartación de aorta y una comunicación interventricular pequeña. El fenotipo que quedó registrado en su historia clínica refería tener frente en tobogán, hendidura antimongoloidea, micrognatia, orejas pequeñas y de implantación baja, pterigium colli, pie bot, sindactilia en los segundos y terceros dedos de los pies, entre otros. No se registraban antecedentes familiares de relevancia. A los 9 meses se pone en evidencia la presencia de hipoacusia bilateral, severa en el oído derecho, leve-moderada en el izquierdo. Se pone de manifiesto una criptorquidia bilateral, que no verá resolución hasta los 7 y 12 años cuando se realicen sendas orquidopexias. Para sus 10 años, estudios bioquímicos mostraron bajos niveles tanto de FSH como de AMH. Este cuadro general se mantiene hasta el momento de la consulta en el servicio, donde viene derivado debido a la combinación de pequeño tamaño testicular (derecho con volumen menor a 1 ml, izquierdo en 1-2 ml, ambos en bolsa) con antecedentes de criptorquidia corregida, a lo cual se sumaba como novedad una escoliosis marcada, hipertelorismo, retraso global del desarrollo, trastorno de ansiedad, pero con un olfato normal por anamnesis (aunque luego el estudio de imágenes mostró agenesia de vía olfatoria con signos de displasia septo-óptica). Esta combinación de fenotipo cardíaco, auricular, genital y del desarrollo, dieron fundamento a la sospecha de un posible Síndrome

CHARGE. En función de estas observaciones, se indicó la realización del estudio genético del paciente el cual arrojó efectivamente una variante NM_017780.4:c.3236C>A (p.Ala1079Asp) en el gen CHD7 en estado de heterocigosis.

Esta variante constituyó todo un desafío, además de una prueba de la rigurosidad de la capacidad de análisis del grupo de trabajo. Este gen que codifica una helicasa de un tamaño colosal (UniProtID: Q9P2D1, 2997 aas) suele arrojar siempre en los estudios de NGS un amplio número de variantes, consecuencia en mayor medida de la enorme extensión del mismo, que de la presencia de variantes causales de cuadros de CHH o de Síndrome de Charge. No obstante, en un análisis convencional la mayoría de las mismas son descartadas rápidamente, no siendo este el caso de la variante en cuestión. En este sentido resulta interesante comenzar analizando la estructura de dominios que presenta CHD7. Como podemos ver en la Figura 11a, pese a su gran extensión, son solamente ciertos fragmentos los correspondientes a dominios bien establecidos (como aquellos que define Pfam), mientras que el resto de la secuencia se encuentra, según métodos de análisis de secuencia, cubierta por regiones desordenadas o enriquecidas en algún tipo de residuo de aminoácido. No resulta sorprendente entender entonces que la mayoría de los falsos positivos se encuentran precisamente en estas últimas regiones. La variante que nos dispusimos a analizar se localizaba en el dominio “SNF2-related domain” (PfamID: PF00176). Este tiene asociada la función de aprovechar la energía proveniente de la hidrólisis de ATP para lograr una disrupción de la interacción ADN-histona, clave para la actividad de CHD7 de helicasa. Esta variante se encuentra ausente en bases de datos de frecuencia alélica poblacional como gnomAD (PM2_Supporting).

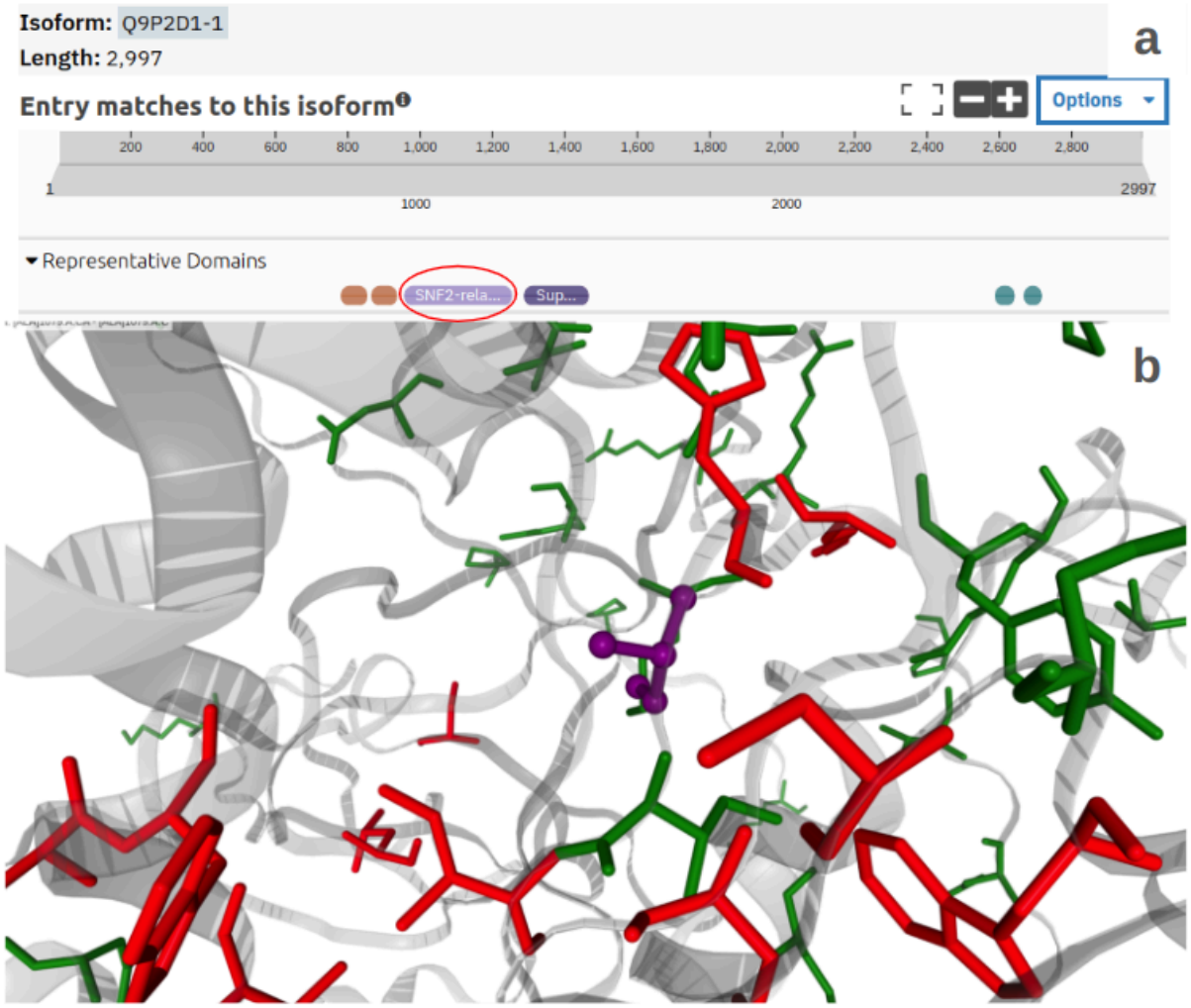

Figura 11: a) Visualización de la arquitectura de dominios de CHD7 según InterPro. b) Visualización del residuo original de alanina (púrpura) en el cristal generado por AlphaFold2, podemos apreciar en rojo aquellos residuos que cuentan con variantes patogénicas reportadas en ClinVar, mientras que aquellas benignas se observan en verde.

Una de las primeras características que analizamos a nivel residuo fue la conservación, donde por medio del algoritmo de Divergencia de Jensen-Shannon apreciamos que el residuo se encontraba dentro del tercer decil de los residuos más conservados. Esta evidencia si bien no es muy concluyente sugiere posible patogenicidad. Por otro lado, el análisis del HMMLogo de Pfam junto al AMS que le dio origen, nos permitió observar que si bien la alanina no se mostraba evolutivamente invariable, en la posición en cuestión y un par adyacentes encontrábamos invariablemente residuos hidrofóbicos. Esto reforzaba nuestra teoría de que el cambio por un aminoácido cargado negativamente como el aspartato, contara con grandes chances de generar una inestabilidad de la estructura de plegamiento. Esta hipótesis se vio reforzada por el análisis de FoldX que derivó en un valor de ΔΔG de 2,23 kCal/mol. También empleamos técnicas de visualización de estructura tridimensional de proteínas haciendo uso de la predicción generada por AlphaFold2 y mapeando sobre la misma al residuo en cuestión como así también aquellos sobre los cuales se han reportado variante de tipo missense patogénicas o benignas. Como podemos apreciar en la Figura 11b, nuestro residuo se encuentra en un entorno donde una gran mayoría de variantes han sido reportadas como patogénicas, y por fuera de este espacio

(por detrás en la imagen), se observa un fondo de variantes benignas más alejadas. El conjunto de toda esta información nos permitió plantear con firmeza que el residuo se encontraba en una región, de mínima, sensible a la variación missense, lo cual coincide con que el gen cuente en gnomAD con un Z-score para variantes missense elevado (PP2). En un rol conservador, no podemos asegurar su rol funcional pues se encuentra alejado de las regiones de interacción con el ADN o el ATP.

Por otro lado, nuestra variante contaba con una predicción patogénica por REVEL de 0,87 (PP3_Moderate). Y en apoyo de estas apreciaciones moleculares y estructurales, se puede observar que este es un gen donde existe un constraint en materia de variantes missense, lo cual ClinGen en general ya cuenta como evidencia. Sin embargo, conociendo de la naturaleza ya mencionada de este gen/proteína en cuanto a su extensión y arquitectura, puntualizando en la región donde se encuentra la variante, se puede apreciar que existe un constraint local de variantes missense en la región de 1kb rodeando a nuestra variante.

La variante fue confirmada por Sanger en el paciente y se confirmó la ausencia en sus padres por medio de la misma metodología, siendo el fenotipo asociado al gen en cuestión de alta especificidad (PM6). El fenotipo del paciente era compatible y específico de Síndrome de Charge, asociado tradicionalmente a variantes en el gen *CHD7* (PP4). El conjunto de toda esta evidencia según la ACMG nos lleva a una probabilidad posterior de patogenicidad de 0,949, la cual nos permite asignar una clasificación de Probablemente Patogénica.

Caso 4.

El cuarto de los pacientes que traemos a estudio corresponde a un individuo al cual se le realizó seguimiento en el Hospital de Niños Ricardo Gutiérrez durante años, pero al cual nunca se le había podido brindar un diagnóstico molecular certero. El caso era de especial interés pues el familiograma resultaba extremadamente florido, comenzando por el hecho de registrarse la presencia de dos hermanos menores con un cuadro fenotípico análogo, además de otros individuos masculinos (primo y tíos-abuelos), que resultaban en la sospecha de un modelo de herencia ligado al X (Figura 12a). Su primera consulta en el servicio se había dado a los 16 años por un retraso puberal, donde se había constatado micropene y criptorquidia bilateral. Se realizó un análisis bioquímico donde se evidenciaron niveles disminuídos de LH, FSH, Testosterona y AMH. La prueba de infusión de GnRH confirmó el diagnóstico de hipogonadismo hipogonadotrófico, momento en el cual se le solicitó ADN para realizar estudios genéticos. Años después su familia volvió a consultar al hospital ya sin el caso índice pero con sus familiares, informando que su criptorquidia no había sido resuelta, que su trastorno de olfacción había sido confirmado por olfatometría y que la administración de tres dosis de testosterona había sido seguida de cambios favorables en su desarrollo. Tanto sus hermanos como su primo habían cursado con cuadros de hipogonadismo hipogonadotrófico, criptorquidia, pubertad retrasada y anosmia, altamente compatible con un diagnóstico clínico de Síndrome de Kallmann. A estos pacientes, se les solicitó también material genético para realizar estudios moleculares. Sin embargo, la primera vez que se realizó el estudio molecular y se interrogaron variantes de tipo SNP o pequeños InDel, no se encontraron variantes causales. No obstante lo cual, en aquel momento no pasó desapercibido que existía un grupo de exones que al revisar las métricas de rendimiento del experimento de secuenciación, sugerían un desempeño insuficiente del proceso. Fue al lograrse la implementación de los algoritmos de detección

de CNVs a partir de datos de exomas/paneles obtenidos por NGS (en nuestro caso, se utilizó el software DECoN) que se logró evidenciar una deleción hemicigota en el cromosoma X que abarcaba a los exones 4, 5 y 6 del gen *ANOS1*. La inspección del archivo BAM permitió visualizar con gran nivel de detalle los puntos de corte por los cuales se dio este evento (Figura 12b).

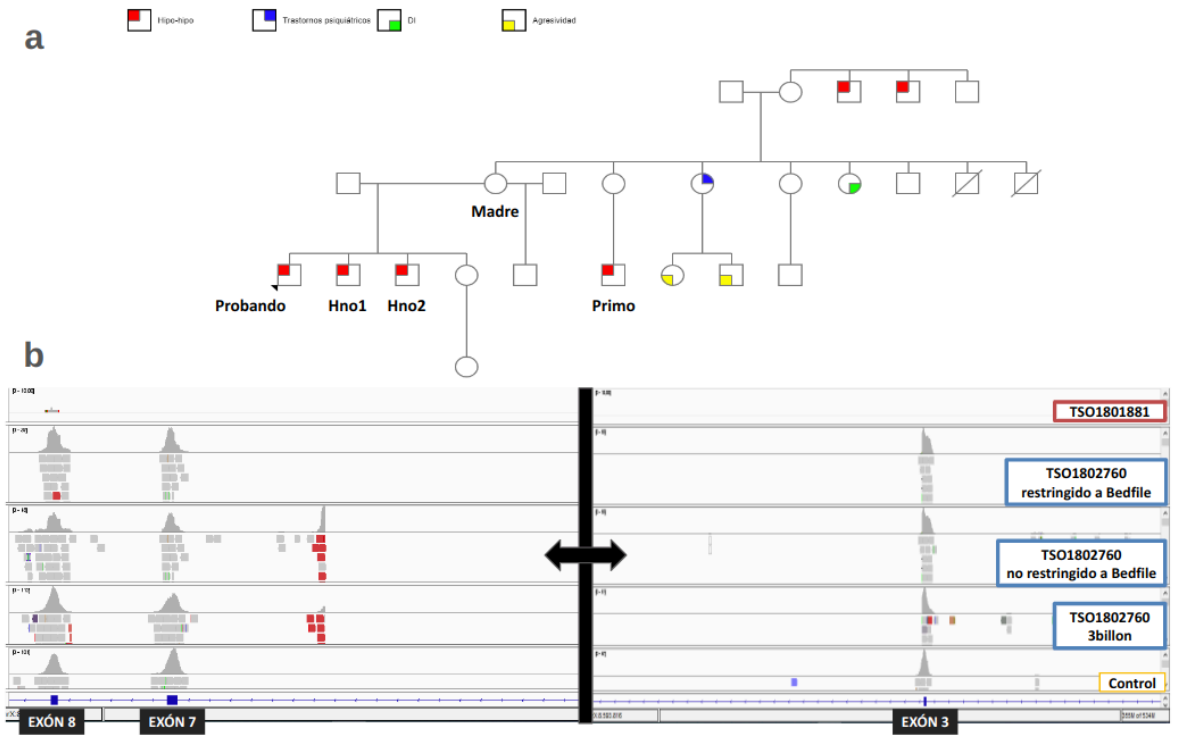

Figura 12: a) Familiograma del caso en estudio. b) Visualización del archivo BAM con el programa IGV en la región de interés del gen *ANOS1*, el orden de los exones se encuentra invertido dado que el gen se encuentra en hebra reverse. TSO1801881 corresponde a la muestra de un paciente con una deleción total del gen *ANOS1*. TSO1802760 corresponde a la muestra de nuestro paciente, se muestra tanto las lecturas comprendidas dentro del kit de captura utilizado (restringido a Bedfile) como por fuera (no restringido a Bedfile). Se presenta también la muestra procesada por un tercero (empresa 3billion). Se agrega una muestra control, proveniente de un paciente afectado por una patología diversa en otro gen.

Analizando la Figura 12b, se puede observar que la región inmediatamente a continuación del exón 3 registra un patrón de lecturas apiladas pero prácticamente cortadas río abajo, indicando que hasta allí dichas lecturas lograron ser mapeadas, pero solo en un fragmento de las mismas. Por otra parte, antes de llegar al exón 7 se registra un cúmulo de lecturas con una baja calidad de mapeo (en color rojo) sobre el intrón, correspondientes a la continuación o fragmentos remanentes de aquellas lecturas cortadas observadas en el exón 3. Este patrón de lecturas “cortadas” es un fuerte indicio de la presencia de una deleción. Con estas informaciones probablemente podríamos haber hecho una muy buena estimación de los puntos de ruptura pero dado que DECoN no está indicado de primera línea a fines diagnósticos, en el laboratorio se acordó seguir adelante con una técnica confirmatoria. Si bien el MLPA es la técnica de biología molecular más recomendada y utilizada para detectar y corroborar presencia de deleciones de fragmentos compatibles con el tamaño de exones promedio, sabiendo que se tenía una buena estimación del tamaño y ubicación de la misma, se procedió a realizar un diseño de PCR clásica para evidenciarla (Figura 13).

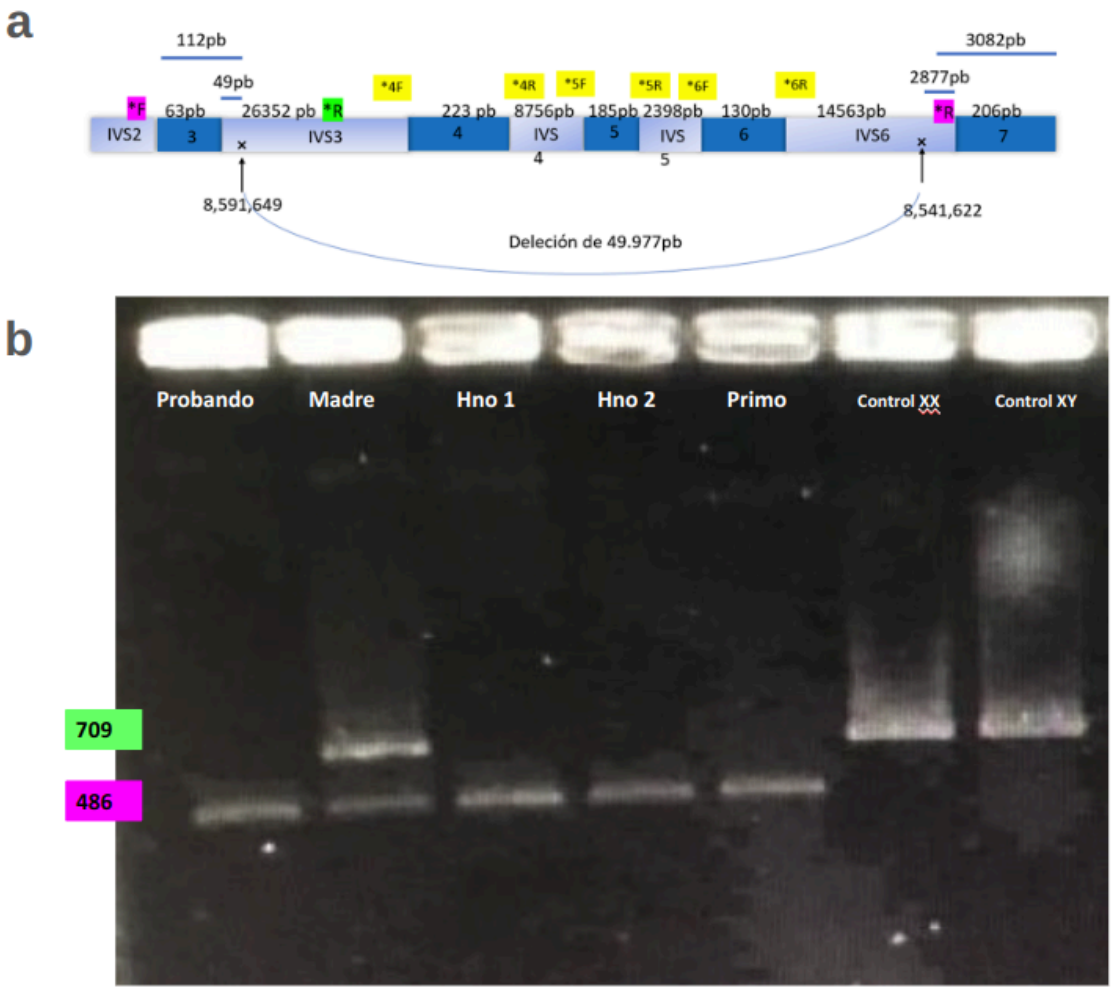

Figura 13: a) Esquema del diseño del experimento de PCR para evidenciar la presencia de la deleción. Las posiciones cromosómicas refieren a la versión 37 del genoma humano de referencia. De encontrarse la deleción se genera un producto de tamaño 486pb, de lo contrario el producto generado es de 709pb. b) Gel de agarosa donde se aprecia los resultados para los individuos de interés.

El experimento de PCR permitió evidenciar que el probando, sus dos hermanos y su primo, todos afectados, presentaban la deleción en estado hemicigota, mientras que la madre del probando era heterocigota para la deleción. La nomenclatura final del CNV por ISCN es arr(GRCh38) Xp22.31(8573581_8623608)x0 mientras que por HGVS será NC_000023.11:g.8573581_8623608del. La clasificación de CNVs detectados en el ámbito clínico sigue una serie de recomendaciones diversa y diferente de aquella utilizada para las variantes de tipo SNP o InDel. No obstante lo cual, siempre es importante recordar que existe un punto de unión fuerte entre ambas, principalmente a nivel del universo de variantes de pérdida de función. En el caso en estudio, podemos apreciar que los puntos de ruptura de la deleción se encuentran dentro del gen, lo cual en el algoritmo de análisis de CNVs encaja en el punto 2E y nos envía a la recomendación de clasificación de variantes de pérdida de función de acuerdo a los criterios de la ACMG. En la misma, ubicándonos en el grupo de deleciones multiexónicas que forman parte de transcriptos biológicamente relevantes, nos permite asignar una evidencia PVS1. De retorno al algoritmo para CNVs, esto nos permite asignar +0,90 puntos, los cuales ya nos dejan en posición de una

clasificación Probablemente Patogénica. Si a esto le sumamos que en esta familia se aprecia un número no inferior a 5 de segregaciones, y que en el algoritmo se asignan 0,15 puntos por segregación (punto 5D), esto automáticamente nos lleva al máximo score de 1 obteniéndose una clasificación final de Patogénica para la variante.

Caso 5.

Por último, el quinto caso corresponde a un paciente que se presentó a la edad de 15 años y 6 meses, con motivo de un escaso desarrollo genital. El mismo había sido evaluado en su infancia por microorquidismo, y aparte de broncoespasmos y alergias, no presentaba otros antecedentes perinatales de relevancia. Su evaluación mostró una genitalia con estadío Tanner G1, con testículo derecho en 1 ml de volumen y el izquierdo en 2 ml, con VP3. Sus niveles de LH, FSH y T eran bajos. Refería anosmia ante el interrogatorio. No contaba con antecedentes familiares de infertilidad y/o de trastornos en la olfacción. El estudio de imágenes por RMN no permitió visualizar adecuadamente los bulbo olfatorios, con surcos olfatorios de escasa profundidad. Su edad ósea era de 13-14 años. Frente a estos hallazgos, se indicó la realización del estudio genético del paciente el cual arrojó dos variantes en genes distintos, ambas en estado de heterocigosis: NM_001126128.2:c.163del en el gen PROK2 y NM_144773.4:c.254G>T (p.Arg85Leu) en el gen PROKR2. Recordemos, como punto central de este caso, que el primer gen codifica para el ligando que es reconocido por el receptor codificado por el segundo gen.

La primera variante, en el gen PROK2, corresponde a una deleción de una única base que genera un corrimiento del marco de lectura normal del transcripto, pudiendo realizarse la predicción del decaimiento del mismo por medio del mecanismo de ARNm *Nonsense-Mediated Decay*, y esto en un gen donde la pérdida de función es un mecanismo conocido de enfermedad (PVS1). La misma cuenta con una frecuencia alélica de filtrado (FAF) de 2,7x10^-4 según informa la base de datos de frecuencia alélica poblacional gnomAD, con ausencia de homocigotas. Este valor se encuentra por debajo de la máxima frecuencia alélica poblacional creíble para pensar a la variante como causal de la patología (PM2_Supporting). La mayoría de los alelos detectados se registra en la población europea no-finlandesa (frecuencia 2,9x10^-4). Se trata de una variante que ha sido ampliamente estudiada, siendo probada su incapacidad de lograr activar a su receptor PROKR2, aún en el caso de lograr escapar el mecanismo de NMD. Además, la variante ha sido observada en una cantidad de pacientes previos con diagnósticos de CHH o KS. Registramos la presencia de al menos cinco pacientes homocigotas en nuestra base de datos de la revisión sistemática de variantes, como así también de cuatro heterocigotas (PM3), uno de los cuales contaba además con una variante missense acompañante en el gen WDR11. La variante fue corroborada por Sanger en estado de heterocigosis en el paciente, y se observó que fue heredada de su madre que era heterocigota para la misma. El conjunto de toda esta evidencia, y de acuerdo a los criterios de ACMG nos lleva a una probabilidad a posteriori de patogenicidad de 0,997, la cual nos permite asignar una clasificación de Patogénica.

Por otro lado, nos encontramos con la segunda variante en PROKR2, también en estado de heterocigosis. La misma cuenta con una frecuencia alélica de filtrado (FAF) de 8,3x10^-4 según informa la base de datos de frecuencia poblacional gnomAD (de nuevo, por debajo del valor de corte previamente calculado), con presencia de un homocigota en población americana mestiza, donde se registra la mayor frecuencia subpoblacional (1,0x10^-3) (PM2_Supporting). Este cambio determina un cambio de Arginina por Leucina

en la posición 85, y si bien presenta una frecuencia compatible con lo que nosotros modelamos como “poco frecuente”, no podemos dejar de ver que los cambios por Histidina, Cisteína o Glicina en la misma posición también cuentan con una interesante cantidad de alelos secuenciados, incluyendo algunos presentes en homocigosis. Todas estas variantes se encuentran reportadas en ClinVar como de interpretación de patogenicidad conflictiva. De hecho, nuestra variante cuenta con dos submissions de significado incierto y una de probablemente benigna. A continuación se intentó investigar la conservación del residuo observando el AMS que dio origen a la definición del dominio 7 transmembrane receptor (rhodopsin family) de Pfam (PfamID: PF00001), construido en base a una diversidad de receptores de los 7 dominios transmembrana. En este alineamiento la arginina no presenta una conservación demasiado alta, si bien podemos apreciar en el HMM Logo una cierta tendencia a la presencia de residuos cargados positivamente (arginina, lisina e histidina). El análisis de conservación por Divergencia de Jensen-Shannon, muestra que el residuo se encuentra en el segundo decilo de los residuos más conservados, utilizados secuencias del cluster UniRef50. Resulta interesante notar que este residuo se encuentra localizado dentro del primer loop intracelular de este receptor acoplado a proteína Gq, mientras que la bibliografía nos dice que tradicionalmente han sido el segundo y el tercer loop los asociados a la activación de dicha proteína. El último reporte de ensayos funcionales conducidos sobre la variante arroja un resultado de baja patogenicidad, definida la misma como una habilidad de señalización Gq entre el 50 y el 80% de la proteína salvaje. Existe una cantidad de pacientes reportados previamente con la variante, entre los cuales podemos encontrar tres con un fenotipo compatible, siendo uno un homocigota acompañado de dos variantes heterocigotas en PLXNA1, y otros dos heterocigotas, uno de los cuales tiene a su vez una variante heterocigota en FGFR1. La predicción bioinformática por el metapredictor REVEL otorga un score de 0,75 (PP3). Observando el algoritmo de comparación entre variantes observadas y esperadas en gnomAD para la presencia de variantes missense, podemos ver que existe una cierta restricción en la zona correspondiente a este loop intracelular, con un Z-score de 3,83 para ese bloque de 1kb. Finalmente, la variante fue corroborada por Sanger en estado de heterocigosis en el paciente, y se observó que fue heredada de su padre que era heterocigota para la misma. El conjunto de toda esta evidencia según la ACMG nos lleva a una probabilidad posterior de patogenicidad de 0,5, la cual nos permite asignar una clasificación de Variante de Significado Incierto.

El hecho de contar con una variante Patogénica heterocigota en PROK2, independientemente de la variante de significado incierto heterocigota en PROKR2, es para muchos laboratorios fundamento suficiente para cerrar el diagnóstico molecular del paciente por tres hechos: la clasificación de la variante, el modelo de herencia asociado al cuadro por OMIM (AD) y la presencia en bibliografía de casos previos con la variante. No obstante, al contar nuestro grupo con expertiz en el estudio de estos genes, no nos confiamos en absoluto y el análisis se decide continuar hasta las últimas consecuencias. Esto está fundamentado principalmente en el hecho de que somos conocedores de la penetrancia incompleta que variantes en estos genes históricamente han presentado. Por otro lado, sabemos que la presentación fenotípica suele ser mucho más definida en casos de homocigosis o heterocigosis compuesta en estos genes, mientras que los simples heterocigotas suelen tener presentaciones más difusas, con fenotipo extra-gonadal extra-olfatorio. Además del hecho que en muchos de estos últimos casos se han encontrado variantes acompañantes en otros genes. Por último, puede resultar obvio en este caso que se trata de un par de genes que codifican para la pareja ligando-receptor, pero la epistasis

entre ambos genes se encuentra ampliamente curada y corroborada por la comunidad científica.

El par constituído por estas dos variantes es un claro candidato a evidenciar un efecto oligogénico, y para ello fue estudiado por medio del software ORVAL en su interfaz de usuario. El análisis arrojó como resultado la existencia de un par de interacción entre ambos, donde la combinación entre ambas variantes fue puntuada con un score de predicción de patogenicidad, VarCoPP, de 0,998. Este score ubica a la combinación en la zona de 99,9% de confianza de causal de enfermedad, como podemos apreciar en la Figura 14a. En la misma podemos indagar cómo son las contribuciones que cada variable hace al veredicto final, como así también los aportes relativos entre las mismas, y de esta manera entender cuáles son las principales variables contribuyentes. Podemos observar que existen cinco variables destacadas, siendo las dos primeras (CADD1 y CADD3) las predicciones que el predictor de patogenicidad de variantes CADD realiza sobre ambas variantes, correspondiendo la primera a la variante en PROKR2 y la segunda a la variante en PROK2. Podemos decir que desde el aspecto meramente molecular, el impacto de ambas variantes parecería estar afectando ambos productos génicos. Por otro lado, vemos que las tres variables de relevancia restantes (BiolDist, BP_similarity y KG_dist) nos hablan sobre las relaciones existentes entre genes. La distancia biológica (BiolDist) nos habla sobre una clara proximidad entre los productos génicos de ambos genes en redes de interacción proteína-proteína. La similitud de proceso biológico (BP_similarity) nos informa que ambos genes se encuentran cercanamente relacionados desde el punto de vista de la ontología GO que define los procesos biológicos en que estos genes se encuentran implicados. La última de estas tres variables (KG_dist) nos informa que en un grafo de conocimiento oligogénico construido in-house por los mismos desarrolladores, ambos genes se encuentran conectados por un bajo número de nodos, es decir, cercanos. Estas tres últimas variables nos dicen que existe una serie de puntos sobre los cuales podemos fundamentar un probable efecto epistático entre genes y, yendo más allá, las variantes que cada uno de ellos porta. Es interesante notar que muchas de las otras variables que forman parte del árbol de decisiones de este algoritmo de tipo Random Forest, están relacionadas a la conservación de estos genes, su sensibilidad a la haploinsuficiencia o su probabilidad de asociarse a uno u otro modelo de herencia mendeliano, todas estas cuestiones que como ya mencionamos no se encuentran demasiado claras aún a ojos de la comunidad científico-médica.

Figura 14: a) Score de predicción de VarCoPP para la combinación en estudio. En rojo podemos observar las contribuciones de las diferentes variables del algoritmo de Random Forest en favor de patogenicidad, mientras que en azul se pueden observar aquellas en contra. b) Veredicto del predictor de efecto digénico, este algoritmo de tipo Random Forest ubica a la combinación como Monogénica acompañada de variante modificadora.

Por otro lado, en la Figura 14b podemos observar el veredicto del predictor de efecto digénico, el cual clasifica a esta combinación como “Monogénica acompañada de una variante modificadora” con una probabilidad de 1. Esto significa que este algoritmo de Random Forest siempre ubicó a la combinación en una hoja correspondiente a esta categoría y nunca en una hoja de “Verdadera digénica” o “Diagnóstico molecular dual”.

Conclusiones

En esta sección de resultados, expusimos el desempeño general que el equipo de trabajo tuvo al aplicar todos los procedimientos pertinentes de análisis genómico sobre los pacientes seleccionados que integran la cohorte de estudio. Nuestro rendimiento diagnóstico estuvo en consonancia con el que podemos encontrar en literatura científica, y de hecho lo superó en muchos casos. Esto habla principalmente del altísimo grado de capacitación de los médicos integrantes del equipo, tanto en lo que respecta al trabajo en consultorio con el paciente reconociendo los indicios de un cuadro compatible como también de su articulación con la genética, biología molecular y bioquímica subyacentes. Esto permitió una selección altamente certera de los pacientes a estudiar, como así también, de los genes compatibles con ser protagonistas de su fisiopatogenia. De todas maneras, existe un dato que no podemos ignorar: aún así, casi la mitad de los casos seleccionados no llega a dicho diagnóstico molecular, y en muchos casos esto se da por eventos que no estamos viendo, o que quizás si vemos pero aún no poseemos las herramientas necesarias para identificarlos.

A continuación, se procedió a desarrollar cinco casos que resultan representativos de todo el trabajo que se realiza de rutina en el servicio cuando un paciente es estudiado por NGS. Como podemos ver, las variantes de tipo SNP o InDel localizadas sobre secuencia codificante (o inmediaciones intrónicas) representan una buena parte de la casuística, y también de los esfuerzos abocados al análisis. La interpretación de estas variantes es laboriosa, implicando no solo el tiempo de buscar evidencia respectiva a la variante en estudio, sino que además existe todo un esfuerzo de actualización constante pues los criterios de ACMG, vía las recomendaciones de ClinGen, permanentemente son refinados y optimizados. Por otro lado, si bien los experimentos de NGS parecen a veces una herramienta ideal, no debemos olvidar que diferentes cambios genéticos ameritan considerar diferentes técnicas que poseen una mayor performance para evidenciarlos. Tal es el caso de los CNVs, donde si bien podemos aprovechar los datos generados por NGS para intentar relevarlos, su identificación precisa requiere el uso de técnicas confirmatorias.

Por último, no olvidemos que tradicionalmente el enfoque mendeliano de estudio de la genómica clínica es el que ha predominado, pero no debemos olvidar que la etiopatogenia de un cuadro puede presentarse con modelos más complejos. Tal es el caso de la oligogenicidad, donde entendemos que las pequeñas contribuciones marginales de un set de dos o más variantes pueden ser responsables de disparar un cuadro. Para realizar el análisis de estas combinaciones debemos cambiar de paradigma, puesto que comienza a ponerse en juego simultáneamente una cantidad elevada de variables que debemos contemplar. También podemos mencionar a la epigenética, que también cuenta con el potencial de explicar causalmente un porcentaje de los casos. Seguir capacitando a los integrantes del equipo de trabajo y expandiendo las facilidades tecnológicas del laboratorio son los pilares sobre los cuales construir un aumento del porcentaje de pacientes que puedan contar con un diagnóstico molecular para el cuadro que presentan.

Sección 3: MotSASi.

Exploración de la base de datos: Eukaryotic Linear Motif resource (ELM database)

El tercer capítulo de resultados de esta tesis corresponde al trabajo realizado en relación con la predicción de patogenicidad en un tipo de región proteica de alto interés funcional en lo referido a la fisiología celular normal , y particularmente a la fisiopatología de enfermedades poco frecuentes: los motivos lineales cortos (SLiMs, del inglés Short Linear Motifs). Los cimientos del trabajo realizado podemos adjudicarlos a la existencia de una base de datos de fundamental importancia, el ELM (Eukaryotic Linear Motif), ya que es allí donde científicos de múltiples disciplinas, estudiosos de múltiples especies, se dirigen a depositar la información referente al descubrimiento y confirmación funcional de "nuevos" motivos en una proteína de dicha especie. A su vez, el ELM funciona como un algoritmo de búsqueda, pues permite al usuario ofrecer como *query* una determinada secuencia proteica, y en función de la información disponible, ofrecerá un abanico de resultados relativos a los motivos descriptos previamente, y a potenciales motivos encontrados en la secuencia.

El ELM cuenta hoy en día con 352 "clases" (o tipos) de motivos que incluyen a 4.272 "instancias" (motivos encontrados en una dada secuencia). Una clase corresponde entonces a un tipo de motivo definido por una expresión regular, que delimita los residuos de aminoácidos que se contempla pueden localizarse en cada posición del mismo. Una instancia por su parte corresponde a una determinada aparición de dicho motivo (compatible con la expresión regular que lo define) en una determinada proteína, y que fuera objeto de análisis y experimentación por parte de un laboratorio de investigación. De las instancias reportadas un 95% son verdaderos positivos, es decir, existe el conocimiento de que fueron corroboradas como motivos funcionales que interaccionan con un dominio, su compañero de unión. Existen actualmente 2.798 interacciones motivo-dominio individualizadas en el ELM. Por otro lado, son escasos los reportes sobre secuencias compatibles con la expresión regular, que son experimentalmente rechazadas como motivos funcionales (lo que sería un falso positivo confirmado).

De fundamental importancia para este trabajo resulta la disponibilidad para una determinada clase de motivo de estructuras cristalográficas que reflejan cómo es la interacción del motivo en cuestión con su domino compañero de interacción (o receptor). En general, estas estructuras cristalográficas nos muestran al motivo como un péptido que encuentra su sitio diana en el receptor que se encuentra completamente representado. Solo un 61% de las clases de ELM (214/352) cuentan con al menos una instancia que tiene asociada una estructura cristalográfica depositada en el Protein Data Bank (PDB). Esto nos habla sobre el limitado cubrimiento estructural que podemos tener de toda una cantidad de interacciones motivo-dominio en el ELM.

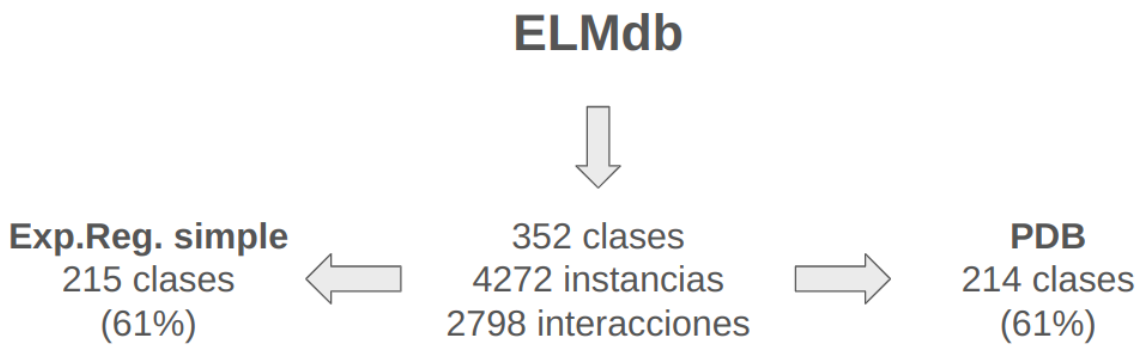

Figura 42. Estructura simple de la base de datos ELM.

Recapitulando el objetivo de MotSASi

Teniendo la presente tesis el foco puesto en lograr mejoras en los réditos diagnósticos genómicos y algoritmos de predicción de patogenicidad, nuestro objetivo principal al diseñar este algoritmo era el de lograr una transferencia real del conocimiento estructural proteico que cultivamos en nuestro laboratorio a la práctica diagnóstica. En el trabajo de 2015 de ACMG-AMP, se menciona un criterio de evidencia, PM1, que enuncia: "Variante localizada en un hot-spot mutacional y/o un dominio funcional bien establecido sin presencia de variación benigna". Si entendemos a los dominios funcionales como aquellas unidades estructurales, funcionales y evolutivas independientes definidas por Pfam, Prosite y tantas otras herramientas, podríamos asignar con justicia los denominados hot-spots a estas unidades funcionales o dominios. Sin embargo, estos dominios son demasiado abarcativos en términos de longitud de secuencia y no se encuentran absolutamente desprovistos de variación benigna, lo que es otro de los requisitos de PM1. Por otro lado, aún dentro de los SLiMs, considerados per-se como motivos, existen variaciones que resultan perfectamente tolerables, y que fueron erróneamente consideradas deletéreas en el pasado dada únicamente su localización al interior de los mismos. El gran desafío planteado entonces en esta tesis, fue el de lograr generar, para cada motivo de tipo SLiM, una matriz de sustitución donde pudiera fácilmente evaluarse el efecto patogénico o benigno de sustituir cada aminoácido por cada uno de los 20 aminoácidos estándar. Esta idea da nombre a nuestro algoritmo, MotSASi (Motif-occurring Single Amino acid Substitution information).

En el armado de estas matrices debería incluirse información proveniente de 3 fuentes:

a) Variantes de interés clínico provenientes de ClinVar, tanto patogénicas como benignas.
b) Variantes de alta frecuencia alélica poblacional de gnomAD, que se consideran benignas.
c) Predicciones termodinámicas de cambio de un aminoácido por otro, evaluado por el software FoldX.

Debemos aclarar en este punto, que si bien el objetivo principal era el de generar estas matrices para su uso en clínica, simultáneamente, se nos presenta un segundo objetivo, que consiste en colaborar en la resolución de otro problema, constituido por el

desafío de lograr una confiable identificación de nuevas instancias de motivos presentes en el proteoma humano.

Cuando se realiza un scanning con la expresión regular de un motivo sobre el conjunto de secuencias proteicas que constituyen el proteoma humano, resulta enorme la cantidad de coincidencias obtenidas. Sin embargo, es posible que estos motivos potenciales sean en su mayoría falsos positivos, resultado producto del enorme universo de búsqueda. Por eso mismo, tradicionalmente se han utilizado filtros para descartar parte de los mismos (como la estructura secundaria, exposición al solvente, conservación, etc). En este trabajo de tesis, nosotros nos valemos de filtros similares en nuestro algoritmo pero agregamos una herramienta clave, la utilización de los tres *inputs* mencionados previamente. De esta manera, nuestro objetivo general radica en utilizar las variantes de interés clínico (además de los otros filtros) en la búsqueda de nuevas instancias (motivos) con alta confianza. Complementando las variantes de ClinVar y gnomAD, recurrimos además a una herramienta predictiva como FoldX, a fines de utilizar las estructuras cristalográficas, y con el objetivo de comprender mejor qué cambios, y en qué posiciones, son mejor o peor tolerados. Una primera idea de lo aquí planteado se plasma en la Figura 43, que se irá terminando de aclarar a lo largo de estos resultados.

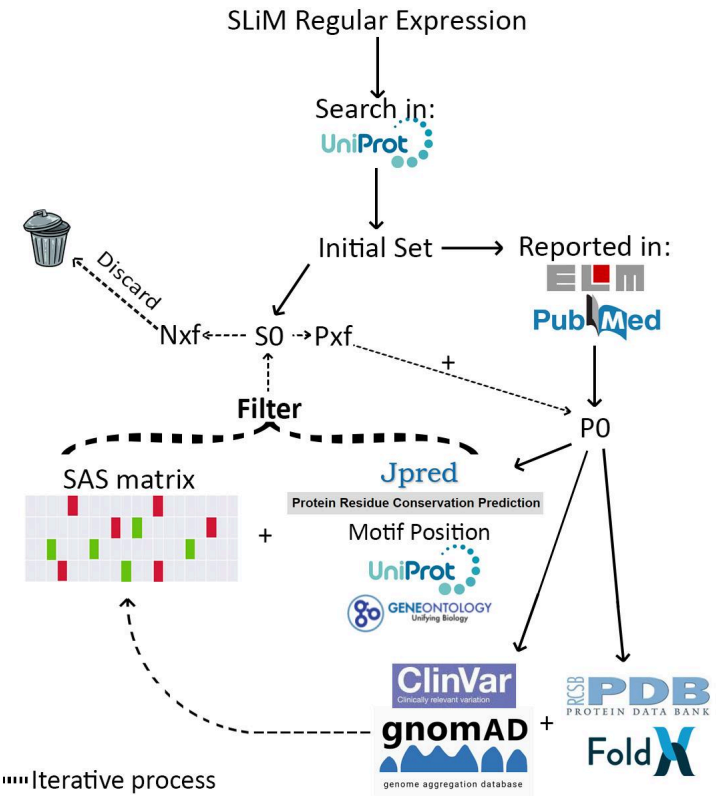

Figura 43. Esquema general de trabajo de MotSASi.

MotSASi novelado: una metodología minuciosa narrada como un workflow

La búsqueda de coincidencias de expresiones regulares de SLiMs en el proteoma

El proceso inicia con la selección de un grupo de clases de motivos del ELM de interés humano en clínica con el cual llevar adelante el análisis. Por una cuestión de

simplicidad, seleccionamos sólo aquellos motivos que presentaban una longitud fija, es decir, el número de residuos en la secuencia del motivo no varía, independientemente de cuáles sean estos. Además, debimos filtrar a los mismos en función de que presentaran instancias experimentalmente comprobadas en el proteoma humano (verdaderos positivos), un mínimo de variantes (5) patogénicas/benignas en los mismos, y al menos una instancia asociada a una estructura cristalográfica del PDB.

A continuación se presenta una lista de los motivos analizados en este primer trabajo acompañados de su expresión regular:

DOC_MAPK_JIP1_4_motif ([RK]P[^P][^P]L.LIVMF),
LIG_PCNA_PIPBox_1_motif ([QM].[^FHWY][LIVM][^P][^PFWYMLIV][FYHL][FYWL]..),
DOC_PP2A_B56_1_motif ([LMFYWIC]..[IVLWC].E),
LIG_NRBOX_motif ([^P]L[^P][^P]LL[^P]),
DOC_MAPK_NFAT4_5_motif, ([RK][^P][^P][LIM].L.[LIVMF].),
DEG_APCC_KENBOX_2_motif (.KEN.),
DOC_ANK_TNKS_1_motif (.R..[PGAV][DEIP]G.),
LIG_PAM2_1_motif_1 (..[LFP][NS][PIVTAFL].A..[FY].[PYLF].),
LIG_PTAP_UEV_1_motif (.P[TS]AP.),
LIG_SH3_2_motif (P..P.KR),
MOD_CDK_SpxK_1_motif (...[ST]P.[KR]),
LIG_SH2_STAP1_motif (Y[DESTA][^GP][^GP][ILVFMWYA]),
DOC_MAPK_MEF2A_6_motif ([RK]...[LIVMP].[LIV].[LIVMF]),
DEG_SCF_TRCP1_1_motif (DSG..[ST]),
DOC_CYCLIN_RxL_1_motif_2 (.[^EDWNSG][^D][RK][^D]L.[FL][EDST]),
LIG_CaM_IQ_9_motif_1 ([ACLIVTM][^P]
[^P][ILVMFCT]Q[^P][^P][RK][^P][^P][^P][^P][RKQ][^P][^P]),
LIG_14-3-3_CanoR_1_motif_4 (R[^DE][^DEPG][ST][^PRIKGN]..[VILMFWYP]),
LIG_PDZ_Class_1_motif ([ST].[VIL]),
LIG_KLC1_WD_1_motif ([LM].W[DE]),
LIG_PTB_Apo_2_motif (NP.[YF])

La construcción de matrices: ClinVar, gnomAD y FoldX

De esta manera comienza el proceso del algoritmo, donde el primer paso es realizar un escaneo del proteoma humano contra la expresión regular que define el motivo en cuestión. Este primer set generado fue denominado Set Inicial, que a su vez fue filtrado para quedarnos únicamente con aquellos potenciales motivos que cuentan con al menos una variante proveniente de ClinVar y/o gnomAD, formando de este modo el Set 0 (S0). A su vez, también se identificaron aquellos motivos que se encontraban anotados como experimentalmente confirmados en la base de datos del ELM, constituyendo el Set Positivo 0 (P0), que no es otra cosa que el conjunto de los verdaderos positivos con los que contamos, y a los cuáles estudiamos para obtener información sobre su estructura secundaria, exposición al solvente, conservación, localización en la secuencia y términos GO asociados. Con estos mismos motivos procedimos a confeccionar las matrices con las variantes de ClinVar y gnomAD, como mostramos en la Figura 44a y 44b.

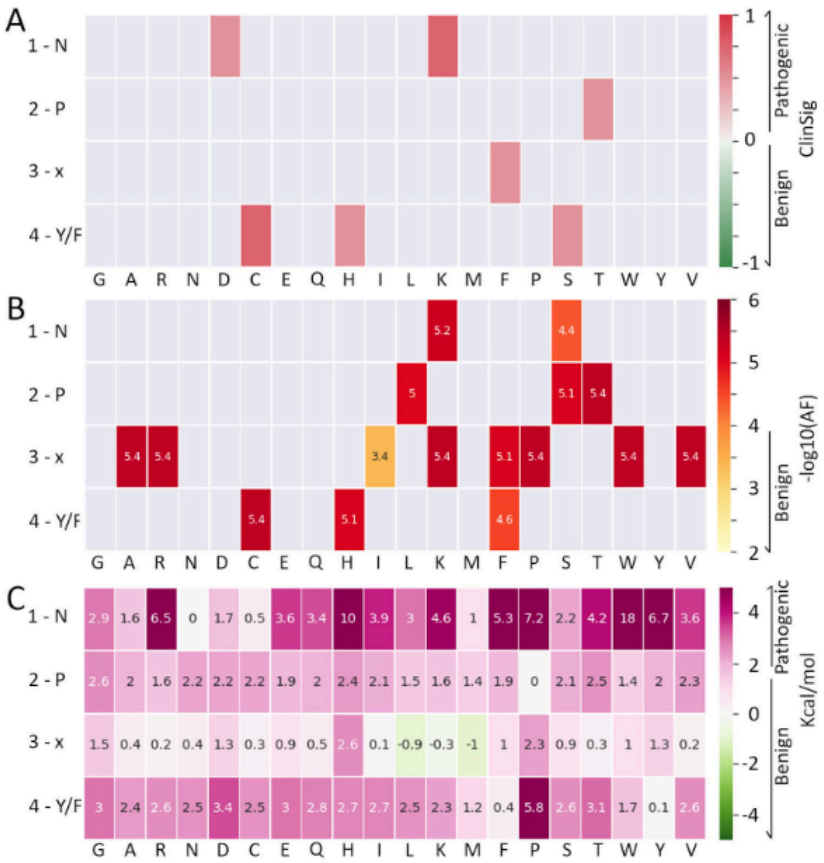

Figura 44. a) Matriz con variantes de ClinVar para el motivo LIG_PTB_Apo_2, tomado como ejemplo. b) Matriz con variantes de gnomAD para el mismo motivo. c) Matriz de estabilidad termodinámica generada a partir del software FoldX.

Además de las matrices de variantes, hicimos uso de las estructuras cristalográficas asociadas a las instancias consideradas como verdaderos positivos, fuimos a buscar las mismas al PDB. A partir de las estructuras de los complejos SLiM-receptor confeccionamos matrices que reflejan el impacto termodinámico en la estabilidad de la unión motivo-dominio para cada cambio posible de aminoácido. Para ello se utilizó el software FoldX, con el comando PositionScan. Obsérvese que en este caso la matriz, como se aprecia en la Figura 44c, se encuentra completa, y no presenta celdas vacías, hecho absolutamente lógico puesto que se trata de una predicción y no de la observación de aparición de variantes en una población, como es el caso de ClinVar y gnomAD, donde las matrices presentan grandes vacíos, ya que hay muchas variantes que aún no han sido observadas.

Los ciclos de iteración

Una vez construidas estas matrices, se inicia una serie de tres ciclos, cada uno de los cuales podrá contar con una o más iteraciones, como explicaremos a continuación. Los ciclos podemos denominarlos:

1) Ciclo rígido: se evalúan solamente las posiciones del motivo que presentan mayor restricción en los aminoácidos permitidos (típicamente 5 o menos). Solo se tomarán en cuenta para el filtro las matrices de ClinVar y gnomAD.

2) Ciclo FoldX: se evalúan solamente las posiciones del motivo que presentan mayor restricción en los aminoácidos permitidos, pero ahora se agrega la matriz de FoldX.
3) Ciclo flexible: se incorporarán al análisis las posiciones que admiten una amplia gama de aminoácidos en dichas posiciones (por ejemplo aquellas que tienen X o ^P en el motivo regular), haciendo de nuevo uso de los tres tipos de matrices.

Comenzando con el Ciclo Rígido, lo que haremos es ingresar al algoritmo aquellos potenciales motivos pertenecientes al Set 0, es decir, coincidencias de la expresión regular en el proteoma humano que presentan al menos una variante en ClinVar o gnomAD. Ahora para cada motivo candidato, lo que haremos es evaluar si estas variantes coinciden o no coinciden con aquellas variantes que incluimos en las matrices de ClinVar y gnomAD que construimos a partir de los controles positivos de P0. Recordar que por encontrarnos aún en el Ciclo Rígido, solo evaluaremos aquellas posiciones en los motivos que toleran menor número de residuos. En el caso de ClinVar es sencillo pues nos remitimos a la clasificación patogénica/benigna de la variante. Sin embargo, esto resulta un poco más complejo a la hora de evaluar variantes de gnomAD. El análisis de esta base de datos nos muestra que aquellas variantes que cuentan con frecuencias alélicas (FA) bajas, pueden ser tanto benignas como patogénicas, mientras que aquellas que superan determinados valores de FA son casi exclusivamente benignas (y cuando no lo sean, probablemente exista un reporte en ClinVar dada la alta circulación y consecuente presencia de casos previamente reportados). La gran pregunta está en donde poner este valor de corte. En la primera implementación de MotSASi, esto se resolvió realizando un análisis de FA para variantes patogénicas y benignas, y se terminó seteando un valor de corte de 0,0001, que si lo expresamos como -log(0,0001) da un valor de 4, por encima del cual existe un predominio claro de variantes benignas. Por lo tanto, al comparar las variantes de gnomAD localizadas en motivos candidatos, si el -log(FA) era ≥ 4 se la consideraba benigna y podía ser comparada a cambios registrados en las matrices de ClinVar y gnomAD.

Si al comparar las variantes en los potenciales motivos con las matrices registramos que existía una coincidencia en su veredicto, benigno o patogénico, entonces a estos motivos se les concede haber pasado favorablemente las matrices de variantes. Por el contrario, si existían discordancias (como por ejemplo, la variante que carga el motivo es benigna por gnomAD pero cuando se confronta con la matriz de ClinVar se observa la presencia de una variante patogénica en el cambio en cuestión), entonces estos motivos se ven rechazados por las matrices de variantes, y se los envía al Set Negativo 1 (N1). Típicamente los motivos candidatos no presentan más de dos variantes patogénicas/benignas (en muchos casos una sola), pero en el caso donde existiera un número superior, se evalúa la predominancia de concordancias o discordancias. También ocurre que, dada la cantidad de celdas vacías en las matrices, muchos motivos con variantes aún no puedan ser evaluados, y son enviados al Set Sobrante 1 (S1).

A aquellos motivos que pasan la instancia de las matrices de variantes, pasan a un segundo control, donde se les comprobará una serie de características, de acuerdo a lo que se detalla a continuación:

1) Conservación: se les solicita que la conservación de los residuos que lo conforman sea al menos tanto o más alta como el menos conservado de los residuos que integran los controles positivos en P0.

2) Estructura secundaria: se realiza una predicción con el software Jpred y se evalúa si existe un buen grado de concordancia entre el motivo y el conjunto de los controles positivos en P0.
3) Términos GO: se le pide al motivo candidato que posea al menos un término GO de aquellos que resultaban más representativos de los controles positivos en P0.
4) Localización en la secuencia proteica: se corrobora si el motivo se localiza en una región análoga a aquellas donde se localizan los motivos en P0.

De esta manera, si un motivo candidato pasa todos estos filtros, podemos decir que el mismo es un motivo funcional con un alto grado de confianza, y lo enviamos al Set Positivo 1 (P1), mientras que si no pasa esta segunda etapa de filtros lo redirigimos a N1.

Lejos de dejarlo aparte, los motivos en P1 serán incorporados al análisis, y utilizaremos las variantes que se localizaban sobre los mismos para enriquecer nuestras matrices de ClinVar y gnomAD. A su vez, tras haber completado una primera iteración en este ciclo, podemos comenzar una nueva iteración, donde todos aquellos motivos cuyas variantes no habían encontrado comparaciones en la primera, pero sí ahora en la segunda iteración, puedan ser evaluados. De esta manera, el proceso se repite hasta que no existen nuevos motivos ingresantes en una nueva iteración. En ese momento se considera que el Ciclo Rígido ha concluido y puede darse inicio al Ciclo FoldX.

En el Ciclo FoldX, la dinámica general es idéntica, pero cambian algunos detalles y aparece un nuevo actor principal clave en la etapa de confrontación de variantes. En primer lugar, los potenciales motivos que ingresarán al ciclo serán aquellos pertenecientes a S1, es decir, todos aquellos sobre los que no pudimos tomar decisiones en el ciclo anterior. Las variantes localizados en estos serán confrontadas como siempre contra las matrices de ClinVar y gnomAD pero luego, serán también confrontadas contra la matriz de FoldX, donde se encontraban todos los impactos en la estabilidad de unión predichos a partir de las estructuras cristalográficas que se habían recabado en PDB vía ELM. Recordemos que en este ciclo seguimos contrastando variantes localizadas en las posiciones rígidas del motivo.

De nuevo, la pregunta interesante es cómo comparar una variante patogénica/benigna obtenida de ClinVar o gnomAD contra un valor termodinámico de (in)estabilidad predicho por FoldX como un ddG (en kCal/mol). Para esto, en la primera implementación de MotSASi se optó por utilizar un valor de corte de 2 kCal/mol elaborado por Radusky et. al. en un trabajo previo [128], producto de un estudio realizado en variantes patogénicas y benignas en proteínas asociadas a Rasopatías. Por encima de este valor se considera que la variante afecta sustancialmente la interacción motivo-dominio. En consonancia, se optó por utilizar simultáneamente otro valor de corte de 1 kCal/mol, por debajo del cual se presume que la variante no generará una inestabilidad significativa de la unión.

Por lo tanto, una variante patogénica perteneciente a un potencial motivo contabilizaría una coincidencia si encontrara en la matriz de FoldX un valor de $ddG \geq 2$ kCal/mol, resultando análogo el razonamiento a variantes benignas que hallen un valor de $ddG \leq 1$ kCal/mol. De la misma manera, podemos pensar las discordancias en función de que estos resultados estuvieran cruzados. Así es como, en definitiva, en el Ciclo FoldX terminamos evaluando a todos los potenciales motivos que contaban con variantes de ClinVar y gnomAD en posiciones rígidas del motivo. Luego del paso de confrontación de variantes, al igual que en el Ciclo Rígido, se procede a controlar las características de los motivos que la pasaron, evaluando como siempre conservación, estructura secundaria, términos GO y posición relativa. Finalmente, de nuevo, haremos uso de las variantes

localizadas en aquellos motivos que pasaron todos los filtros para enriquecer las matrices de ClinVar y gnomAD. Al final del Ciclo FoldX, tendremos a todos estos nuevos potenciales motivos evaluados y los asignaremos a sus respectivos sets, tal cual lo hicimos en el Ciclo Rígido, solo que ahora estos se llaman P2, N2 y S2.

Por último, los motivos remanentes en S2 ingresarán al tercer y último ciclo, el Ciclo Flexible, donde completamos nuestros análisis y evaluaciones de los potenciales motivos pues habilitaremos ahora las comparaciones de variantes que se localizan en las posiciones flexibles del motivo. El proceso es absolutamente análogo a los ciclos anteriores, pero los sets que obtendremos al final del mismo, serán P3, N3 y S3.

La salida de MotSASi reflejada y cuantificada en ejemplos

En la Tabla 2 se presentan los resultados obtenidos por MotSASi para tres motivos elegidos como ejemplos del proceso, donde podemos apreciar cómo los potenciales motivos se van distribuyendo en los diferentes sets a medida que transcurren los ciclos.

Nombre del motivo	LIG_PTB_Apo_2	LIG_PDZ_Class_1	LIG_KLC1_WD_1
Consensus Sequence	NP.[YF]	[ST].[VIL]$	[LM].W[DE]
Initial Set	1824	1022	1903
S0	178	136	122
P0	18	71	9
N1	25	5	1
P1	18	75	9
S1	135	56	112
N2	88	14	54
P2	19	86	9
S2	71	36	59
N3	149	40	103
P3	29	96	19
S3	0	0	0

Tabla 2. Progresión de la distribución de los potenciales motivos localizados en el proteoma humano a partir de la expresión regular en los respectivos sets ciclo a ciclo.

Producto de todo este proceso, nos encontraremos al final con dos resultados fundamentales. Por un lado, contamos con una cantidad de nuevas instancias de alta confianza de motivos en el proteoma humano (aquellos que fueron coincidentes con las matrices de variantes observadas y Foldx). Esto resulta clave ya que justamente estos sitios son funcionalmente importantes, y puede brindarnos información no sólo en tanto en lo respectivo al estudio de la normal fisiología o fisiopatología de enfermedades, sino que

también será de gran utilidad en el caso de registrar sobre los mismos variantes en casos a estudiar de genómica clínica. Pero por otro lado, y de la mano de esto último, ante la ocurrencia de una determinada variante en un motivo, también contaremos al final con una matriz de análisis con la cual buscar predecir el efecto final de una variante missense novel en el mismo. Un ejemplo de esta matriz podemos verlo en la Figura 45.

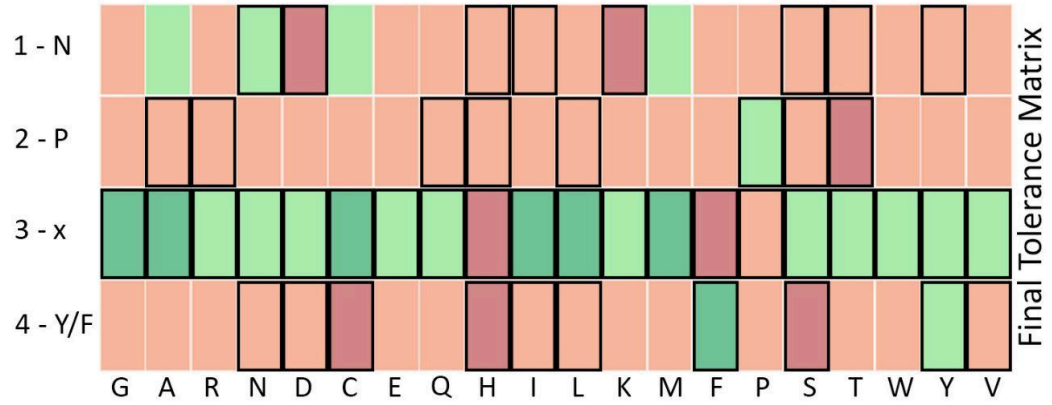

Figura 45. Matriz final de tolerancia del motivo LIG_PTB_Apo_2 a variantes missense en el mismo. El código de colores indica Rojo: disruptivo, Naranja: probablemente disruptivo, Verde claro: probablemente tolerado, Verde oscuro: tolerado.

Como podemos apreciar, concluimos en última instancia con nuestros objetivos iniciales. La utilización de estas matrices de tolerabilidad al cambio missense representan una herramienta útil a laboratorios clínicos avocados al diagnóstico molecular. Representan a su vez una oportunidad de profundización en la aplicación de criterios de evidencia ACMG, particularmente PM1, como así también lo relativo a la utilización de predictores (PP3/BP4). Sin embargo, somos conscientes de las limitaciones naturales del método, y fue eso lo que decidimos abordar en una segunda etapa de trabajo con MotSASi.

MotSASi 2.0: sorteando el obstáculo estructural utilizando AlphaFold2

Recapitulando, para poder estudiar un determinado motivo por nuestro algoritmo, debíamos contar con 1) controles positivos en humanos en la base de datos del ELM, 2) un número mínimo de variantes patogénicas/benignas en los mismos y 3) al menos una estructura cristalográfica de la interacción motivo-dominio depositada en el PDB. El primer requerimiento se cumple casi automáticamente en el momento que una clase de motivo aparece en la base de datos, pues en general no tarda en reportarse una instancia del mismo en humanos, aún cuando su descubrimiento se hizo en otra especie. El segundo requerimiento es absolutamente una limitante presente y real, no obstante lo cual, dada la amplia difusión de las tecnologías de secuenciación masiva, el número de variantes de interés clínico va exponencialmente en aumento, y por lo tanto cada vez se sortea más fácilmente este obstáculo. El tercer requerimiento constituía por lejos la peor de las limitaciones. Si bien cada vez se avanza más en técnicas orientadas a dilucidar estructuras proteicas, desde la difracción de rayos X hasta la crioelectromicroscopía, la disponibilidad de las mismas representando la interacción motivo-dominio es absolutamente limitada, más aún, en aquellos tiempos donde se diseñó la primera implementación de MotSASi. Por esto mismo, en esta segunda etapa decidimos valernos de los avances realizados en los últimos

años en materia de predicción de estructura tridimensional de proteínas en función de su información de secuencia. El algoritmo ganador de esta carrera fue el denominado AlphaFold2 (AF2), que a fuerza de resultados viene logrando cada vez más dejar en el olvido al clásico modelado por homología.

En sus últimas versiones, AF2 permite ingresar como *query* más de un polipéptido al mismo tiempo, y trabaja con los mismos para generar una predicción no solo del plegado individual de cada uno, sino que a su vez modela como será la unión entre ambos. De esta manera, nuestra hipótesis de trabajo se orientó a evaluar si AF2 podía ser la fuente que nos proveyera de estructuras representativas de la interacción motivo-dominio con el fin de analizar aquellas clases de motivos del ELM que contaban con esa limitante (es decir, para las cuales no se encontraba disponible una estructura cristalográfica).

El primer paso en esta búsqueda fue entonces el de lograr determinar si AF2, tomando como *query* las secuencias del péptido correspondiente al motivo y el polipéptido correspondiente al dominio, era capaz de replicar las estructuras cristalográficas experimentales con las cuáles ya contábamos. Los motivos utilizados en esta etapa fueron, por ende, aquellos que ya habían sido estudiados en la primera implementación de MotSASi. Se lograron resultados positivos en esta etapa, como podemos ejemplificar en la Figura 46a, pero también debemos decir que en muchos casos las predicciones no resultaron exitosas como en la Figura 46b. El principal problema en la predicción de AF2 no es la predicción de la estructura del receptor en general, sino algunas veces una incorrecta ubicación del motivo con respecto al mismo.

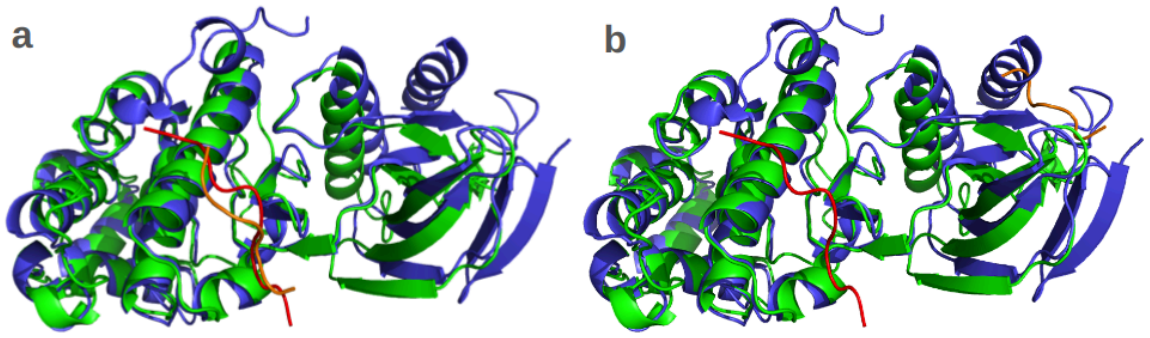

Figura 46. a) Predicción correcta de AF2 superpuesta con la estructura cristalográfica depositada en el PDB (PDB ID: 1UKH) que representan la interacción motivo-dominio para el motivo DOC_MAPK_JIP1_4. b) Predicción incorrecta de AF2 para el mismo motivo. En rojo y verde se aprecian los polipéptidos correspondientes a motivo y dominio en 1UKH, en naranja y azul las correspondientes a motivo y dominio productos de AF2.

Para los 17 motivos utilizados en la validación de AF2 como generador de predicciones de la interacción motivo-dominio, se generaron 10 modelos por motivo, variando entre ellos levemente algunos parámetros como la cantidad de ciclos de refinamiento de la predicción. En términos generales, es decir, tras realizar análisis exploratorios de mera observación, podemos decir que el desempeño fue correcto, pudiendo siempre generarse al menos una estructura de alta similitud a aquella que contábamos como referencia obtenida del PDB. Algo interesante para comentar es que de este análisis se pone de manifiesto rápidamente que AF2 no tiene en absoluto problemas para predecir el plegado del dominio de interacción (la estructura del receptor), de hecho la confianza atribuida en dicha predicción es altísima. Sobre el plegado del péptido no

podemos conjeturar demasiado puesto que su extensión es mínima, en general permanece desordenado, con mínimas expresiones de estructura secundaria. El verdadero desafío, y aquello donde en mucho de los modelos generados por AF2 fallaban, era en la ubicación misma del motivo con respecto al dominio. Recordemos, que existe una diferencia importante de tamaño entre motivos y dominios, contabilizando estos últimos muchísimos más aminoácidos. El verdadero desafío es entonces discriminar cuando AF2 ha logrado ubicar correctamente el motivo en torno al dominio.

Las estructuras cristalográficas de referencia y las predicciones de AF2 fueron reducidas a sus átomos comunes y se utilizó la métrica RMSD para compararlas en términos cuantitativos. Para el global de los modelos de AF2 generados, el RMSD obtenido fue de 2,71 +/- 1,32 en relación a sus respectivas referencias. Contando con estos datos por detrás, por medio de la inspección visual individual de cada modelo, se fue clasificando a las mismas en interacciones que eran un fiel reflejo de la interacción motivo-dominio y aquellas que no lo eran. Cuando evaluamos el conjunto de las predicciones óptimas, encontramos que estas presentaban un RMSD de 2,43 +/- 0,78 (registrando un 32% de las mismas un RMSD ≤ 2), mientras que aquellas erróneas se encontraban en 4,45 +/- 2,39. Como podemos ver, si bien presenta tendencias, el RMSD no es un gran clasificador de predicciones en óptimas y erróneas.

Contar con predicciones fidedignas de la interacción motivo-dominio es lo que en última instancia resuelve el problema de la disponibilidad de estructuras cristalográficas que, como ya habíamos explicado en la primera parte, es nuestro punto de partida para poder generar la matriz de FoldX de sustitución estructural. Esta era una herramienta clave a la hora de confrontar las variantes presentes en los potenciales motivos y moverlos secuencialmente a los sets P2, N2, P3 y N3. Como lo que comparamos son los valores de ddG en la matriz, resulta indispensable analizar si los valores obtenidos analizando una predicción óptima, coinciden con aquellos resultantes de procesar la estructura cristalográfica de referencia depositada en el PDB. En la Figura 47, podemos observar las matrices correspondientes para un motivo de ejemplo.

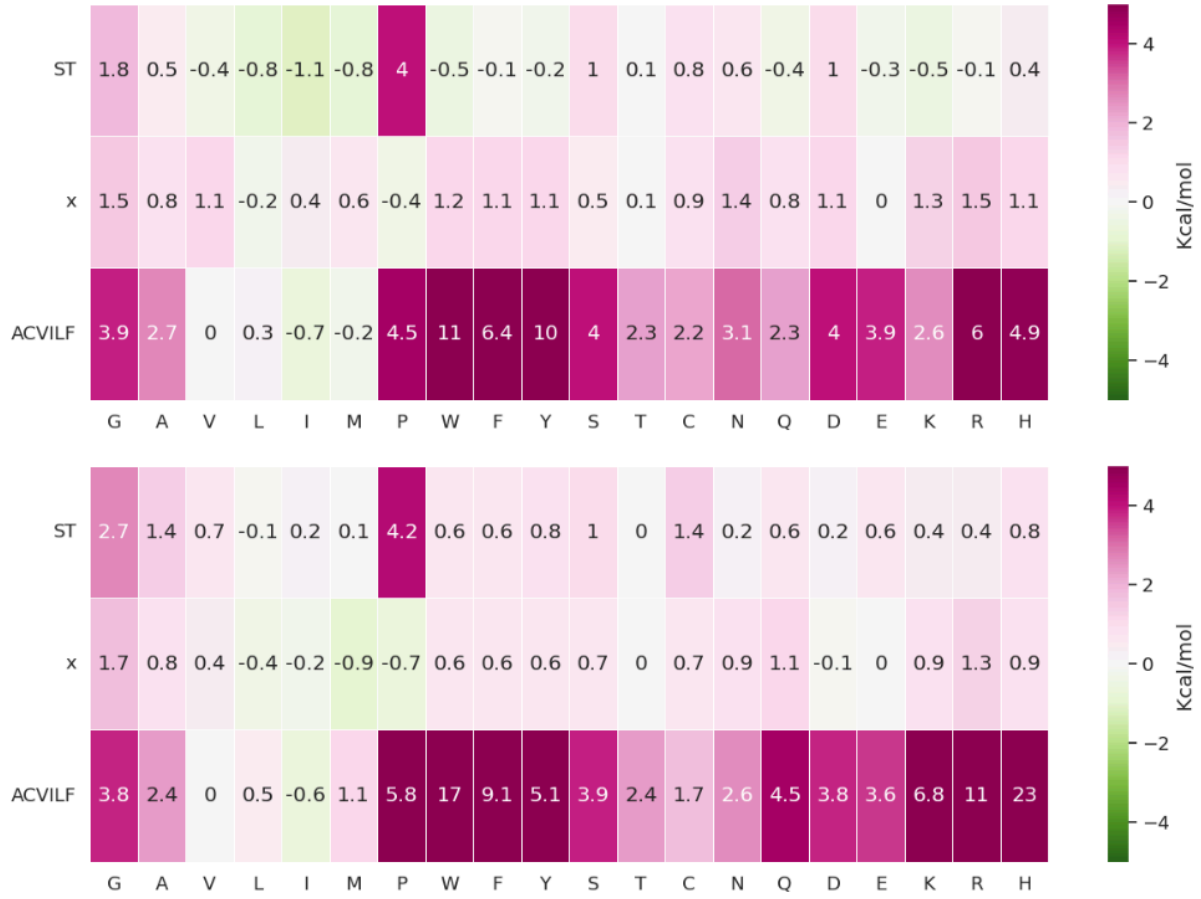

Figura 47. a) Matriz de estabilidad, medida en ddG (kCal/mol), para la estructura cristalográfica de referencia depositada en PDB (PDB ID:3GJ9), generada a partir del software FoldX, comando PositionScan, para el motivo LIG_PDZ_Class_1. b) Matriz de estabilidad análoga realizada sobre una de las predicciones óptimas emitidas por AF2.

A simple vista, podemos observar que existe un alto grado de similitud entre ambas matrices, no obstante lo cual decidimos abordar esta comparación cuantitativamente con dos métricas comunes. En primer lugar, calculamos el coeficiente de correlación de Pearson para los valores de ddG de estabilidad obtenidas con los modelos de AF2 y con las estructuras cristalográficas de referencia. Para aquellas predicciones evaluadas como óptimas, se obtuvo un coeficiente de correlación de Pearson de 0,65 +/- 0,22, mientras que la misma métrica para las predicciones descartadas era de sólo 0,27 +/- 0,26.

En segundo lugar, nos interesaba saber cómo era la coherencia existente entre matrices, es decir, entender si lo que una matriz planteaba como un cambio tolerado también lo era en la otra matriz. Para discriminar entre cambios tolerados o no-tolerados, debemos hacer uso de los valores de corte de ddG, similar a lo que planteamos anteriormente para la primera implementación de MotSASi. Para obtener el umbral de corte, entre las variantes toleradas y aquellas que no, tomamos 225 variantes de ClinVar (reportadas como patogénicas o benignas y al menos una estrella de Review Status) y de gnomAD (benignas de alta frecuencia) localizadas en instancias de motivos documentadas en el ELM que tuvieran un PDB asociado, y les calculamos su ddG de estabilidad utilizando FoldX. Nos decidimos por un valor de corte de 2,1 kCal/mol el cual logra un compromiso entre sensibilidad y especificidad (83% y 84% respectivamente), con una exactitud del 84%. En la Figura 48 podemos ver un histograma que expone esta discriminación.

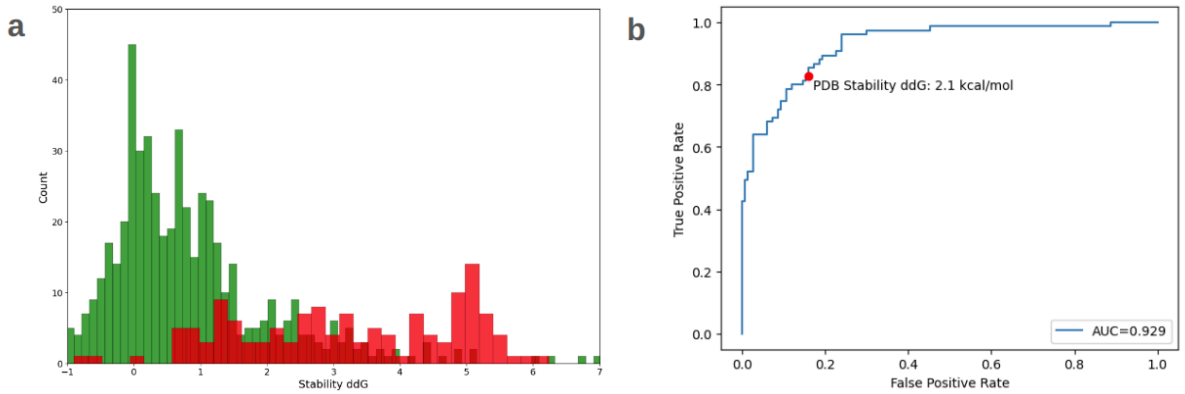

Figura 48. a) Histograma con las variantes de ClinVar y gnomAD utilizadas en la determinación del valor de corte de ddG de estabilidad. En rojo figuran aquellas variantes patogénicas, en verde figuran las benignas. b) Curva ROC (AUC=0,929) generada en el proceso de determinación del valor de corte de ddG para PDB, el punto rojo corresponde al valor de corte de 2,1 kCal/mol (S: 83%, E:84%).

Tomando este valor de corte, se procedió a analizar las matrices de ddG de aquellas estructuras cristalográficas tomadas de PDB (consideradas como la referencia) y las predicciones óptimas generadas por AF2 (consideradas como los valores a testear). Dada la diferente naturaleza de ambos tipos de estructuras, nos pusimos como primer objetivo determinar un valor de corte para las estructuras predichas por AF2, siguiendo la misma estrategia que en el punto anterior. En la Figura 49 pueden apreciarse el histograma y la curva ROC que nos permiten establecer, en este caso, un valor de corte de 1,4 kCal/mol para predicciones de AF2 en un compromiso de sensibilidad y especificidad (82% y 81%, respectivamente) con una exactitud del 81%.

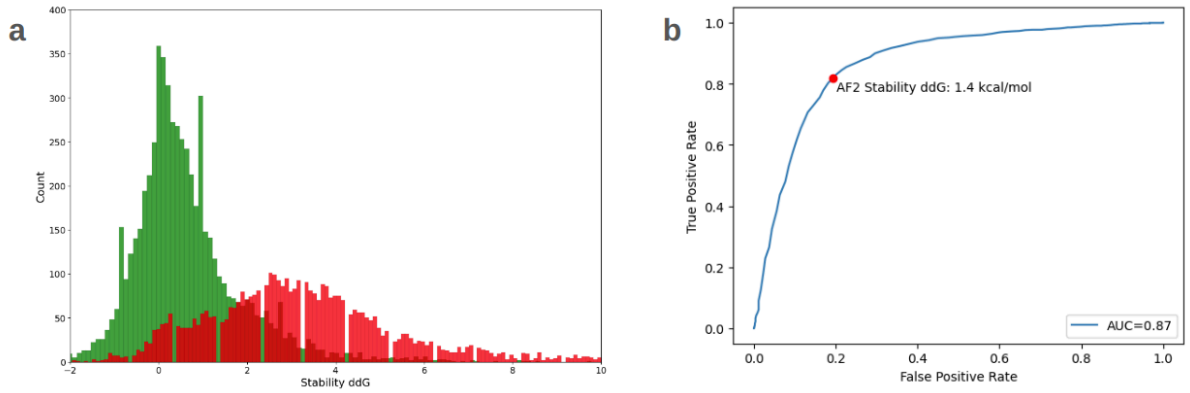

Figura 49. a) Histograma con los valores de ddG de estabilidad calculado en las estructuras cristalográficas predichas por AF2 para cambios previamente clasificados como tolerados o no-tolerados en estructuras depositadas en el PDB. En rojo figuran aquellos cambios no-tolerados, en verde figuran los tolerados. b) Curva ROC (AUC=0,870) generada en el proceso de determinación del valor de corte de ddG para AF2, el punto rojo corresponde al valor de corte de 1,4 kCal/mol (S: 82%, E:81%).

Finalmente, decidimos hacer una evaluación externa de la capacidad predictiva de las matrices de FoldX obtenidas con AF2, con las variantes patogénicas o benignas, provenientes de ClinVar o gnomAD, residentes en los motivos utilizados en la validación, y ver cómo estas podrían ayudar a elegir entre predicciones óptimas y erróneas de AF2. Esto

nos brinda una suerte de validación independiente de las métricas utilizadas en la clasificación de las predicciones en óptimas y erróneas, es decir, energía de unión motivo-dominio y confianza de AF2. Para efectuar la comparación entre variantes de ClinVar y gnomAD y las matrices termodinámicas de las predicciones, el valor de corte utilizado fue naturalmente el calculado para predicciones de AF2. A modo de ejemplo, incluimos en la Figura 50 el resultado de este análisis para un determinado motivo.

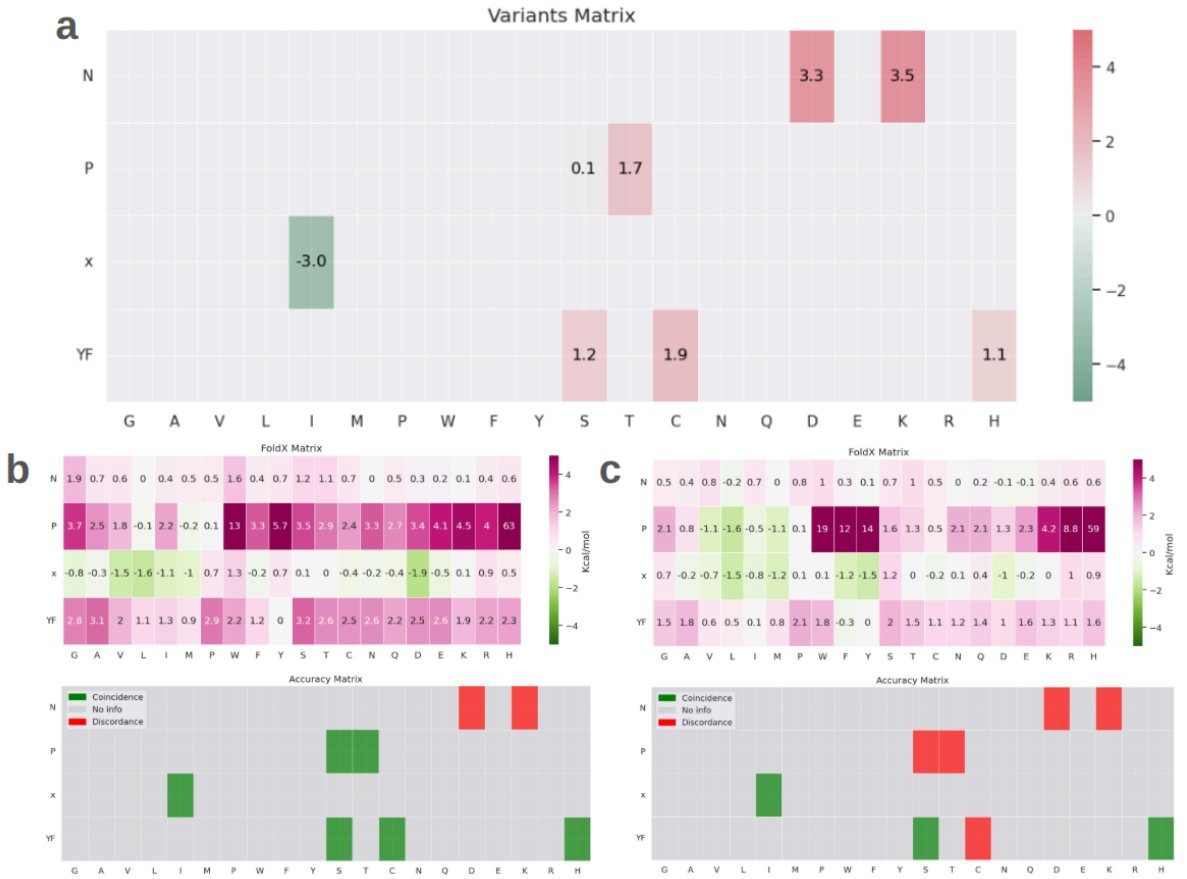

Figura 50. a) Matriz de patogenicidad/benignidad de variantes de ClinVar y gnomAD para el motivo LIG_PTB_Apo_2. En rojo se muestran las variantes patogénicas mientras que las benignas son mostradas en verde. b) Matrices para una predicción óptima de AF2. c) Matrices para una predicción errónea de AF2. En ambas, arriba, matriz de estabilidad indicando ddG en kCal/mol, debajo, matriz de coincidencias en la predicción de efecto. En verde se muestran las coincidencias mientras que en rojo están las disidencias.

Más allá de que como ya mencionamos la cantidad total de variantes analizadas en cada motivo derivadas de gnomAD y ClinVar es baja, como puede observarse en el panel inferior de la Figura 50b, cuando realizamos la comparación entre la matriz de variantes patogénicas/benignas y comparamos contra la matriz de estabilidad, observamos una gran cantidad de coincidencias (6/8) para las predicciones óptimas, representadas las coincidencias como celdas de color verde. Por su parte, en la Figura 50c vemos cómo predominan las disidencias entre ambas matrices (5/8), representadas como celdas de color rojo. Obsérvese que ambas matrices energéticas poseen un cierto grado de similitud, derivado de las diferencias fisicoquímicas entre aminoácidos y el impacto que conlleva en general su cambio, por esto cobra interés observar su desempeño al intentar clasificar variantes previamente reportadas en toleradas o no-toleradas. De la mano de esto, el hecho

que AF2 aplique su algoritmo y nos termine entregando una predicción para la interacción motivo-dominio, por más que la misma sea clasificada como errónea (como el caso de la Figura 50c), poseerá un cierto grado de coherencia biológica, y por lo tanto una parte de las variantes analizadas serán correctamente predichas en su efecto (en la Figura 50c, 3/8 variantes son coincidencias).

Se procedió a realizar la evaluación agregando todas las variantes en los motivos utilizados en la validación, calculando una métrica de exactitud global. La misma para las predicciones óptimas fue del 75% mientras que para las predicciones erróneas fue del 63%. Si bien la diferencia no es muy grande, es claro que la localización correcta del motivo impacta en la clasificación de las variantes.

Preparación de *inputs* y determinación de parámetros necesarios para la implementación de MotSASi 2.0

Una vez que contamos con los datos de validación mencionados previamente, procedimos a realizar las predicciones correspondientes a clases de motivos que no contaban previamente con estructuras cristalográficas asociadas depositadas en PDB. Decidimos quedarnos con aquellas clases de motivos pertenecientes a las diversas categorías de ELM (LIG, CLV, DOC, DEG, MOD, TRG) que tengan longitud fija y que no tengan ningún PDB asociado. Estos filtros dan como resultado un conjunto de 43 clases de motivos para analizar que agrupan un total de 198 instancias de motivos. De las mismas, 24 presentan al menos una instancia comprobada experimentalmente en humanos (sumando 100 instancias en el agregado) y fueron consideradas factibles de abordar con el pipeline de MotSASi.

Para aquellas clases de motivos que no contaban previamente con estructura cristalográfica, recurrimos a AF2 para generar las correspondientes predicciones. El problema radica ahora en que a la hora de analizar una predicción de AF2 realizada para un nuevo motivo sin estructura cristalográfica previa, no poseemos una referencia para clasificarla en óptima o errónea. Por lo tanto, el desafío constaba de analizar las predicciones confeccionadas en la validación previa con el fin de poder detectar una o más variables que sirvieran para clasificar predicciones generadas para estos nuevos motivos. Nuestros esfuerzos se orientaron a estudiar propiedades relativas a la topología y la termodinámica de la reacción de binding entre motivo y dominio. Analizando una cantidad significativa de variables, y sobreviviendo a la tentación de implementar un modelo de machine learning de muchas variables, concluimos el proceso con la elección de dos métricas:

1) ddG unión motivo-dominio: energía de interacción entre las cadenas del motivo y el dominio, el mismo fue calculado utilizando el software FoldX.
2) Media de pLDDT de los residuos del motivo implicados en la interfase con el dominio: pLDDT es un score de confianza de predicción calculado por residuo, es un dato que AF2 genera automáticamente para sus predicciones.

En la Figura 51 podemos observar como ambas variables se comportan en el grupo de las predicciones óptimas y erróneas.

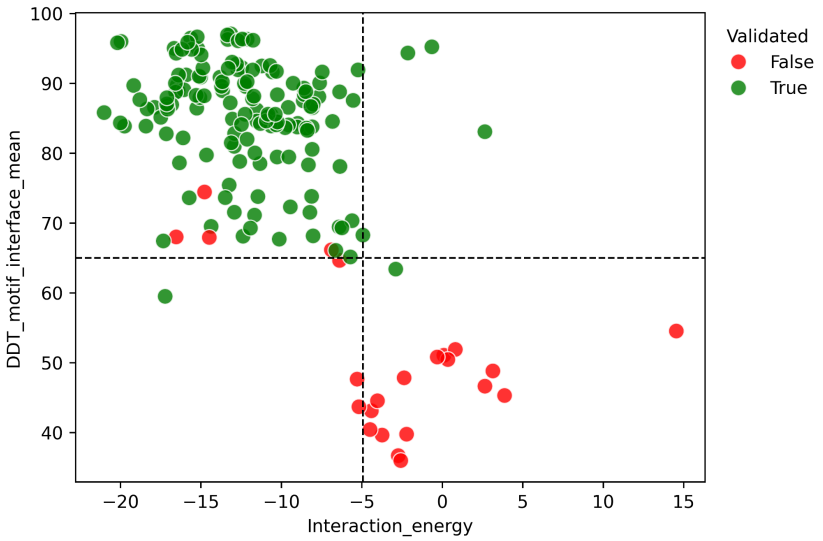

Figura 51. Diagrama de dispersión que muestra la relación entre la media de pLDDT de los residuos del motivo (eje y) y la energía de interacción en kCal/mol (eje x). Cada punto representa una predicción de AF2: los puntos verdes corresponden a predicciones clasificadas como reproducibles según una inspección manual y su acuerdo con la estructura cristalográfica correspondiente en el PDB, mientras que los puntos rojos representan predicciones incorrectas. Las líneas punteadas indican los valores de corte seleccionados para distinguir entre predicciones correctas o erróneas por AF2.

Como podemos ver, ambas métricas poseen una capacidad discriminativa interesante entre predicciones óptimas y erróneas, y aplicadas en conjunto nos permiten discriminar con facilidad, y un alto nivel de confianza, cuáles serán aquellos modelos generados por AF2 que podemos utilizar en el contexto de esta segunda implementación de MotSASi. Los valores de corte propuestos fueron -5 kCal/mol para la energía de unión (ddG en kCal/mol) y 65 para la media de pLDDT para los residuos del motivo implicados en la interfase. De esta manera, una predicción que registre un valor de ddG de binding por debajo del primer valor de corte y una media de pLDDT por encima del segundo valor de corte será utilizada para correr el pipeline de MotSASi.

Ponderando la evidencia aportada por ClinVar, gnomAD y FoldX en su ingreso a matrices de clasificación

Con respecto a la implementación de esta segunda versión de MotSASi, se realizaron algunas optimizaciones que vale la pena destacar. Por un lado, nuestro grupo de trabajo venía observando algunos usos y costumbres replicadas hacía tiempo en lo referido al tratamiento de las variantes en los algoritmos bioinformáticos. Por mencionar algunas:

1) Cuando se trabaja con variantes de ClinVar, siempre se las filtra en primera instancia por su estado de revisión, es decir, solo se utilizan variantes con un grado mínimo de evidencia. Una vez filtradas, en función de su veredicto se les asigna un puntaje (por ejemplo, “Patogénica” = 1, “Patogénica/Probablemente patogénica” = 0,75, “Probablemente patogénica” = 0,5).
2) Si se trabaja con variantes de gnomAD, suele utilizarse un valor de corte universal para la frecuencia, por encima del cual se establece que las variantes son muy probablemente benignas.

3) A la hora de evaluar un cambio, de contar con datos provenientes de ClinVar y gnomAD y también con predicciones sobre el mismo, se decide trabajar únicamente con los datos depositados en las mencionadas bases de datos.
4) Si se observan incongruencias entre los veredictos provistos por las bases de datos (veredicto patogénico por ClinVar y benigno por gnomAD) en variantes poco estudiadas, en general, al evaluar enfermedades poco frecuentes, el dato de frecuencia suele mostrarse correcto tarde o temprano.

Con todas estas premisas en mente, y con la expresa voluntad de contribuir a mejorarlas, decidimos idear un sistema de comparación de datos aportados por bases de datos (ClinVar y gnomAD), junto con las predicciones (FoldX) donde todos estos ingresan al algoritmo y cuentan a priori con posibilidades de poner a disposición sus aportes. Para lograrlo, decidimos traducir sus veredictos a un sistema de puntos que fueran reflejo de la confianza que nos entrega cada uno de los mismos. Para esto, los pasos a seguir fueron:

1) En primer lugar, decidimos tomar al conjunto de variantes missense de ClinVar y por medio de un script in-house nos quedamos con el estado de revisión de las mismas, como así también cuántas cargas había tenido en la base de datos. Esta información se tradujo al sistema de puntos asignando 1 por cada estrella del código de estado de revisión de ClinVar y 0,1 por cada carga. Los puntos asignados son positivos en caso de ser reportada la variante dentro de categorías patogénicas, negativos en caso de benignidad.
2) En segundo lugar, con el propósito de lograr relevar una mayor cantidad de variantes benignas de gnomAD, tomamos a los genes que codifican a las proteínas humanas contenidas en Swiss Prot y que cuentan con asociaciones a enfermedades poco frecuentes de tipo mendeliano por OMIM, y observamos cuántas variantes patogénicas tenían en ClinVar. Si estas eran menores a diez, directamente trabajamos a estos genes con un valor de corte universal calculado en base a frecuencias de variantes patogénicas y benignas de ClinVar residentes en genes asociados a modelo de herencia autosómico recesivo, el cual resultó de -log(FA) = 4,1. En caso de contar con más de 10 variantes patogénicas, procedimos a ordenar las mismas de mayor a menor frecuencia, e iteramos en orden decreciente de frecuencia por las mismas analizando su estado de revisión y cargas en la base de datos. Si la variante contaba con una gran cantidad de evidencia (tres o cuatro estrellas de estado de revisión de ClinVar, o una gran cantidad de submission) o no era del todo seguro su reporte como patogénica (muchas cargas de significado incierto o incluso benigno), se iba descendiendo en el valor de frecuencia hasta lograr un valor óptimo. La idea subyacente es que una variante patogénica en el mundo de las enfermedades poco frecuentes con un valor de frecuencia por encima de lo normal, probablemente ya cuente con una cantidad de reportes de patogenicidad, por lo que la misma información de ClinVar será la que la expondrá como patogénica, y que no limite la posibilidad de relevar otras variantes como benignas. Este algoritmo nos permitió individualizar el valor de corte de frecuencia por cada gen. Luego, los valores de frecuencia fueron traducidos a un sistema de puntos como se hizo con ClinVar. En este caso se optó por una función exponencial asintótica a -4 (determinado como el mínimo de puntaje asignable a una variante de gnomAD) y que pasara por el valor x = 4 en el valor de y = -2 de la función, como exponemos en la Figura 52. Observemos que los puntos asignados son siempre negativos.

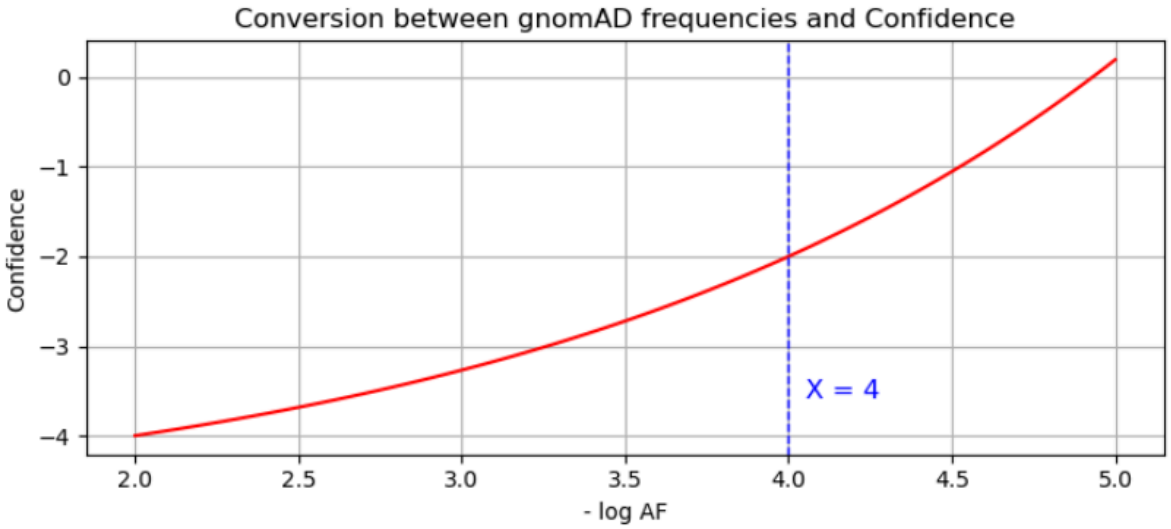

Figura 52. Función exponencial que traduce las frecuencias de gnomAD con el sistema de puntos utilizado para evaluar la confianza en el respectivo dato.

3) Para los datos de las predicciones provenientes de FoldX, el tratamiento del dato fue similar al que se realizó con las frecuencias de gnomAD, pero se pensó una función asintótica a +/- 2,5 (máximos puntajes atribuibles a predicciones de FoldX patogénicas y benignas, respectivamente). En el caso de analizar predicciones derivadas de estructuras depositadas en el PDB, se utilizó una función que corta al eje x en 2,1 kCal/mol (en consonancia con el valor de corte determinado). De la misma manera, en caso de trabajar con modelos derivados de AF2, la función corta al eje x en 1,4 kCal/mol. Obsérvese que los puntos asignados son positivos en caso de veredicto deletéreo mientras que son negativos en caso de ser el mismo no-deletéreo.

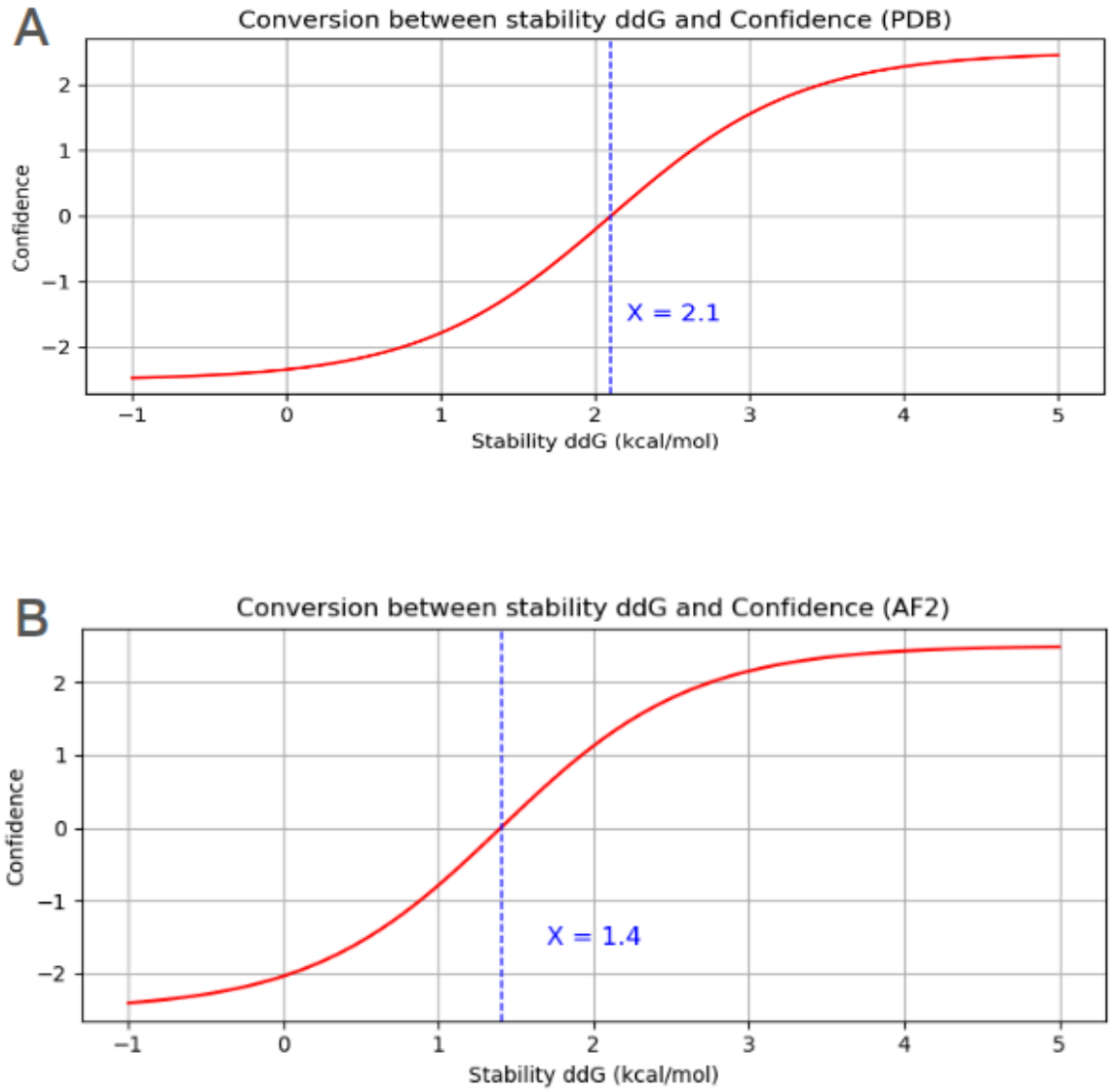

Figura 53. a) Función exponencial que traduce las predicciones termodinámicas de FoldX efectuadas sobre estructuras cristalográficas depositadas en el PDB (en kCal/mol) con el sistema de puntos utilizado para evaluar la confianza en el respectivo dato. b) Función análoga para predicciones derivadas de modelos de AF2.

De esta manera, a medida que los datos provenientes de una u otra fuente van ingresando en la matriz, en lugar de arbitrariamente excluir uno u otro, lo que se hace es dejarlos competir entre ellos en función de la confianza que tenemos en los mismos, traducida en puntos. Podemos ver en la Figura 54 cómo se genera una matriz generada por MotSASi en este proceso. Si para un dado cambio de aminoácido sólo una fuente de evidencia se pronuncia al respecto, entonces el mismo accede a la Matriz Final. Si dos o más fuentes de evidencia distintas coinciden en su veredicto, se procede a sumar ambas y ese valor arriba a la Matriz Final. En caso de contar con evidencias opuestas, se procede a realizar la suma de puntos positivos y negativos por separado, se comparan, y permanece aquel valor superior.

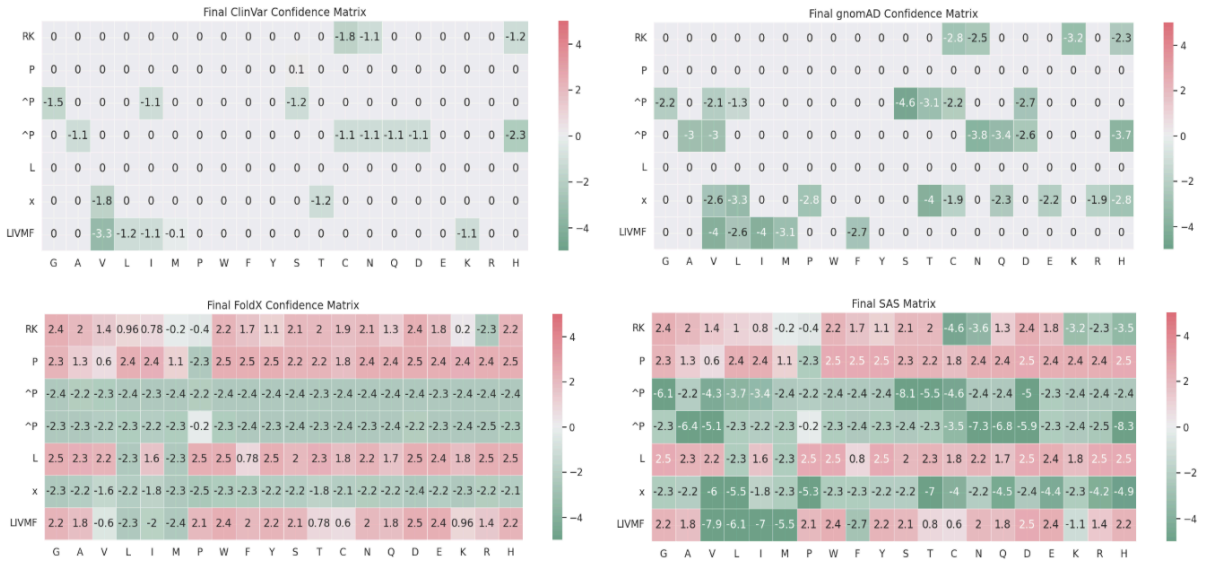

Figura 54. Ejemplo de matriz generada por MotSASi reflejando el sistema de puntos de confianza implementado para el motivo DOC_MAPK_JIP1_4. Las matrices de ClinVar, gnomAD y FoldX son confrontadas entre sí y producto de la comparación de los valores absolutos que encontramos en cada posición emergen los valores que son plasmados en la Matriz Final.

En segundo lugar de las diferencias implementadas en esta versión de MotSASi, se buscó utilizar al máximo las estructuras depositadas por el equipo diseñador de AF2 para las proteínas que componen el proteoma humano en SwissProt. Particularmente, se agregó un filtro correspondiente a la exposición al solvente que tuviera el motivo encontrado evaluándose en ese momento. Los SLiMs son unidades funcionales que para poder interactuar con su dominio correspondiente deben encontrarse expuestas en regiones con una gran flexibilidad, generalmente desordenadas. Por lo tanto, evaluamos dicha exposición luego de una minuciosa calibración del veredicto otorgado por el software freesasa. Esto nos permitió agregar una capa extra de confianza para aquellos motivos candidatos que pasaran todas las instancias de MotSASi.

La performance general de MotSASi 2.0

En este trabajo, se procesaron 51 clases de motivos con el pipeline de MotSASi, de los cuales 27 contaban con estructura cristalográfica depositada en PDB (los mismos que en la primera implementación de MotSASi), mientras que 24 correspondían a motivos que previamente no contaban con las mismas. En la Tabla 2 podemos apreciar las contribuciones de nuevas instancias de motivos que ambos grupos hacen al panorama general de los SLiMs, incluídas las instancias de motivos previamente reportadas en el ELM. Así mismo, se hace un desglosado de aquellas clases de motivos e instancias de las mismas con apariciones en la base de datos OMIM.

	MotSASi	PDB	AF2	OMIM
# Clases ELM	51	27	24	42
# Instancias ELM	405	305	100	199
# Coincidencias RE proteoma	377.170	165.481	211.689	124.171
# Instancias sin variantes (% del total)	315.957 (84%)	137.444 (83%)	178.513 (84%)	98.554 (79%)
# Instancias negativas (% del total)	52.482 (14%)	22.218 (13%)	30.264 (14%)	22.047 (18%)
# Instancias positivas (% del total)	8.731 (2%)	5.819 (4%)	2.912 (1%)	3.570 (3%)
# Proteínas	5.027	3.896	2.360	1.694
Enriquecimiento contra inst. ELM	X 21,56	X 19,08	X 29,12	X 17,94

Tabla 3. Número de tipos de motivos (# Clases ELM), SLiMs validados experimentalmente (# Instancias ELM), coincidencias de expresiones regulares en el proteoma humano de SwissProt (# Coincidencias RE proteoma) y su categorización: coincidencias sin variantes en ClinVar/gnomAD (# Instancias sin variantes), falsos positivos (# Instancias negativas) y SLiMs de alta confianza (# Instancias positivas). La fila "Proteínas" muestra el número total de proteínas humanas de SwissProt que contienen al menos un motivo por categoría. Las columnas especifican los conjuntos de datos analizados: MotSASi (conjunto de datos completo), PDB (motivos con estructuras cristalográficas), AF2 (motivos modelados con AlphaFold2) y OMIM (motivos en proteínas asociadas a enfermedades en OMIM).

Respecto de las instancias depositadas en el ELM, la implementación de MotSASi en motivos que ya cuentan con una estructura cristalográfica genera un enriquecimiento de instancias de motivo (o sea motivos identificados en proteínas específicas) de 19X (5.819/305), mientras que aquellas que fueron estudiadas en base a predicciones lo hacen en un factor de 29X (2.912/100). Vale la pena mencionar en este punto que resulta bastante más meritoria la predicción de nuevas instancias realizada en los nuevos motivos incorporados al análisis gracias a las predicciones, puesto que el promedio de instancias depositadas en el ELM para las mismas es de 4,17 instancias / clase mientras que para las clases de motivos con estructuras cristalográficas depositadas en el PDB, el mismo es de

11,3 instancias / clase. Si bien la cantidad de instancias aportadas por los motivos estudiados gracias a predicciones de AF2 es menor, su desempeño en términos de cuánto expande la frontera de conocimiento de instancias en el proteoma es mejor, y lo hace en clases de motivos que se encontraban menos estudiadas al día de la fecha.

Comparación de MotSASi contra AlphaMissense: ventajas claves

Por otra parte, los aportes realizados por MotSASi no se dan únicamente en términos de las instancias nuevas de motivos de alta confianza que puede hacer a la discusión general de identificar sitios funcionalmente importantes en el genoma. También aporta, a partir de sus matrices elaboradas gracias a la información de variantes previamente reportadas, y las predicciones termodinámicas, un conjunto de predicciones sobre todos los posibles cambios missense que pueden darse en los mismos. De esa manera, logramos aportar a la discusión un total de 1.041.656 cambios missense predichos con un veredicto acompañante, listos para funcionar como material de anotación para variantes detectadas en el contexto clínico de diagnóstico molecular de enfermedades poco frecuentes.

Nos interesó particularmente compararnos contra un predictor que combine información estructural además de las tradicionales informaciones de entrada como la conservación, estructura secundaria predicha, dominios presentes, etc. En ese contexto, resultó particularmente interesante una herramienta como lo es AlphaMissense (AM) [129]. Para realizar la comparación, procedimos a correr el pipeline sobre la totalidad de motivos pero en este caso dejando aparte un 20% de las variantes patogénicas obtenidas de ClinVar y las variantes benignas obtenidas de gnomAD, constituyendo este conjunto un set de evaluación independiente. El mismo se compuso de 2.628 variantes, incluyendo 2.449 variantes benignas de gnomAD y 179 variantes patogénicas de ClinVar. Luego, procedimos a evaluar el desempeño de AlphaMissense y MotSASi sobre las mismas. MotSASi superó a AlphaMissense en las tres métricas evaluadas: precisión (98% frente a 76%), puntaje F1 (0,99 frente a 0,85), y coeficiente de correlación de Matthews (MCC) (0,83 frente a 0,33). Es evidente que, aunque AlphaMissense es una excelente herramienta de predicción general para la patogenicidad de variantes, el análisis detallado de las estructuras subyacentes en un contexto biológico, como lo realiza MotSASi, mejora significativamente la precisión general de las predicciones.

	Veredicto MotSASi Patogénico	Veredicto MotSASi Benigno
ClinVar Patogénico	130	49
gnomAD Benigno	5	2444

	Veredicto AM Patogénico	Veredicto AM Benigno
ClinVar Patogénico	152	27
gnomAD Benigno	607	1842

Tabla 4. Matrices de confusión que comparan el rendimiento de MotSASi y AlphaMissense (AM) en un conjunto de validación de variantes.

El análisis detallado de la matriz de confusión muestra que la clave de la mejora observada en MotSASi radica en su capacidad para evitar clasificar variantes benignas conocidas

como patogénicas. El conjunto de datos naturalmente desequilibrado, con muchas más variantes benignas, como se observa a menudo en ejemplos del mundo real, resulta en que AlphaMissense clasifica erróneamente 607 variantes benignas como patogénicas, mientras que MotSASi sólo presenta 5 casos de estos posibles falsos positivos. Como era de esperarse, para las variantes patogénicas conocidas, aunque MotSASi no detecta 49 casos, AlphaMissense falla solo en 27. En términos generales, parece que MotSASi tiende a ser más conservador que AlphaMissense respecto a la clasificación patogénica, lo cual está alineado con las recomendaciones actuales de ClinGen.

Conclusiones

Este capítulo de resultados tiene quizás como máxima enseñanza el hecho de que combinando diferentes conceptos teóricos y herramientas bioinformáticas, se puede modelar un problema biológico con alto grado de precisión. La coincidencia de secuencias proteicas humanas con la expresión regular de clases de motivos, junto a los estudios de conservación, exposición al solvente o estructura secundaria, permiten establecer requisitos mínimos para considerar un posible candidato. La utilización de datos provenientes de varias fuentes −en nuestro caso de bases de datos de variantes de interés clínico y frecuencia alélica poblacional− junto a predicciones termodinámicas como las de FoldX permite indagar en los pormenores al interior del mismo, y a su vez reverberar el proceso. Gracias a este trabajo, se resuelve más de un problema al mismo tiempo. Por un lado, se incrementa la cantidad de secuencias compatibles con representar SLiMs funcionales. Por otro lado, se brindan herramientas concretas, las matrices de sustitución finales, para su utilización en la interpretación de patogenicidad de variantes residentes en los mismos. Nótese que ambos datos resultan imprescindibles a la hora de dicha interpretación, encuadrados dentro de criterios enunciados por las recomendaciones de la ACMG (PM1, referido a sitios funcionales relevantes y PP3/BP4, predicciones bioinformáticas de variantes).

Otro aspecto clave en la implementación de MotSASi es el de llenar los vacíos que el algoritmo como tal pudiera tener. Con respecto a las variantes, es cuestión de tiempo que cada vez contemos con más y más datos de las mismas, gracias a la expansión general que el rubro genómico experimenta. Pero por otro lado, se demostró que uno puede tomar las herramientas de inteligencia artificial desarrolladas en los últimos años, como AlphaFold2, y utilizarlas para resolver problemas biológicos más complejos, como entender dónde y cómo es la interacción de un motivo con respecto a un dominio. Por medio de una validación intuitiva con respecto a las estructuras cristalográficas ya resueltas y depositadas en el PDB, logramos evaluar la confianza en los modelos generados por AF2 y así estudiar clases de motivos que de otra manera no hubiera sido posible.

Por último, resulta interesante destacar que la predicción bioinformática de patogenicidad de variantes experimenta un impulso sensible con estos desarrollos. La mayoría de los predictores desarrollados busca en general resolver un determinado problema planteado en términos de su efecto molecular (variantes missense, de splicing, codones stop prematuros). Si bien extremadamente útiles, no todas las variantes en estos conjuntos involucran el mismo problema biológico, y justamente un adecuado modelado del mismo (interacción SLiM-receptor), combinado con múltiples enfoques de su análisis (genómico y estructural), permite hallar respuestas más precisas y confiables.

Capítulo 4: Discusión

Revisión Sistemática

En el capítulo 1 de resultados se expuso el trabajo relativo a la revisión sistemática. Producto de la misma se originaron dos publicaciones, una orientada específicamente a los fenotipos clínicos emergidos de la misma, y una segunda orientada al panorama genético relevado. No podemos dejar de observar que ambas constituyen dos caras de una misma moneda, y que es justamente su interrelación la que adquiere una mayor riqueza y valor agregado. La relación genotipo-fenotipo es de vital interés para la comunidad médica y científica, donde el estudio asociativo entre gen-patología ya está hilando al punto de trabajar en encontrar asociaciones entre variantes de determinado efecto molecular o localizadas en dominios específicos de proteínas, con sub-presentaciones clínicas específicas. En una patología con una documentación notable de pacientes en los últimos 50 años, resulta muy valioso contar con un compendio genotípico-fenotípico exhaustivo del mismo.

En materia de análisis de variantes genéticas por medio de las recomendaciones ACMG, debemos comenzar por decir que contar con una revisión sistemática donde se establecieron criterios de inclusión/exclusión bien definidos para hipogonadismo hipogonadotrófico, resulta fundamental a la hora de aplicar un criterio de evidencia igualmente tradicional como importante, a saber, contar casos previos de pacientes afectados con la patología que presentaron la misma variante (PS4 y PM3). Quizás puedan pasar un poco desapercibidas como problemáticas, pero existen una serie de situaciones a las cuales como analistas nos enfrentamos cuando resolvemos un caso. Una de particular sensibilidad es el de reconocer una cierta especificidad del fenotipo asociado al gen donde se localiza la variante en estudio, siendo este patognomónico (PP4) o no. Si bien no existe una guía general para la contabilización de esta evidencia emitida por ACMG, sí encontramos una serie de grupos especializados en diferentes patologías que han adoptado una serie de consideraciones a la hora de contar casos previos. Por ejemplo, el grupo especialista en variantes en el gen CDH1 establece una serie de criterios clínicos a la hora de evaluar un caso reportado previamente. Más interesante aún, el grupo especializado en variantes en PTEN, utiliza guías clínicas de puntuación de fenotipo para decidir si cada caso reportado previamente lo contabilizará como un caso completo (1 punto) o, si su compatibilidad fenotípica con el cuadro es parcial, como medio caso (0,5 puntos). Evidentemente, la evaluación clínica del caso que estamos analizando contar como evidencia no es un dato para nada menor, el hecho de contar con una base de datos que sabemos está altamente enriquecida en casos de hipogonadismo hipogonadotrófico, curada por nuestro equipo médico, nos da la confianza para entender que los datos presentes pueden utilizarse a la hora de asignar el nivel de evidencia para este criterio.

En lo respectivo al nivel de evidencia que asignamos, precisamente debemos tener claro que el CHH es una enfermedad poco frecuente, y es en función de esto que nos comparamos con otras patologías para setear los mismos (2, 4, 8, 16 casos para asignar niveles de PS4_Supporting, PS4_Moderate, PS4 y PS4_VeryStrong, respectivamente). Obsérvese que el nivel de astringencia que se plantea, respecto de la utilización convencional de PS4, donde de contarse con reportes previos de la variante, sin una cuantificación precisa, automáticamente se asignaba el nivel Strong. Todo esto se ve reflejado en los resultados y en las decisiones que tomamos en la revisión sistemática.

Muchas veces se percibe en el tratamiento de las variantes que se realiza en diversos trabajos científicos, que existen criterios que sólo pueden aplicarse al tipo específico de variante para el cual se escribieron en 2015 en el paper original de ACMG.

Esta misma organización se convirtió en los últimos años en la principal protagonista de un proceso de retroceso de las mismas reglas que había establecido. Quizás uno de los grandes conceptos que debió militar fue el hecho de que, al analizar una variante presente en un individuo, el objeto de estudio no es precisamente la aparición de esa variante en ese individuo específico, sino que es la variante en sí misma, como entidad abstracta la que debemos estudiar, y que su aparición en uno u otro individuo, son simplemente instancias físicas de la misma, donde se encuentra rodeada de entornos genéticos diversos.

De la misma manera, algo interesante se gestó con respecto a las etiquetas PS1, PS4 y PM5. Sobre PS4 ya hemos hablado, se refiere a apariciones de la variante en estudio en casos previamente reportados. PS1, por su lado, es una etiqueta que refiere a variantes missense (una de las peores pesadillas de los analistas) patogénicas reportadas previamente, en la misma posición de la proteína y donde el cambio nucleotídico es diferente pero el cambio aminoácido es idéntico. Por último, PM5 se utiliza para variantes missense patogénicas, previamente reportadas en la misma posición de la proteína, donde tanto el cambio nucleotídico como el aminoacídico son diferentes. No cabe ninguna duda que en aquellos años, donde las herramientas de análisis bioinformático eran más limitadas (particularmente a los profesionales no familiarizados con la programación), las variantes missense representaban un problema frecuente y complejo, con lo cual mucha de la energía se puso en su análisis. Estas etiquetas, que uno puede tomar simplemente como meros ítems a tildar en función de la información presente en bases de datos biológicas o motores de búsqueda, representan una hipótesis biológica concreta, es decir, que el cambio missense está generando una alteración concreta del producto proteico del gen. Nótese que es una hipótesis de trabajo extremadamente amplia pues no estamos diciendo nada en lo referido al mecanismo de patogenicidad asociado, como podría ser la alteración de la estabilidad del polipéptido, de la unión de un sustrato, de la interacción con un partner proteico o nucleotídico, las opciones son múltiples. Sin embargo, las etiquetas originales tenían esta premisa subyacente (nótese que incluso, en el paper de 2015 de ACMG, se ponía como requisito comprobar que el mecanismo de enfermedad no fuera una alteración del splicing, ya que la lógica de la etiqueta carecería de sentido), y la gran pregunta que debemos hacernos es si la misma no es aplicable a otros tipos de variantes. Por ejemplo, todos sabemos que la deleción de la fenilalanina 508 en CFTR es patogénica porque no se da el target del transportador en membrana. Si estudiamos una variante que deleciona los residuos 507, 508 y 509 del mismo (nótese que abarca a la deleción patogénica previamente reportada), ¿no constituirá esto una evidencia en favor de la patogenicidad de nuestra variante? ACMG viene pensando esto hace tiempo, y uno de los avances más importantes que hizo al respecto fue en materia de análisis de variantes de splicing, donde la presencia de variantes en las mismas regiones canónicas de splicing podían constituir evidencia en pos de la patogenicidad de la variante en estudio. Entendiendo este panorama, se comprende porqué el hecho de contar con una base de datos de refinado curado clínico, constituye una ventaja pues nos permite tomar decisiones inteligentes que nos permitan sumar o desestimar evidencias a determinadas variantes en estudio.

En un ateneo reciente, analizando una variante nonsense sensible a NMD que solo presentaba a su favor su efecto molecular y baja frecuencia (compatible con una clasificación Probablemente patogénica por ACMG), se planteó una consideración especial por parte de una médica del servicio que manifestó su descontento, pues en posiciones inmediatamente adyacentes a la misma se contaba con reportes de otras variantes nonsense clasificadas como patogénicas. Su observación era justa, ¿cómo podía ser que esa información, claramente relevante, no contara como evidencia a favor de la

patogenicidad de la variante que cargaba su paciente? En definitiva, el mecanismo molecular era idéntico, y se comprobaba que la vigilancia contra stops prematuros por NMD era operativo en dicho exón. Resulta absolutamente razonable que esta información sea considerada, como así también, que la misma aparezca en el informe, pues les es de valor tanto al médico como al paciente.

Finalmente, existen algunas piezas de evidencia que nos hubiera gustado incorporar al análisis de revisión sistemática, pero que lamentablemente no pudimos, debido a diversas razones, entre las que podemos mencionar como de relevancia: 1) que ante la ponderación esfuerzo/rendimiento, decidimos no seguir adelante con el análisis, 2) que hubiera una ausencia de valor agregado a los datos ya recolectados, 3) se evidenciara una disponibilidad limitada a ciertos datos. Como ejemplo de los mismos, podríamos citar el caso de las ocurrencias de novo, o la segregación de la variante en diversos integrantes de la familia. Muchísimas veces, principalmente en los grandes estudios de cohortes de familias, está información no se encuentra disponible, por lo que aunque hubiésemos querido, no podríamos haber accedido a la misma para todas las variantes. Por otro lado, también podemos argumentar que si a los pacientes que presentaron la variante previamente ya los estamos contabilizando en PS4, entonces corremos el riesgo de estar realizando conteo doble de evidencia.

Otra fuente de evidencia que en general resulta muy atractiva es la presencia de ensayos funcionales relativos a la variante en estudio. Históricamente, la misma constituía un nivel de evidencia fuerte (PS3) en caso de arrojar un resultado deletéreo sobre el producto del gen. Sin embargo, debemos decir que la última recomendación de ClinGen sobre ensayos funcionales fue extraordinariamente estricta y conservadora con los resultados de ensayos funcionales, bajando drásticamente los niveles de evidencia asignados. Analizando el global de la evidencia funcional hallada en literatura, la enorme mayoría de los reportes no cuenta con la cantidad de controles, réplicas, variantes adicionales testeadas, como para llegar a cumplir estos estándares, o a lo sumo aportar una pieza de evidencia de nivel Supporting. No obstante lo cual, no vemos este hecho para nada como una pérdida irremediable, por dos motivos: en primer lugar, a la hora de interpretar una variante con el objetivo de informar, el analista debe obligatoriamente realizar un rastrillaje de los motores de búsqueda, bases de datos biológicas y bibliografía referidos a la variante. Y dado el alto grado de conectividad entre las mismas, resulta absolutamente factible que llegue rápidamente a la información y pueda verla por sí mismo. Repetimos, en la mayoría de los casos, o no podrá sumar evidencia por no cumplirse los requisitos de la recomendación, o el máximo nivel que pueda asignar no pasará de una etiqueta de nivel Supporting (PS3_Supporting). En segundo lugar, se podría argumentar con razón, que independientemente de la inclusión de esta evidencia o no en el informe, podría resultar muy útil la aparición de la etiqueta PS3, a la hora de priorizar la variante encontrada en el contexto de la búsqueda de la variante causal. Claramente, el hecho de que una variante en algún momento haya resultado de interés clínico para la comunidad científica la promovió en su estudio funcional, dando como consecuencia que, independientemente de la calidad del ensayo y su resultado, resulte una hipótesis atractiva a estudiar. Pues bien, en ese sentido, una reciente publicación emergida de nuestro laboratorio arrojó un resultado muy interesante, y es que, a la hora de evaluar un estudio en el que se comparaba la performance entre dos algoritmos de anotación de variantes usando criterios ACMG, la utilización de la etiqueta PP5 (variante reportada como patogénica en bases de datos pero sin acceso a su fundamentación) ajustada por la cantidad de evidencia (en definitiva, presencia de entrada patogénica en ClinVar y la cantidad de contribuciones de información

a la misma), funcionaba como un excelente proxy para emular los aportes dados por etiquetas como PS3 (ensayos funcionales) o PS4 (pacientes reportados previamente con la variante), y así lograr coincidir con las clasificaciones que expertos habían operado sobre un set de variantes extremadamente curadas. Esto puede hacerse absolutamente independiente de la información que luego aparezca en el informe, donde de hecho, la recomendación de ACMG es que la etiqueta PP5 (y su análogo benigno, BP6) quede en desuso. Por lo tanto, entendemos que a la hora de priorizar, existen otras maneras de poner variantes causales en franca exposición.

El análisis de firmas de oligogenicidad no logró aportar un conocimiento adicional en lo relativo a la etiopatogenia del cuadro, sin embargo, no es tampoco un resultado que esperábamos a priori encontrar. La realidad es que la mayoría de los trabajos de los cuales extrajimos combinaciones de variantes realizaron experimentos de secuenciación con paneles o exomas, investigando genes ya previamente asociados a hipogonadismo hipogonadotrófico. En general, al igual que en multitud de otras enfermedades genéticas, muchos de los genes estudiados surgieron como consecuencia de encontrarse en un primer pool de genes con variantes con mayor frecuencia, que luego se intentó ampliar por medio de análisis topológicos en redes génicas o vías de señalización. También debemos decir que nos encontramos con filtrados muy laxos de las variantes informadas, casi nulidad de análisis del efecto combinado de variantes en genes específicos (muchos extraídos de papers que informaban cohortes de pacientes) y, al final del día, un número aún bajo de combinaciones. No pudimos hallar un enriquecimiento en combinaciones siguiendo un determinado patrón, pero lo que sí nos quedamos, es un registro de las combinaciones reportadas en modo general en literatura. Y más importante aún, utilizando este set de combinaciones relevadas de publicaciones donde los pacientes fueron comprobados con los criterios de inclusión y exclusión para la revisión sistemática, pudimos observar que una herramienta como VarCoPP presentó un 92% de sensibilidad en su tarea de identificarlas como posibles causales de enfermedad. Esta sensibilidad aceptable nos permite trabajar con cierta confianza a la hora de abordar el estudio de un nuevo caso, utilizando este algoritmo como una buena herramienta de priorización de las combinaciones presentes en el mismo. Pensando en una futura aplicación de estas herramientas a casos donde no se encuentren alteraciones genéticas que puedan explicar el cuadro en términos monogénicos (por ejemplo mediante SNPs, indels o CNVs implicando un único gen), resulta de vital importancia una herramienta que permita sistematizar la priorización de combinaciones, invitando a la comunidad a relevar las mismas para una posterior interpretación en detalle.

Recordemos que el algoritmo VarCoPP y la plataforma donde se encuentra implementado, ORVAL, no son utilizados como criterio diagnóstico en lo más mínimo. Como ya mencionamos previamente, sería equivalente a dar un diagnóstico en función de una predicción deletérea de un predictor bioinformático como REVEL o SpliceAI para una enfermedad monogénica. Lo que en última instancia reviste verdadero valor es el trabajo de curación posterior en el cual se analizan los distintos tipos de evidencia disponibles. En ese sentido, podemos decir que muchas de las combinaciones que relevamos en nuestra revisión sistemática ya fueron incorporadas por OLIDA a la base gracias al enorme trabajo realizado por sus curadores en sus inicios, que realizaron un trabajo complementario al nuestro analizando publicaciones conteniendo asociaciones a oligogenicidad en general. Por lo tanto, la mayoría de combinaciones en publicaciones en condiciones de ingresar a OLIDA con cierto grado de curación ya fueron analizadas y curadas. Eso además implica que las relaciones sinérgicas entre los genes de los mismos ya fueron curadas, y por lo

tanto estarán disponibles a la hora de evaluar combinaciones de variantes localizadas en los mismos.

Referido a este último punto, vale la pena aclarar que en general las combinaciones de variantes residían en genes previamente asociados a la patología, y como ya mencionamos, muchos de ellos fueron progresivamente descubiertos por pertenecer a las mismas vías de señalización o mostrar cercanía en redes génicas, es decir, en general se documentaron genes en epistasis directa. Estas mismas combinaciones fueron las que a su vez se usaron al entrenar VarCoPP. Por lo tanto, se plantea el interrogante sobre nuestra capacidad, y la del algoritmo, para poder identificar epistasis indirectas entre genes, como por ejemplo, variantes localizadas en genes que pertenecen a procesos biológicos diferentes pero que tienen un impacto fisiopatológico común (por ejemplo, variantes en genes ligados a migración neuronal y fitness metabólico que combinadas alteran la normal fisiología de la neurona GnRH). En este sentido, los análisis topológicos u ontológicos pueden mostrar insuficiencia, y la única alternativa a contar con pruebas concretas de dichas sinergias se remite a la realización de ensayos funcionales. Hoy día esta es una de las grandes limitantes en el mundo de la oligogenicidad, pues conllevan la realización de múltiples pruebas paralelas en un sistema biológico adecuado, in vivo o in vitro. Pensemos que mínimamente deberían testear: a) WT, b) Knock-out para el gen A, c) Knock-out para el gen B, d) Knock-out para los genes A y B, e) Variante en gen A, f) Variante en gen B, g) Variantes en genes A y B. Estamos hablando de una cantidad de recursos e infraestructura considerables a los cuales no todos los laboratorios tienen acceso. Sin embargo, es esta una fuente de información de primera mano que no solo permite que combinaciones con variantes en genes con epistasis indirecta puedan llegar a altos scores de confianza, sino que además redundará en información representativa de las mismas que a su vez sirva para un mejor entrenamiento de algoritmos como VarCoPP.

Con la perspectiva de que los análisis de combinaciones oligogénicas logre una estandarización adecuada y pueda aspirar a figurar en los informes diagnósticos con entidad propia, y no como meros hallazgos circunstanciales y casuales hallados en el caso, resulta fundamental contar con las herramientas adecuadas para lograr esta curación. Un paso muy importante fue dado por OLIDA, no solo al establecer la base de datos de combinaciones oligogénicas, sino que dentro de la misma establece asociaciones que potencialmente pueden ser transferidas a una multitud de combinaciones de variantes, como la evidencia de sinergia en alteraciones en combinaciones de genes. Por otro lado, tenemos que decir que existen herramientas aún no disponibles abiertamente al público como la consulta de frecuencia poblacional de combinaciones de variantes en poblaciones de referencia. Hoy día contamos con una base de datos de referencia a estos fines como lo es gnomAD, la cual hoy día podemos fácilmente consultar por una variante particular. Sin embargo, la estructuración de un sistema de *querys* para combinaciones de variantes es un objetivo infinitamente más desafiante desde el punto de vista computacional, pues pensando el producto cartesiano de variantes, hace escalar las dimensiones del problema exponencialmente. Los intentos hechos hasta la fecha por gnomAD por dar alguna solución parecida, como consultar frecuencia de combinaciones dentro del mismo gen (pensado para enfermedades monogénicas), resultan insuficientes a nuestros propósitos. Soluciones alternativas como buscar la combinación en los cerca de 2.500 experimentos que integran el Proyecto 1000 Genomas, aporta una evidencia útil pero donde no podemos dejar de lamentar la omisión de un cúmulo riquísimo de información contenida en gnomAD (principalmente la versión 4 que cuenta con más de de 730.000 exomas y 76.000 genomas). Reducir el universo de variantes consideradas al menos a aquellas localizadas

en secuencia codificante, de baja frecuencia, no-sinónimas, podría reducir momentáneamente la complejidad del problema como estrategia para contar rápidamente con esta información, no obstante lo cual todos comprendemos las limitaciones que impondría al análisis.

Cohorte de pacientes

El trabajo inherente a la presente tesis, al igual que la enorme mayoría de aquellas referidas al campo de las ciencias exactas y naturales, se basa principalmente en aportes realizados al saber biológico y también médico, a fines de optimizar nuestro conocimiento de la genética molecular buscando una aplicación que, en este caso, pretende lograr mejores diagnósticos moleculares de enfermedades poco frecuentes. Sin embargo, el trabajo realizado en estos años queda marcado por el desafío que representó el armado desde los cimientos de toda una nueva estructura, en un hospital público, de un servicio que brindó, y hoy en día brinda de manera regular, modernos estudios moleculares, no solo al servicio de Endocrinología Pediátrica del Hospital de Niños Ricardo Gutiérrez, sino a todos los servicios del mismo. Este hecho, que parece sencillamente un dato anecdótico, sienta las bases del éxito: que una iniciativa como la llevada a cabo finalice en el establecimiento de un efectivo centro de estudios moleculares de alta complejidad, entre ellos NGS (desarrollado principalmente en esta tesis), pero también de otras técnicas como Array-CGH que resultan complementarios al mismo.

Montar este tipo de servicios representa un desafío inmenso, pues lejos está la disponibilidad de recursos económicos de ser el único obstáculo, al que se suman como factores adicionales: el establecimiento físico del mismo, la logística de recolección de muestras, la puesta a punto de los protocolos de procesamiento de las mismas, el procesamiento informático de los datos producto de la secuenciación, por solo mencionar algunos de los problemas sensibles fundamentales de resolver en la preanalítica. Luego de todo eso, nunca debemos olvidar mencionar la interacción obligada entre los individuos más ligados al análisis bioinformático con los médicos derivadores, tratantes o indicadores del estudio. En relación con la capacidad de reconocer a los pacientes candidatos al estudio de NGS, la confianza siempre fue máxima con los profesionales médicos pues bien se sabe de la calidad de formación que reciben aquellos que trabajan en nuestros nosocomios públicos. Es decir, el rol fenotipificador y seleccionador se encontraba cumplido. Pero por otro lado, tenemos que observar que en el sistema de educación universitaria, lamentablemente los programas de estudio de la carrera de Medicina aún se mantienen reticentes a incorporar materias referidas a Biología Molecular en general, y ni hablemos de Genómica Clínica. Frecuentemente, los médicos recién laureados deben concurrir por su cuenta a recibir estos conocimientos en diversos cursos de post-grado.

En ese contexto, una de las tareas más importantes y exitosas llevadas a cabo por nuestro grupo de trabajo fue el de no separar aguas entre médicos y analistas de NGS, sino que se consolidaron equipos de trabajo multidisciplinarios, con transmisión de conocimiento bidireccional. Esta interacción se materializó, en el caso del grupo especializado en Hipogonadismo Hipogonadotrófico, en ateneos semanales donde se discutió paciente por paciente y se lograron consensos de trabajo en lo referido a la entrega del informe final al paciente. También se realizaron capacitaciones específicas para los médicos a fines de establecer un lenguaje y criterios comunes en lo referido al análisis de variantes genéticas. La consolidación de este equipo especializado, es un factor clave en la tasa de éxito

diagnóstico registrada, alrededor del 50%, compatible con las mayores tasas de éxito registradas tanto para la enfermedad como para la Genómica Clínica en general.

Sin embargo, no sería fiel a la realidad dar un panorama donde no se registraron dificultades, o incluso errores. La puesta a punto de las corridas de NGS no fue para nada sencilla, siendo que durante un tiempo considerable se debieron implementar múltiples correcciones en lo referido a la elección de la estrategia de secuenciación (se inició con exoma clínico, y luego derivó en un panel para la patología estudiada y a otras que se estudian en el hospital), al armado de las bibliotecas, al número de clusters generados en el experimento, a la cobertura de los genes de interés, el seteo de umbrales de sensibilidad y especificidad para el llamado de variantes, etc. Por otro lado, la capacitación continua del recurso humano ligado a los análisis es una tarea personal y tiempo-intensiva. A esto se suma que la arquitectura genética de la patología se vislumbraba borrosa dado los postulados que por entonces se planteaban en lo referido a la oligogenicidad. Todos obstáculos que fueron incidiendo principalmente a nivel temporal, donde múltiples procesos sufrieron demoras significativas. Como ejemplos, podríamos citar el análisis de variantes de baja calidad que nunca confirmaron por Sanger, o el temor a dar un informe negativo a un paciente con variantes con grado de evidencia intermedia o que no cumplieran con la cigosidad esperada. Esto derivó lógicamente en tiempos de retorno de análisis (en inglés TAT, turn around time) visiblemente dilatados. Sin embargo, podemos decir que la mejora continua del grupo permitió lubricar los procesos de logística, procesamiento, interpretación y elaboración de informe, reduciendo de manera significativa estos tiempos.

Yendo al resultado de los pacientes analizados, las implicancias de un equipo de trabajo consolidado sigue reflejando una lógica en lo referido a sus tasas de éxito. El número global indica un 49% de diagnósticos moleculares positivos para pacientes con hipogonadismo hipogonadotrófico. Sin embargo, si realizamos una diferenciación entre pacientes que manifestaron signos y síntomas extra-gonadales extra-olfatorios manifiestos (muchas veces denominados hipogonadismos hipogonadotróficos sindrómicos), y aquellos que no lo hicieron (CHH), vemos que entre los primeros la tasa de éxito alcanza valores en torno al 65% mientras que en el segundo grupo se registran tasas de éxito del 44%. Lo que llama la atención, es que, si dentro de esta cohorte de pacientes diferenciamos entre pacientes con o sin fenotipo olfatorio (Síndrome de Kallmann o CHH normósmico -CHHn-, respectivamente), la tasa de éxito permanece en el mismo número. Interesantemente, en diversas publicaciones anteriores, siempre los pacientes con Sindrome de Kallmann presentaban mayores rendimientos diagnósticos, en torno al 50%, mientras que los paciente con CHHn se posicionan en torno al 35%. Más allá de que nuestra casuística dista aún mucho de poseer una potencia estadística óptima, no deja de llamar la atención la aproximación de estos números.

Pensemos que en el momento en que un paciente comienza a ser estudiado y se evidencia la presencia conjunta de un retraso puberal junto con un cuadro de anosmia o hiposmia, la sospecha diagnóstica se dirige ineludiblemente a estos cuadros donde la neurona GnRH no ha logrado transitar óptimamente su migración y establecimiento en el hipotálamo. Podemos decir que es una presentación patognomónica de los mismos, y en virtud de su estudiada etiología genética, será por lo tanto muy probable que se indique su estudio por NGS. Pero la situación se torna diferente cuando un retraso puberal se presenta sin manifestaciones olfatorias u extra-gonadales en general. Comienza a volverse mucho más difícil distinguir entre un cuadro de etiología genética, y las múltiples etiologías adicionales que pueden originarlo, como causas orgánicas (régimen de transfusiones crónicas, tratamiento crónico con corticoides, desviaciones de la función tiroidea), u otras

responsables de HH funcionales (trastornos de la alimentación, exceso de ejercicio físico, estrés crónico). En este sentido, contar con un equipo médico a la altura, riguroso en el estudio clínico, bioquímico y por imágenes del paciente, permite plantear una hipótesis de trabajo harto más concreta a la hora de estudiarlo por NGS. Es decir, un buen trabajo fenotipificador permite seleccionar mejor los pacientes, lo cual redunda en mayores tasas de éxito para grupos como el nuestro donde se estudian Trastornos del Desarrollo Sexual, además de una optimización en la utilización de los recursos materiales y humanos disponibles.

En el capítulo 2 de esta tesis se desarrollan 5 casos, seleccionados básicamente en función de lo atractivo al análisis que los mismos pudieran resultar. Sin embargo, no podemos dejar pasar por alto que los mismos no son una representación fiel de la frecuencia del tipo de alteraciones del conjunto de los casos resueltos por nuestro grupo. Los casos donde se detectaron CNVs fueron menores al 20% del total de los mismos. Los casos donde se detectaron dos variantes con fuerte evidencia de interacción entre sí para postular un modelo oligogénico fueron menores al 20%. En general, las variantes detectadas fueron en su mayoría del tipo missense y se localizaban en un único gen, considerándose causales con alta confianza y un modelo de herencia típicamente mendeliano.

Algo interesante para destacar es que incluimos como ejemplo un caso de una variante missense en el gen CHD7. Este gen, que codifica para una helicasa, constituye un bello caso de estudio dadas sus particularidades que nos dejan comprender porqué es tan importante saber diferenciar el momento en el cual dejar de utilizar las incontables y redundantes herramientas bioinformáticas que nos ofrecen en modo masivo los distintos servidores, y tomar la posta de llevar adelante un estudio mucho más detallado de la variante en cuestión. Este gen codifica para una helicasa de gran tamaño, estamos hablando de una proteína de casi 3.000 residuos de aminoácidos. Ya de por sí, esto trae como primera consecuencia que se vuelve increíblemente común encontrar al menos una variante missense por caso, que obviamente deberemos estudiar. Pero por otro lado, sabemos que en este gen la gran mayoría de variantes patogénicas reportadas son de tipo truncante, es decir, variantes nonsense, frameshift o de splicing. Esto guarda alguna relación con el hecho de que, tratándose de una proteína, aparenta ser más probable que la misma logre acomodar las inestabilizaciones energéticas generadas por variantes missense de mejor manera, y que resulte este panorama en el que variantes que hagan decaer el transcripto entero por NMD sean las que predominen como responsables de su pérdida de función.

Sin embargo, cada tanto se nos presentan casos como el descrito, donde una variante missense localizada en un sitio con una implicancia funcional concreta pueda mostrarse explicativa del fenotipo observado. CHD7 presenta dominios (en tanto definidos por Pfam) que, agrupados, en términos de secuencia, no llegan a cubrir el 20% de la misma. Uno de ellos es el SNF2-related domain que le permite a la helicasa, por medio de la hidrólisis de ATP, desestabilizar la interacción ADN-histona permitiendo que factores de transcripción puedan acceder al primero. Existe una tentación muy difundida entre los analistas genómicos de asociar la localización de una variante missense en un dominio funcional con un rol funcional del residuo que codifica, colocando consecuentemente la etiqueta PM1 de ACMG. Esto no siempre es así, y la evidencia demuestra que ni todas las variantes missense residentes en dominios son patogénicas, ni todas las variantes missense residentes fuera de ellos son benignas. Uno debe tener cuidado al aplicar esta evidencia, y el secreto para hacerlo es tener absolutamente en claro el mecanismo de

patogenicidad que uno postula para su variante, y cómo el mismo es abarcado por las herramientas con las que uno cuenta. Muchas personas deciden agregar una capa extra de evidencia para etiquetar PM1, consistente en comprobar la conservación del residuo dentro del dominio, observando los AMS de Pfam o sus respectivos logos. Sin embargo, no toman en cuenta que la conservación es uno de los outputs con contribución marginal más importante a los predictores (y metapredictores) bioinformáticos que suelen utilizarse, REVEL a la cabeza en lo que respecta a la variante missense. De esta manera, se corre el riesgo de estar contando dos veces la misma evidencia (una vez para PM1 y otra para PP3), razón por la cual muchas veces no se deja que un subconjunto de etiquetas sobrepase un determinado nivel de evidencia al sumarse. La etiqueta PM1 me habla sobre un rol funcional del residuo de interés, y esto mecanísticamente podríamos ubicarlo en responsabilidad en la actividad catalítica, estabilización del sustrato, interacciones proteína-proteína, modificaciones post-traduccionales importantes a la función, o incluso estabilización energética del core proteico. Cuanto más relevante sea el residuo a la función, más conservado se encontrará, y más probable que capture ese efecto por medio de predictores. Lo realmente interesante surge cuando consideramos elementos funcionalmente importantes no tan claros, donde la pertenencia a un dominio, o la conservación no terminan de reflejar el impacto de su alteración, como ocurre en los motivos proteicos, que abordaremos posteriormente.

Lo que sí podríamos considerar perfectamente, es cuál es la sensibilidad que esa región (en este caso el dominio) pudiera tener a la variación missense. En ese sentido, las guías actuales recomiendan utilizar el Z-score para variantes missense de gnomAD a la hora de utilizar la etiqueta PP2 (en caso del mismo superar el valor de 3,09). Esta métrica está confeccionada en base a un algoritmo que genera un modelo de cantidad de variantes missense esperadas en la secuencia del gen según la tasa mutacional promedio del proteoma, y comparar la misma contra las variantes missense observadas poblacionalmente. La misma métrica advierte sobre las limitaciones de su uso en proteínas de gran tamaño. Pero lo que resulta más sorprendente aún, es que no se consideren las sensibilidades diferenciales que los distintos dominios pueden poseer a la variación missense, en su calidad de unidad estructural, funcional y evolutiva independiente. Particularmente en CHD7, este dato puede vislumbrarse al inspeccionar visualmente herramientas de gnomAD como el Regional Missense Constraint, donde se aprecia que la restricción a variantes missense se concentra en las regiones correspondientes a los dominios. Sin embargo, muchos analistas se apresuran a utilizar la etiqueta pues los mismos softwares automatizados de interpretación (VarSome, FranKlin) lo hacen. Cabe aclarar que esto no constituye una crítica a los mismos, que son herramientas muy provechosas para abordar el estudio de variantes, recopilando información y acercando sugerencias, que luego los profesionales deben ocuparse de aceptar o rechazar. En el caso de la variante que se trató en la presente tesis, un exhaustivo análisis estructural con una buena dosis de evidencias genómicas permite tranquilamente utilizar esta evidencia, que no lo olvidemos, no deja de ser basada en un parámetro estadístico, donde se observa la relación entre variantes esperadas y observadas. Por eso mismo, entendemos que su nivel de evidencia sea Supporting, dado que el significado biológico que atribuímos es más limitado.

El quinto caso clínico desarrollado corresponde a un paciente que presentó una combinación de variantes en heterocigosis en la dupla ligando-receptor conformada por PROK2-PROKR2. Estos constituyen la base del sistema de las prokineticinas, tradicionalmente asociado al proceso biológico de quimiotaxis, tan necesario para que la

neurona GnRH pueda alcanzar su destinación en el hipotálamo. Hace ya mucho tiempo se vienen observando casos/informes donde la presencia de variantes heterocigotas aisladas en estos genes, patogénicas por ACMG, son consideradas diagnósticas en su conjunto y dichos casos clasificados como resueltos/positivos. Sin embargo, muchas de esas variantes son encontradas en la misma cigosidad en una cantidad considerable de individuos sanos, ante lo cual la respuesta inevitable que no tarda en llegar es atribuirlo a la penetrancia incompleta de las mismas. No refutamos esta aseveración, pero nos interesa dejar bien planteado el hecho de que la penetrancia incompleta no habla de una característica inherente de la variante, sino de cómo la misma es condicionada por otros factores genéticos (o epigenéticos) y/o ambientales. En este sentido, resulta muy interesante cómo los modelos oligogénicos, descritos en la presente tesis, vienen a llenar parte de este vacío de conocimiento.

En el caso de nuestro paciente, mostramos cómo un algoritmo (VarCoPP) logró rankear a nuestra combinación por encima de las demás, y la lógica que presentó por detrás, ponderando en su justo valor la contribución hecha tanto por el impacto molecular de las variantes individuales como por el peso de la evidencia atribuida a la relación epistática presente entre los productos proteicos de ambos genes (ligando-receptor). Algunos grupos han querido generar publicaciones justificando causalidad en función de los veredictos dados por estas herramientas, o la cantidad de combinaciones asignadas con scores compatibles con generación de patología. Nada más alejado de la filosofía que ACMG/AMP viene predicando hace años. Establecer patogenicidad de la combinación por el veredicto de estas herramientas es el análogo monogénico de decir que una variante missense es patogénica, solamente porque obtuvo una predicción bioinformática deletérea (es decir, cumplir con la etiqueta PP3).

Al día de hoy no existe una guía o recomendación universalmente aceptada para el reporte de combinaciones oligogénicas en pacientes, pero consideramos de gran valor el aporte hecho por el Laboratorio de Inteligencia Artificial de la Universidad Libre de Bruselas. En sus esfuerzos por delinear las definiciones necesarias para abordar el estudio de casos oligogénicos, incluyendo el desarrollo de VarCoPP, construyeron una base de datos, OLIDA, donde poder realizar la carga de combinaciones que cuenten con un alto grado de curación. Además, una vez cargadas, se somete a las mismas a un protocolo riguroso para evaluar la confianza que podemos tener en la misma de ser causal, traducido en un score que toma valores enteros entre 0 y 3.

La evidencia considerada abarca:

1) información genética familiar (FAMmanual) centrada en la segregación de las variantes consideradas,

2) información genética estadística (STATmeta) destinada a evaluar la posibilidad de que la combinación haya sido detectada en el paciente por mera casualidad,

3) información funcional de la combinación de variantes (VARmeta) que considera toda la información experimental o predicha sobre el efecto conjunto de las variantes, y

4) información funcional de la combinación de genes (GENEmeta), que hace lo propio pero a nivel de genes, con el fin de evidenciar efectos sinérgicos por alteraciones en los mismos.

Nuestra combinación, por ejemplo, alcanzaría scores de FAMmanual = 2, STATmeta = 1, VARmeta = 1 y GENEmeta = 2. Si seguimos los árboles de decisión que el protocolo de

OLIDA nos brinda, vemos como ambos scores funcionales se combinan en un FUNmeta = 2, que luego se combinan para generar un FINALmeta score de confianza de 2. La interpretación que extraemos de OLIDA del mismo es que “la combinación estudiada presenta una buena segregación genética y sinergia funcional entre los genes y variantes involucrados, mostrando un efecto en el fenotipo estudiado, pero donde el mecanismo descrito no es aún claro o lo suficientemente fuerte para dar cuenta de oligogenicidad”. Resulta muy interesante observar cómo esta definición refleja la situación evidenciada por nuestro caso, puesto que contamos con información sumamente valiosa sobre el efecto molecular individual que ambas variantes pueden tener, su frecuencia poblacional es consistente con una variante patogénica, una segregación que acompaña un modelo de herencia oligogénico acorde (los padres heterocigotos para las variantes individuales no presentaron fenotipo mientras que el paciente portador de ambas sí lo desarrolló), y existen pruebas de la sinergia en la alteración simultánea de ambos genes.

De hecho, resulta muy valioso observar que la información referida a la epistasis entre genes estudiada en otros casos y publicaciones, es extrapolada a nuestro caso de estudio con un alto grado de confianza, y esto es justamente consecuencia de todo este celoso proceso de curación. Evidentemente, el reporte de casos siguiendo patrones oligogénicos no solo aporta información sobre las variantes y los modelos de herencia asociados sino que también profundiza en conocimientos de las combinaciones génicas que luego pueden ser extrapolados a otros casos con variantes en los mismos. En última instancia, lo que se genera es un círculo virtuoso donde el mayor conocimiento sobre la etiopatogenia oligogénica redunda en mayor número de casos detectados, cuyo reporte a su vez enriquece cada vez más las bases de datos para ir a analizar casos nuevos. Preservar el estado de curación de las futuras nuevas entradas será clave para impulsar adecuadamente un creciente análisis de estos casos utilizando las mencionadas herramientas. De todas maneras, no podemos negar que lo que aportará la consolidación final del segmento será la realización de experimentos adecuados para evaluar las sinergias funcionales entre genes y variantes.

MotSASi

Por último, en nuestro tercer y último capítulo de resultados, abordamos el camino recorrido en lo referente al estudio de motivos lineales cortos en proteínas humanas. Puede que luego de leer los dos primeros capítulos, el lector piense “Que extraño, luego de dos capítulos iniciales dedicados a estudiar pacientes de una patología, me encuentro con un gran capítulo dedicado a desarrollar un paquete de software para estudiar un grupo particular de variantes”. Es perfectamente válido. En el primer capítulo, observamos cómo pudimos generar información para optimizar la toma de decisiones y el estudio de futuros pacientes que cursan con el fenotipo. En el segundo capítulo los casos fueron pocos pero desarrollados a fondo, como cada paciente que estudiamos en nuestro hospital. Y en este tercer capítulo, nos abrimos al espectro general de patologías mendelianas y nos dedicamos a profundizar en dirección de las variantes missense que se localizan en estas unidades funcionales denominadas motivos. Podemos observar cómo se cubre un abanico de situaciones donde el número de pacientes y su fenotipo presentado son inversamente proporcionales a la especificidad de los tipos de cambios abordados.

El hecho de haber abordado el estudio de elementos funcionales en proteínas es una consecuencia directa de la tradición y expertise del laboratorio comandado por el director de esta tesis. El hecho de haber elegido los motivos lineales cortos como primer

objeto de estudio es prácticamente una casualidad. Uno de nuestro compañeros descubre una variante missense de significado incierto en el contexto de un caso de genómica clínica, casi por casualidad se encuentra con que la misma se localiza en un motivo, procede a realizar un experimento en su laboratorio y corrobora que la alteración de dicho motivo impide a la proteína alcanzar su destinación final en la membrana plasmática. Este fue el punto de partida que, pandemia mediante, dio lugar a meses y meses de trabajo para generar la primera implementación de MotSASi. Sin embargo, algo que debemos decir, es que en la comunión de genómica clínica y análisis estructural que encarna nuestro grupo, probablemente hubieran podido figurar como prioridades otros objetos de estudio: sitios activos, sitios de unión a sustratos, topologías de interacción proteína-proteína entre dominios, etc. Pues bien, los motivos constituyeron nuestro comienzo, y serán los otros nuestros próximos objetivos en la ambición de brindar a la comunidad científica una herramienta que le permita echar claridad sobre un criterio de evidencia que nos involucra particularmente: PM1, cuáles son los residuos funcionalmente importantes en una proteína, y cuál es el nivel de evidencia que podemos atribuirle a los mismos. Como ya dijimos anteriormente, en los últimos años los analistas han recurrido a todo tipo de estrategias, en general involucrando las interfaces de usuario que los diferentes browsers les ofrecen, a la hora de asignar esta evidencia. De esta manera, localización en dominios, alto grado de conservación, representan algunos de los enfoques más recurrentes. Mención aparte merecen los renombrados hot-spots. Muchas veces, a la gente le resulta muchísimo más práctico comprender que un fragmento de una secuencia es funcionalmente importante. Sin embargo, no podemos dejar de observar que la representación de una proteína en una secuencia es un modelo unidimensional, y que la biología, la termodinámica y la cristalografía se han encargado por mucho tiempo de estudiar cómo se pliegan las proteínas. Producto de estos estudios es que llegamos hoy a contar con modelos tridimensionales que nos muestran como residuos, que quizás se encuentran considerablemente separados en la secuencia del polipéptido, se encuentran en el espacio una vez dada la reacción de plegado. Las estructuras cristalográficas depositadas en PDB, como así también las predicciones generadas por AF2, constituyen entonces herramientas fundamentales para entender por ejemplo qué residuos son los que conforman el bolsillo de un sitio activo, qué residuos son los que estabilizan a diversos sustratos que deben aproximarse para que se de una reacción química entre ellos, qué residuos son los que participan activamente de las interacciones entre dos dominios de dos proteínas diferentes. Por lo tanto, en lo referido a estos elementos funcionales y hot-spots, si bien pueden resultar útiles para identificar regiones con mayor sensibilidad a la variación, y pueden aproximar bastante bien sitios funcionalmente importantes, la especificidad de los mismos no resulta óptima pues sabemos que siempre pueden encontrarse variantes benignas en los fragmentos entre residuos claves. Sin embargo, cuando analizamos motivos, quizás sea justamente en estos que encontramos los verdaderos hot-spots, pues estos efectivamente constituyen elementos de secuencia ininterrumpida, al estar definidos por pequeñas expresiones regulares. Claro, el problema es lograr identificarlos siendo tan cortos. Por todos estos motivos, nuestro grupo considera que el aporte que puede realizar a solucionar este problema es máximo.

A diferencia de otros elementos funcionales, los motivos tienen una serie de dificultades extra harto difíciles de superar: en general no se encuentran en dominios, sino que están en regiones altamente desordenadas de secuencia, flexibles, lo cual les permite interaccionar con su dominio de unión. Están definidos por expresiones regulares, que a veces pueden ser más o menos degeneradas. Por lo tanto, el desafío era doble, porque a

diferencia de sitios activos o sitios de unión a sustrato, que en general se localizan en dominios (que al día de hoy es extremadamente factible que podamos identificar por medio de Pfam junto a su función), con los motivos debemos realizar un esfuerzo superior. Primero, para saber dónde están. Segundo, encontrar argumentos suficientes para poder decir que sean motivos funcionales y que no estén simplemente coincidiendo con la expresión regular por azar. Y en tercer lugar, residuo por residuo, analizar en virtud de las informaciones obtenidas de bases de datos de variantes clínicas, de frecuencia, y la aplicación de termodinámica sobre estructuras cristalográficas, un veredicto para cada uno de los cambios posibles. Como podemos ver, los primeros pasos constituyen los requisitos indispensables para asignar PM1, en este caso, este es un motivo funcional con un alto grado de confianza. A continuación, procede la segunda parte del análisis en la cual, en virtud de los datos precedentes, emitimos un veredicto sobre el carácter deletéreo de la variante en lo referido a la unión motivo-dominio. Si nos detenemos un segundo a pensar, este veredicto es absolutamente análogo a aquellos que los tradicionales predictores y metapredictores bioinformáticos emiten, como REVEL en el caso de las variantes missense. Por lo que para una implementación en clínica, de corroborarse que la variante missense en estudio se localiza en un motivo funcional de alta confianza, proponemos etiquetar PM1, y a continuación, de brindar MotSASi un veredicto sobre el carácter deletéreo de la variante, proceder a etiquetar PP3 o BP4 en función de si el mismo orienta a la patogenicidad o a la benignidad. Cuando se analiza la implementación de etiquetas referidas a efectos moleculares deletéreos de las variantes, como PM1, PP3, PM4 (deleción o inserción in-frame en el polipéptido) y PP2 (región con sensibilidad elevada a la variación missense), dado que existe un grado alto de superposición entre ellas, en general se adopta como medida precautoria no permitir que la suma de las mismas exceda un nivel de evidencia Strong (correspondiente a dos Moderate o cuatro Supporting). Pues bien, tomando esto en cuenta, y considerando las matrices con puntuación de confianza patogénica/benigna que entregan, proponemos el siguiente sistema. Aquellos veredictos patogénicos que superen un nivel de corte de puntuación de 2 puntos, asignar un nivel de evidencia moderado a la etiqueta de predictores (PP3_Moderate). Para puntuaciones entre 0 y 2 asignar simplemente un nivel Supporting (PP3). En caso de ser benigno el veredicto, en caso de encontrarse entre 0 y -2, asignar nivel supporting de evidencia (BP4) y en caso de ser menor a los -2 puntos, asignar un nivel moderado (BP4_Moderate). De esta manera, lo que hacemos es respetar el máximo nivel de evidencia que se le asigna a estas etiquetas, y al mismo tiempo, en caso de estar muy seguros de la naturaleza benigna del veredicto de MotSASi, neutralizamos el hecho de encontrarse la variante en un sitio funcional (PM1) con la predicción benigna (BP4_Moderate).

Una aclaración que debemos realizar con respecto a MotSASi, es que los resultados que presentamos en esta tesis no corresponden a la totalidad de los motivos depositados en la base de datos del ELM. No debemos olvidar que MotSASi fue diseñado como un algoritmo donde todo lo que debe dársele como *input* es el nombre del motivo y su expresión regular, y adicionalmente, en caso de haber generado predicciones de AF2 para completar el análisis, los archivos correspondientes a las mismas. Sin embargo, no debemos dejar de resaltar que existen requerimientos y filtros impuestos que limitan la cantidad de motivos estudiables automáticamente por MotSASi. Por ejemplo, se solicita la presencia de al menos una instancia del motivo ya debidamente documentada en el ELM en especie humana. Es a partir de las instancias de motivos ya registradas que se establecen los filtros correspondientes para permitirnos identificar nuevas posibles instancias de motivo de alta confianza para seleccionar candidatos que progresivamente iremos estudiando.

Además, las variantes provenientes de ClinVar y gnomAD presentes en los mismos son las que enriquecen el análisis, haciendo que no se dependa únicamente de las predicciones termodinámicas de FoldX. En caso de que una persona contara con una variante presente en un comprobado o potencial motivo nuevo, poco estudiado, y quisiera utilizar MotSASi, podemos claramente forzar el análisis para evaluar la factibilidad del motivo de ser funcional, incluso de generar un veredicto (que corresponderá a FoldX probablemente), pero no es la filosofía propia de MotSASi que busca generar predicciones por consenso, tratando de integrar la mayor cantidad de información e instancias de motivo posibles.

Por otro lado, generar las predicciones de AF2 dista mucho de ser una simple tarea donde remito al software las secuencias correspondientes al motivo y al dominio para que genere las predicciones. Al contrario, múltiples son los problemas que encontramos en este camino, y que en la mayoría de los casos requieren que un profesional con criterio tome cartas en el asunto. Si bien nosotros tomamos la información del ELM, en principio curada en alto grado, y a partir de ella generamos un archivo fasta para pedir a AF2 la predicción estructural, muchas veces ocurre que la base de datos se encuentra incompleta, y si bien se enuncia que existe una instancia de motivo en una posición determinada de una proteína específica, muchas veces no fue cargado a la base el par de interacción motivo-dominio. Esto complica la generación del fasta, con lo cual debemos ir a la bibliografía específica para suplir esta falta de prolijidad en la base, y elaborar artesanalmente los archivos para enviar a AF2. Esto también ocurre cuando la subida de la información al ELM se hizo con un transcripto de referencia proteico muy antiguo, y no coinciden las numeraciones, que nosotros siempre con mucho celo comprobamos. Problemas análogos ocurren cuando por ejemplo se definen los dominios con algoritmos alternativos, no-Pfam. Pero estas dificultades técnicas no son ni por asomo las más interesantes, que resultan ser de naturaleza biológica. Algunas situaciones típicas con las que nos hemos encontrado incluyen por ejemplo a los motivos que cuentan con algún residuo modificado, en general fosforilaciones. AF2 al día de hoy no permite introducir residuos fosforilados en su *input*, lo cual nos impide de contar con predicciones que incluyan modificaciones post-traduccionales. Este obstáculo quisimos sortearlo durante las etapas de validación introduciendo en dicha posición un residuo fosfomimético (por ejemplo, en lugar de colocar la fosfotirosina correspondiente, introducimos un glutamato que replica al menos el carácter negativo de la carga), pero los resultados no fueron para nada buenos, con lo cual, al menos provisoriamente, no pudimos analizar estas clases de motivos, que representan un conjunto de número interesante. Otro problema que encontramos fue el de aquellos motivos que poseen una estequiometría de unión diferente de la tradicional 1:1, es decir, el motivo reconoce un sitio de binding en una unidad del dominio. Algunas veces, el motivo interacciona con un dominio que se encuentra repetido dos o muchísimas veces en su proteína, y la unión se da en términos de que el primero se aloja en un surco conformado por dos unidades del segundo que se encuentran muy próximas. La información de estequiometría de interacción se sumó en las últimas versiones del ELM, con lo que uno puede estar más prevenido de estos sucesos. En el contexto de la validación quisimos replicar estas interacciones introduciendo una secuencia que contuviera dos repeticiones del dominio y los resultados fueron muy buenos, logrando AF2 emular perfectamente la interacción. Por lo tanto, si se cuenta con información sobre la posibilidad de este tipo de circunstancia, AF2 aparece como una herramienta atractiva.

Contar con la disponibilidad de predictores como MotSASi genera una contribución a la interpretación de variantes generando enfermedad por medio de patrones monogénicos como también de combinaciones de las mismas siguiendo un patrón oligogénico. A este

respecto, debemos decir que en MotSASi, en general, las variantes con las cuales se entrenó al modelo fueron analizadas con un enfoque monogénico, por lo cual uno debería plantearse la posibilidad de modificar determinados valores de corte (relajando a los mismos) a fines de ganar sensibilidad en la detección de variantes compatibles con patrones oligogénicos, donde en la base de datos OLIDA se documentan variantes con frecuencias de hasta 3,5% y efectos moleculares más moderados. Esto va de la mano de la apertura del universo de mecanismo de patogenicidad que plantea MotSASi con respecto a predictores bioinformáticos previos para variantes missense. La mayoría de estos últimos se basaban en una serie de variables como la localización en dominio, conservación, estructura secundaria, impacto termodinámico del cambio, etc, que quizás no terminaban de captar 100% el problema de secuencias como los motivos, más flexibles, que evolucionaron muy rápido recientemente y cumplen su rol funcional por medio de unirse a un dominio.

Dado que nuestro enfoque se centra fuertemente en el componente estructural del problema de la unión motivo-dominio, decidimos compararnos contra un algoritmo como lo es AlphaMissense, que básicamente basa su predicción en las predicciones de estructura tridimensional generadas a partir de la secuencia de entrada, acompañado de un fuerte componente de conservación gracias a los MSA que genera, y refinado con información de frecuencia alélica poblacional en humanos y primates. El paper correspondiente a esta herramienta indica que la misma posee una menor exactitud en regiones desordenadas tal y como las define AF2 en base al score de confianza asociado a dichos residuos. Probablemente parte de esta menor performance tenga que ver con el hecho de que si bien la mayor parte del proteoma desordenado se encuentra enriquecido en variación benigna, secuencias funcionales como los motivos se encuentran en el mismo, y ciertas variantes pueden tener un impacto deletéreo en la unión motivo-dominio que ameritan su clasificación como patogénicas. La comparación realizada entre MotSASi y AlphaMissense fue llevada adelante de tal manera de no favorecer totalmente al primero con variantes utilizadas en su entrenamiento, siendo utilizado un set de evaluación independiente. Nuestro algoritmo demostró desempeñarse de mejor manera a la inteligencia artificial desarrollada por Google, principalmente en el aspecto de no dar como patogénicas a variantes que no son deletéreas para la unión motivo-dominio (es decir, evitar falsos positivos).

Referencias

1. FADEPOF. [cited 27 Dec 2024]. Available: https://fadepof.org.ar/epof.php
2. Lee K. The World Health Organization (WHO). Routledge; 2008.
3. Prakash V, Moore M, Yáñez-Muñoz RJ. Current Progress in Therapeutic Gene Editing for Monogenic Diseases. Mol Ther. 2016;24: 465–474.
4. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43: D789–98.
5. https://www.primeraedicion.com.ar/wp-content/uploads/2022/05/ENSERio-LATAM_Capitulo-Argentina_FADEPOF_final.pdf
6. Genetics, Disease Prevention and Treatment FAQ. In: Genome.gov [Internet]. [cited 28 Dec 2024]. Available: https://www.genome.gov/FAQ/Genetics-Disease-Prevention-and-Treatment

7. Manolio TA. Incorporating Whole-Genome Sequencing Into Primary Care: Falling Barriers and Next Steps. Annals of Internal Medicine. 2017. p. 204. doi:10.7326/m17-1518
8. Meng Q, Yu J. Next Generation DNA Sequencing Technologies and Applications. Next Generation Sequencing and Whole Genome Selection in Aquaculture. 2010. pp. 35–56. doi:10.1002/9780470958964.ch3
9. Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. Genome Medicine. 2015. doi:10.1186/s13073-015-0199-2
10. Diagnostic Clinical Genome and Exome Sequencing. New England Journal of Medicine. 2014. pp. 1169–1170. doi:10.1056/nejmc1408914
11. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11: 31–46.
12. Frank M, Prenzler A, Eils R, Graf von der Schulenburg J-M. Genome sequencing: a systematic review of health economic evidence. Health Econ Rev. 2013;3: 29.
13. Willyard C. New human gene tally reignites debate. Nature. 2018;558: 354–355.
14. van Nimwegen KJM, van Soest RA, Veltman JA, Nelen MR, van der Wilt GJ, Vissers LELM, et al. Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing.
15. Ku C-S, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Ann Neurol. 2012;71: 5–14.
16. Stranneheim H, Wedell A. Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. Journal of Internal Medicine. 2016. pp. 3–15. doi:10.1111/joim.12399
17. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013;369: 1502–1511.
18. Hoppman-Chaney N, Peterson LM, Klee EW, Middha S, Courteau LK, Ferber MJ. Evaluation of oligonucleotide sequence capture arrays and comparison of next-generation sequencing platforms for use in molecular diagnostics. Clin Chem. 2010;56: 1297–1306.
19. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55: 641–658.
20. Precone V, Del Monaco V, Esposito MV, De Palma FDE, Ruocco A, Salvatore F, et al. Cracking the Code of Human Diseases Using Next-Generation Sequencing: Applications, Challenges, and Perspectives. Biomed Res Int. 2015;2015: 161648.
21. Xiao T, Zhou W. The third generation sequencing: the advanced approach to genetic diseases. Transl Pediatr. 2020;9: 163–173.
22. Zhang R. Oxford Nanopore sequencing and library construction v1. protocols.io. doi:10.17504/protocols.io.btcwnixe
23. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of Next-Generation Sequencing Technologies. Analytical Chemistry. 2011. pp. 4327–4341. doi:10.1021/ac2010857
24. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30: 418–426.
25. Wong L-JC. Next Generation Sequencing Based Clinical Molecular Diagnosis of Human Genetic Disorders. Springer; 2017.
26. Meur G, Simon A, Harun N, Virally M, Dechaume A, Bonnefond A, et al. Insulin gene mutations
resulting in early-onset diabetes: marked differences in clinical presentation, metabolic status, and

pathogenic effect through endoplasmic reticulum retention. Diabetes. 2010;59: 653–661.
27. Yubero D, Artuch R. NGS for Metabolic Disease Diagnosis. EJIFCC. 2018;29: 227–229.
28. Guan Y-F, Li G-R, Wang R-J, Yi Y-T, Yang L, Jiang D, et al. Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. Chin J Cancer. 2012;31: 463–470.
29. Gagan J, Van Allen EM. Next-generation sequencing to guide cancer therapy. Genome Med. 2015;7: 80.
30. Roychowdhury S, Van Allen EM. Precision Cancer Medicine: Challenges and Opportunities. Springer Nature; 2020.
31. Blue GM, Winlaw DS. Next Generation Sequencing in Congenital Heart Disease: Gene Discovery and Clinical Application. Journal of Next Generation Sequencing & Applications. 2015. doi:10.4172/2469-9853.1000113
32. Odibo AO, Krantz DA. Prenatal Screening and Diagnosis, An Issue of the Clinics in Laboratory Medicine, E-Book. Elsevier Health Sciences; 2016.
33. Muzzey D. Understanding the Basics of NGS in the Context of NIPT. Noninvasive Prenatal Testing (NIPT). 2018. pp. 7–24. doi:10.1016/b978-0-12-814189-2.00002-5
34. Pös O, Budis J, Kubiritova Z, Kucharik M, Duris F, Radvanszky J, et al. Identification of Structural Variation from NGS-Based Non-Invasive Prenatal Testing. International Journal of Molecular Sciences. 2019. p. 4403. doi:10.3390/ijms20184403
35. Ma L, Jakobiec FA, Dryja TP. A Review of Next-Generation Sequencing (NGS): Applications to the Diagnosis of Ocular Infectious Diseases. Seminars in Ophthalmology. 2019. pp. 223–231. doi:10.1080/08820538.2019.1620800
36. Ampofo K, Pavia A, Blaschke AJ, Schlaberg R. Detection of Respiratory Pathogens in Parapneumonic Effusions by Hypothesis-free, Next-Generation Sequencing (NGS). Open Forum Infectious Diseases. 2017. pp. S17–S17. doi:10.1093/ofid/ofx162.044
37. Harb WJ. Gastrointestinal Polyposis Syndromes. Inherited Cancer Syndromes. 2011. pp. 105–125. doi:10.1007/978-1-4419-6821-0_5
38. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16(th) IHIW : review of HLA typing by NGS. Int J Immunogenet. 2013;40: 72–76.
39. Mack SJ, Milius RP, Gifford BD, Sauter J, Hofmann J, Osoegawa K, et al. Minimum information for reporting next generation sequence genotyping (MIRING): Guidelines for reporting HLA and KIR genotyping via next generation sequencing. Hum Immunol. 2015;76: 954–962.
40. Zhang X. Exome sequencing greatly expedites the progressive research of Mendelian diseases. Front Med. 2014;8: 42–57.
41. Sawyer SL, Hartley T, Dyment DA, Beaulieu CL, Schwartzentruber J, Smith A, et al. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clin Genet. 2016;89: 275–284.
42. Grody WW, Nakamura RM, Kiechle FL, Strom C. Molecular Diagnostics: Techniques and Applications for the Clinical Laboratory. Academic Press; 2009.
43. Genomic and Personalized Medicine. Academic Press; 2012.
44. Kanzi AM, San JE, Chimukangara B, Wilkinson E, Fish M, Ramsuran V, et al. Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. Front Genet. 2020;11:544162.
45. McKenna, Aaron & Hanna, Matthew & Banks, Eric & Sivachenko, Andrey & Cibulskis, Kristian & Kernytsky, Andrew & Garimella, Kiran & Altshuler, David & Gabriel, Stacey & Daly, Mark & DePristo, Mark. (2010). The Genome Analysis Toolkit: A MapReduce

framework for analyzing next-generation DNA sequencing data. Genome research. 20. 1297-303. 10.1101/gr.107524.110.
46. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform. 2012;3: 40.
47. Sergey Nurk et al. ,The complete sequence of a human genome. Science 376,44-53(2022). DOI:10.1126/science.abj6987
48. A global reference for human genetic variation. Nature. 2015;526: 68–74.
49. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337: 64–69.
50. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536: 285–291.
51. The Genome Aggregation Database (gnomAD). [cited 20 Feb 2022]. Available: https://www.nature.com/immersive/d42859-020-00002-x/index.html
52. Karczewski, Konrad & Francioli, Laurent & Grace, Tiao & Cummings, Beryl & Alföldi, Jessica & Wang, Qingbo & Collins, Ryan & Laricchia, Kristen & Ganna, Andrea & Birnbaum, Daniel & Gauthier, Laura & Brand, Harrison & Solomonson, Matthew & Watts, Nicholas & Rhodes, Daniel & Singer-Berk, Moriel & England, Eleina & Seaby, Eleanor & Kosmicki, Jack & Xavier, Ramnik. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 581. 434-443. 10.1038/s41586-020-2308-7.
53. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45: D840–D845.
54. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.
55. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry. 2010;19: 227–229.
56. Ghosh, Rajarshi & Harrison, Steven & Rehm, Heidi & Plon, Sharon & Biesecker, Leslie. (2018). Updated recommendation for the benign stand-alone ACMG/AMP criterion. Human Mutation. 39. 1525-1530. 10.1002/humu.23642.
57. Tripathi, Siddhant & Sharma, Yashika & Kumar, Dileep. (2024). Unraveling APOE4's Role in Alzheimer's Disease: Pathologies and Therapeutic Strategies. Current protein & peptide science. 10.2174/0113892037326839241014054430.
58. Boehm CD, Kazazian HH Jr. The molecular basis of genetic disease. Curr Opin Biotechnol. 1990;1: 180–187.
59. Strasser BJ, Fantini B. Molecular diseases and diseased molecules: ontological and epistemological dimensions. Hist Philos Life Sci. 1998;20: 189–214.
60. Prokopcova, Aneta & Baloun, Jiri & Švec, Xiao & Šenolt, Ladislav. (2022). Non‐coding RNAs in diseases with a focus on osteoarthritis. WIREs RNA. 14. 10.1002/wrna.1756.
61. Bampi, Giovana & Ramalho, Anabela & Santos, Leonardo & Wagner, Johannes & Dupont, Lieven & Cuppens, Harry & de Boeck, Christiane & Ignatova, Zoya. (2020). The Effect of Synonymous Single-Nucleotide Polymorphisms on an Atypical Cystic Fibrosis Clinical Presentation. Life. 11. 14. 10.3390/life11010014.
62. Kurosaki, Tatsuaki & Popp, Maximilian & Maquat, Lynne. (2019). Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nature Reviews Molecular Cell Biology. 20. 10.1038/s41580-019-0126-2.

63. Richards, Sue & Bale, Sherri & Bick, David & Das, Soma & Gastier-Foster, Julie & Grody, Wayne & Hegde, Madhuri & Lyon, Elaine & Spector, Elaine & Voelkerding, Karl & Rehm, Heidi. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in medicine : official journal of the American College of Medical Genetics. 17. 10.1038/gim.2015.30.
64. Li, Quan & Wang, Kai. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. The American Journal of Human Genetics. 100. 10.1016/j.ajhg.2017.01.004.
65. Rehm, Heidi & Berg, Jonathan & Brooks, Lisa & Bustamante, Carlos & Evans, James & Landrum, Melissa & Ledbetter, David & Maglott, Donna & Martin, Christa & Nussbaum, Robert & Plon, Sharon & Ramos, Erin & Sherry, Stephen & Watson, Michael. (2015). ClinGen — The Clinical Genome Resource. The New England journal of medicine. 372. 10.1056/NEJMsr1406261.
66. Strande, Natasha & Riggs, Erin & Buchanan, Adam & Ceyhan-Birsoy, Ozge & DiStefano, Marina & Dwight, Selina & Goldstein, Jenny & Ghosh, Rajarshi & Seifert, Bryce & Sneddon, Tam & Wright, Matt & Milko, Laura & Cherry, J. & Giovanni, Monica & Murray, Michael & o'daniel, Julianne & Ramos, Erin & Santani, Avni & Scott, Alan & Berg, Jonathan. (2017). Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. The American Journal of Human Genetics. 100. 10.1016/j.ajhg.2017.04.015.
67. Whiffin, Nicola & Minikel, Eric & Walsh, Roddy & O'Donnell-Luria, Anne & Karczewski, Konrad & Ing, Alexander & Barton, Paul & Funke, Birgit & Cook, Stuart & MacArthur, Daniel & Ware, James. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. Genetics in medicine : official journal of the American College of Medical Genetics. 19. 10.1038/gim.2017.26.
68. SVI Recommendation for Absence/Rarity (PM2) - Version 1.0 [Internet]. [cited 31 Dec 2024]. Available: https://clinicalgenome.org/site/assets/files/5182/pm2_-_svi_recommendation_-_approved_sept2020.pdf
69. Tayoun, Ahmad & Pesaran, Tina & DiStefano, Marina & Oza, Andrea & Rehm, Heidi & Biesecker, Leslie & Harrison, Steven. (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Human Mutation. 39. 10.1002/humu.23626.
70. Walker, Logan & Hoya, Miguel & Wiggins, George & Lindy, Amanda & Vincent, Lisa & Parsons, Michael & Canson, Daffodil & Bis-Brewer, Dana & Cass, Ashley & Tchourbanov, Alexander & Zimmermann, Heather & Byrne, Alicia & Pesaran, Tina & Karam, Rachid & Harrison, Steven & Spurdle, Amanda. (2023). APPLICATION OF THE ACMG/AMP FRAMEWORK TO CAPTURE EVIDENCE RELEVANT TO PREDICTED AND OBSERVED IMPACT ON SPLICING: RECOMMENDATIONS FROM THE CLINGEN SVI SPLICING SUBGROUP. medRxiv : the preprint server for health sciences. 10.1101/2023.02.24.23286431.
71. Pejaver, Vikas & Byrne, Alicia & Feng, Bing-Jian & Pagel, Kymberleigh & Mooney, Sean & Karchin, Rachel & O'Donnell-Luria, Anne & Harrison, Steven & Tavtigian, Sean & Greenblatt, Marc & Biesecker, Leslie & Radivojac, Predrag & Brenner, Steven & Tayoun, Ahmad & Berg, Jonathan & Cutting, Garry & Ellard, Sian & Kang, Peter & Karbassi, Izabela & Topper, Scott. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. The American Journal of Human Genetics. 109. 10.1016/j.ajhg.2022.10.013.

72. PS2/PM6: Recommendation for de novo PS2 and PM6 ACMG/AMP criteria (Version 1.1) [Internet]. [cited 31 Dec 2024]. Available: https://clinicalgenome.org/site/assets/files/3461/svi_proposal_for_de_novo_criteria_v1_1.pdf
73. SVI Recommendation for in trans Criterion (PM3) - Version 1.0 [Internet]. [cited 31 Dec 2024]. Available: https://clinicalgenome.org/site/assets/files/3717/svi_proposal_for_pm3_criterion_-_version_1.pdf
74. Biesecker, Leslie & Harrison, Steven. (2018). The ACMG/AMP reputable source criteria for the interpretation of sequence variants. GENETICS in MEDICINE. 20. 10.1038/gim.2018.42.
75. Tavtigian, Sean & Greenblatt, Marc & Harrison, Steven & Nussbaum, Robert & Prabhu, Snehit & Boucher, Kenneth & Biesecker, Leslie. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. GENETICS in MEDICINE. 20. 10.1038/gim.2017.210.
76. Tavtigian, Sean & Harrison, Steven & Boucher, Kenneth & Biesecker, Leslie. (2020). Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. Human Mutation. 41. 10.1002/humu.24088.
77. Riggs, Erin & Andersen, Erica & Cherry, Athena & Kantarci, Sibel & Kearney, Hutton & Patel, Ankita & Raca, Gordana & Ritter, Deborah & South, Sarah & Thorland, Erik & Pineda Alvarez, Daniel & Aradhya, Swaroop & Martin, Christa. (2019). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genetics in Medicine. 22. 1-13. 10.1038/s41436-019-0686-8.
78. Lu JT, Campeau PM, Lee BH. Genotype-phenotype correlation. Obstet Gynecol Surv. 2014;69:728–730.
79. Home - OMIM - NCBI. [cited 31 Dec 2024]. Available: https://www.ncbi.nlm.nih.gov/omim
80. Papa R, Doglio M, Lachmann HJ, Ozen S, Frenkel J, Simon A, et al. A web-based collection of genotype-phenotype associations in hereditary recurrent fevers from the Eurofever registry. Orphanet J Rare Dis. 2017;12: 167.
81. Tisdale A, Cutillo CM, Nathan R, Russo P, Laraway B, Haendel M, et al. The IDeaS initiative: pilot study to assess the impact of rare diseases on patients and healthcare systems. Orphanet J Rare Dis. 2021;16: 429.
82. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare diseases are there? Nat Rev Drug Discov. 2020;19: 77–78.
83. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nature Reviews Genetics. 2018. pp. 253–268. doi:10.1038/nrg.2017.116
84. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28: 165–173.
85. Paz JFL. Aproximación a las dimensiones y factores asociados al contexto de enfermedades raras o poco frecuentes. Revista Iberoamericana de Bioética. 2021. pp. 01–13. doi:10.14422/rib.i15.y2021.006
86. Festa, Adalgisa & Umano, Giuseppina & Giudice, Emanuele & Grandone, Anna. (2020). Genetic Evaluation of Patients With Delayed Puberty and Congenital Hypogonadotropic Hypogonadism: Is it Worthy of Consideration?. Frontiers in Endocrinology. 11. 253. 10.3389/fendo.2020.00253.

87. Karges, Beate & Neulen, Joseph & Nicolas, de & Wolfram, Karges. (2012). Genetics of Isolated Hypogonadotropic Hypogonadism: Role of GnRH Receptor and Other Genes. International journal of endocrinology. 2012. 147893. 10.1155/2012/147893.
88. Rohayem, Julia & Alexander, Emma & Heger, Sabine & Nordenström, Anna & Howard, Sasha. (2024). Mini-Puberty, Physiological and Disordered: Consequences, and Potential for Therapeutic Replacement. Endocrine reviews. 45. 10.1210/endrev/bnae003.
89. Cassatella, Daniele & Howard, Sasha & Acierno, James & Xu, Cheng & Papadakis, Georgios E. & Santoni, Federico & Dwyer, Andrew & Santini, Sara & Sykiotis, Gerasimos & Chambion, Caroline & Meylan, Jenny & Marino, Laura & Favre, Lucie & Li, Jiankang & Liu, Xuanzhu & Zhang, Jianguo & Bouloux, Pierre-Marc & De Geyter, Christian & Paepe, Anne & Pitteloud, Nelly. (2018). Congenital Hypogonadotropic Hypogonadism and Constitutional Delay of Growth and Puberty Have Distinct Genetic Architectures. European Journal of Endocrinology. 178. EJE-17. 10.1530/EJE-17-0568.
90. Versbraegen, Nassim & Fouché, Aziz & Nachtegael, Charlotte & Papadimitriou, Sofia & Gazzo, Andrea & Smits, Guillaume & Lenaerts, Tom. (2019). Using game theory and decision decomposition to effectively discern and characterise bi-locus diseases. Artificial Intelligence in Medicine. 99. 10.1016/j.artmed.2019.06.006.
91. Gazzo, Andrea & Danneels, Dorien & Cilia, Elisa & Bonduelle, Maryse & Abramowicz, Marc & Dooren, Sonia & Smits, Guillaume & Lenaerts, Tom. (2015). DIDA: A curated and annotated digenic diseases database. Nucleic Acids Research. 10.1093/nar/gkv1068.
92. Nachtegael, Charlotte & Gravel, Barbara & Dillen, Arnau & Smits, Guillaume & Nowe, Ann & Papadimitriou, Sofia & Lenaerts, Tom. (2022). Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database. Database. 2022. 10.1093/database/baac023.
93. Papadimitriou, Sofia & Gravel, Barbara & Nachtegael, Charlotte & Baere, Elfride & Loeys, Bart & Vikkula, Miikka & Smits, Guillaume & Lenaerts, Tom. (2022). Toward reporting standards for the pathogenicity of variant combinations involved in multilocus/oligogenic diseases. Human Genetics and Genomics Advances. 4. 100165. 10.1016/j.xhgg.2022.100165.
94. Versbraegen, Nassim & Gravel, Barbara & Nachtegael, Charlotte & Renaux, Alexandre & Verkinderen, Emma & Nowe, Ann & Lenaerts, Tom & Papadimitriou, Sofia. (2023). Faster and more accurate pathogenic combination predictions with VarCoPP2.0. BMC Bioinformatics. 24. 10.1186/s12859-023-05291-3.
95. K. Van Roey, B. Uyar, R.J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T.J. Gibson, N.E. Davey, Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation, Chem. Rev. 114 (2014) 6733–6778. https://doi.org/10.1021/cr400585q.
96. P. Tompa, N.E. Davey, T.J. Gibson, M.M. Babu, A Million Peptide Motifs for the Molecular Biologist, Mol. Cell. 55 (2014) 161–169. https://doi.org/10.1016/j.molcel.2014.05.032.
97. Manjeet Kumar, Sushama Michael, Jesús Alvarado-Valverde, András Zeke, Tamas Lazar, Juliana Glavina, Eszter Nagy-Kanta, Juan Mac Donagh, Zsofia E Kalman, Stefano Pascarelli, Nicolas Palopoli, László Dobson, Carmen Florencia Suarez, Kim Van Roey, Izabella Krystkowiak, Juan Esteban Griffin, Anurag Nagpal, Rajesh Bhardwaj, Francesca Diella, Bálint Mészáros, Kellie Dean, Norman E Davey, Rita Pancsa, Lucía B Chemes, Toby J Gibson, ELM—the Eukaryotic Linear Motif resource—2024 update, Nucleic Acids

Research, Volume 52, Issue D1, 5 January 2024, Pages D442–D455, https://doi.org/10.1093/nar/gkad1058.
98. A. Etxebarria, A. Benito-Vicente, L. Palacios, M. Stef, A. Cenarro, F. Civeira, H. Ostolaza, C. Martin, Functional Characterization and Classification of Frequent Low-Density Lipoprotein Receptor Variants, Hum. Mutat. 36 (2015) 129–141. https://doi.org/10.1002/humu.22721.
99. E. Kalay, A.P.M. de Brouwer, R. Caylan, S.B. Nabuurs, B. Wollnik, A. Karaguzel, J.G.A.M. Heister, H. Erdol, F.P.M. Cremers, C.W.R.J. Cremers, H.G. Brunner, H. Kremer, A novel D458V mutation in the SANS PDZ binding motif causes atypical Usher syndrome, J. Mol. Med. 83 (2005) 1025–1032. https://doi.org/10.1007/s00109-005-0719-4.
100. Martín, Mariano & Brunello, Franco & Modenutti, Carlos & Nicola, Juan & Marti, Marcelo. (2022). MotSASi: Functional short linear motifs (SLiMs) prediction based on genomic single nucleotide variants and structural data. Biochimie. 197. 10.1016/j.biochi.2022.02.002.
101. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering, and energy sciences. Nucleic Acids Res. 2021;49(D1):D437–D451. doi:10.1093/nar/gkaa1038.
102. J. Delgado, L.G. Radusky, D. Cianferoni, L. Serrano, FoldX 5.0: working with RNA, small molecules and a new graphical interface, Bioinformatics. 35 (2019) 4168–4169. https://doi.org/10.1093/bioinformatics/btz184.
103. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153
104. Chen, S.et al. A genomic mutational constraint map using variation in 76,156 human genomes Nature, 625, 92–100 (2024). https://doi.org/10.1038/s41586-023-06045-0.
105. Jumper, J. et al. “Highly accurate protein structure prediction with AlphaFold.” Nature, 596, pages 583–589 (2021). DOI: 10.1038/s41586-021-03819-2.
106. Grinspon, Romina & Ropelato, María & Gottlieb, Silvia & Keselman, Ana & Martínez, Alicia & Ballerini, María & Domené, Horacio & Rey, Rodolfo. (2010). Basal Follicle-Stimulating Hormone and Peak Gonadotropin Levels after Gonadotropin-Releasing Hormone Infusion Show High Diagnostic Accuracy in Boys with Suspicion of Hypogonadotropic Hypogonadism. The Journal of clinical endocrinology and metabolism. 95. 2811-8. 10.1210/jc.2009-2732.
107. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.
108. Fowler, Anna. (2022). DECoN: A Detection and Visualization Tool for Exonic Copy Number Variants. 10.1007/978-1-0716-2293-3_6.
109. Cingolani, Pablo & Platts, Adrian & Wang, Le & Coon, Melissa & Nguyen, Tung & Luan, Wang & Lu, Xiangyi & Ruden, Douglas. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 6. 80-92. 10.4161/fly.19695.
110. Ioannidis, Nilah & Rothstein, Joseph & Pejaver, Vikas & Middha, Sumit & McDonnell, Shannon & Baheti, Saurabh & Musolf, Anthony & Li, Qing & Holzinger, Emily & Karyadi, Danielle & Cannon-Albright, Lisa & Teerlink, Craig & Stanford, Janet & Isaacs, William & Cooney, Kathleen & Lange, Ethan & Schleutker, Johanna & Carpten, John & Sieh, Weiva.

(2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. The American Journal of Human Genetics. 99. 877-885. 10.1016/j.ajhg.2016.08.016.
111. Jaganathan, Kishore & Panagiotopoulou, Sofia & McRae, Jeremy & Fazel Darbandi, Siavash & Knowles, David & Li, Yang & Kosmicki, Jack & Arbelaez, Juan & Cui, Wenwu & Schwartz, Grace & Chow, Eric & Kanterakis, Efstathios & Gao, Hong & Kia, Amirali & Batzoglou, Serafim & Sanders, Stephan & Farh, Kyle. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell. 176. 10.1016/j.cell.2018.12.015.
112. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. doi:10.1093/nar/gkq603.
113. Moher, David & Shamseer, Larissa & Clarke, Mike & Ghersi, Davina & Liberati, Alessandro & Petticrew, Mark & Shekelle, Paul & Stewart, Lesley & Altman, Douglas & Booth, Alison & Chan, An & Chang, Stephanie & Clifford, Tammy & Dickersin, Kay & Gøtzsche, Peter & Grimshaw, Jeremy & Groves, Trish & Helfand, Mark & Higgins, Julian & Whitlock, Evelyn. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic reviews.
114. https://franklin.genoox.com - Franklin by Genoox
115. Kopanos, Christos & Tsiolkas, Vasilis & Kouris, Alexandros & Chapple, Charles & Aguilera, Monica & Meyer, Richard & Massouras, Andreas. (2018). VarSome: The Human Genomic Variant Search Engine. Bioinformatics (Oxford, England). 35. 10.1093/bioinformatics/bty897.
116. Holtgrewe, Manuel & Beule, Dieter. (2016). VCFPy: a Python 3 library with good support for both reading and writing VCF. The Journal of Open Source Software. 1. 10.21105/joss.00085.
117. Harrison, Peter & Amode, M & Austine-Orimoloye, Olanrewaju & Azov, Andrey G & Barba, Matthieu & Barnes, If & Becker, Arne & Bennett, Ruth & Berry, Andrew & Bhai, Jyothish & Bhurji, Simarpreet & Boddu, Sanjay & Branco Lins, Paulo & Brooks, Lucy & Ramaraju, Shashank Budhanuru & Campbell, Lahcen & Martinez, Manuel & Charkhchi, Mehrnaz & Chougule, Kapeel & Yates, Andrew D. (2023). Ensembl 2024. Nucleic Acids Research. 52. 10.1093/nar/gkad1049.
118. Szklarczyk, Damian & Kirsch, Rebecca & Koutrouli, Mikaela & Nastou, Katerina & Nlp, Farrokh & Hachilif, Radja & Gable, Annika & Fang, Tao & Doncheva, Nadezhda & Pyysalo, Sampo & Bork, Peer & Jensen, Lars J & von Mering, Christian. (2022). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research. 51. 10.1093/nar/gkac1000.
119. The UniProt Consortium , UniProt: the Universal Protein Knowledgebase in 2025, Nucleic Acids Research, 2024;, gkae1010, https://doi.org/10.1093/nar/gkae1010
120. Rius, A., Aguirre, N., Erra, L., Brunello, F. G., Biagioli, G., Zaiat, J., & Marti, M. A. (2025). Study of the impact of ClinGen Revisions on ACMG/AMP variant semi-automatic classification for Rare Diseases diagnosis. Clinica Chimica Acta, 566, 120065. https://doi.org/10.1016/j.cca.2024.120065.
121. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013;30(4):772–780. doi:10.1093/molbev/mst010.

122. J.A. Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics. 23 (2007) 1875–1882. https://doi.org/10.1093/bioinformatics/btm270.
123. Pollastri, Gianluca & Przybylski, Dariusz & Rost, Burkhard & Baldi, Pierre. (2002). Improving the Prediction of Protein Secondary Strucure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. Proteins. 47. 228-35. 10.1002/prot.10082.
124. Mitternacht S (2016) FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Res. 5: 189 Available at: http://dx.doi.org/10.12688/f1000research.7931.1.
125. The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. Genetics. 2023 May 4;224(1):iyad031. DOI: 10.1093/genetics/iyad031.
126. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics, J. Mol. Graph. 14 (1996) 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.
127. M. Waskom, seaborn: statistical data visualization, J. Open Source Softw. 6 (2021) 3021. https://doi.org/10.21105/joss.03021.
128. Radusky, Leandro & Modenutti, Carlos & Delgado, Javier & Bustamante, Juan & Vishnopolska, Sebastian & Kiel, Christina & Serrano, Luis & Marti, Marcelo & Turjanski, Adrian. (2018). VarQ: A Tool for the Structural and Functional Analysis of Human Protein Variants. Frontiers in Genetics. 9. 620. 10.3389/fgene.2018.00620.
129. Jaganathan, Kishore et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell. 176. 10.1016/j.cell.2018.12.015.