



Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales.
Carrera de Cs. Biológicas

Tesis de Licenciatura Carrera ciencias biológicas

Estudio del impacto de las modificaciones de ClinGen a la clasificación de variantes de acuerdo a los criterios del American College of Medical Genetics (ACMG) en el contexto del diagnóstico de Enfermedades Poco Frecuentes (EPoFs).

Ana Rius Garrido

Director: Dr. Marcelo A. Martí.

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

Junio 2024

Estudio del impacto de las modificaciones de ClinGen a la clasificación de variantes de acuerdo a los criterios del American College of Medical Genetics (ACMG) en el contexto del diagnóstico de Enfermedades Poco Frecuentes (EPoFs).

Ana Rius Garrido
Director: Dr Marcelo Martí

Licenciatura en Ciencias Biológicas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina.

RESUMEN:

Con el rápido desarrollo de técnicas de secuenciación masiva, surge una disciplina orientada al análisis de variantes génicas para el diagnóstico genético (o molecular), en particular en Enfermedades Poco Frecuentes (EPoFs), que consiste en determinar cuál es la (o las) variantes que serían responsables del desarrollo de la patología.

En el año 2015, el Colegio Americano de Genética Médica y Genómica (ACMG) desarrolló un conjunto de recomendaciones para ponderar la evidencia presentada por cada una de las variantes, y de este modo llegar a una clasificación estandarizada. En los últimos 5 años, The NIH-funded Clinical Genome Resource (ClinGen), se ocupó de la revisión de estos criterios.

La presente tesis de licenciatura, se propuso analizar de manera profunda el impacto de las modificaciones ClinGen-Rev sobre la clasificación de variantes, comparándola con la clasificación original ACMG-2015. Tras analizar conjuntos de variantes etiquetadas y clasificadas por ambos criterios, se logró ver un impacto del 8% en el veredicto clínico de las mismas, observando que ClinGen-Rev logra una correcta categorización el 90% de las variantes, una mejora significativa en relación al 67% alcanzado por ACMG-2015. A su vez, hemos estudiado en profundidad las implicancias en el proceso de priorización y selección de variantes, y fue posible identificar las modificaciones mayormente responsables del cambio.

Study of the impact of ClinGen modifications on variant classification revisiting the criteria of the American College of Medical Genetics (ACMG) in the context of Rare Diseases (RDs) diagnosis

ABSTRACT:

With the rapid development of massive sequencing technologies, the analysis of genetic variants for clinical diagnosis has emerged, particularly in the context of Rare Diseases (RDs). Diagnosing these diseases involves identifying the genetic variants responsible for the pathology development.

In 2015, the American College of Medical Genetics (ACMG) established a set of recommendations to assess the evidence associated with each variant, aiming to achieve a standardized classification. Over the past 5 years, ClinGen, the NIH-funded Clinical Genome Resource, has reviewed these criteria.

This thesis examines the impact of ClinGen-Rev modifications on variant classification, comparing them with the ACMG-2015 original recommendations. After analyzing sets of genetic variants using both criteria, we observed an 8% impact on the clinical verdict of these variants. It was found that ClinGen-Rev modifications correctly categorized 90% of the variants, representing a significant improvement compared to the 67% achieved by ACMG-2015. Additionally, the implications on the prioritization and selection process of variants were deeply studied, identifying some modifications of significant importance.

INTRODUCCIÓN

Uno de los logros científicos más importantes del siglo XXI fue la secuenciación del genoma humano, un avance que ha revolucionado la comprensión de nuestra biología. La primera versión, presentada en 2001, fue el resultado del Proyecto del Genoma Humano (The Human Genome Project¹), una colaboración internacional encargada de secuenciar, identificar y catalogar los genes presentes en el ADN humano. De allí surge el **genoma humano de referencia**, que es una secuencia consenso (no es la secuencia de ADN de una sola persona) aceptada del genoma de nuestra especie, que los investigadores utilizan como estándar para comparar con los genomas de cada individuo, y/o de otras especies.

Cada persona tiene su propio genoma, el cual es aproximadamente 99.9% idéntico al genoma de referencia, el 0.1% restante, comprende distintos tipos de diferencias que se distribuyen a lo largo del genoma y que técnicamente se denominan variantes². La amplia mayoría de las variantes tienen poco o ningún efecto en la apariencia externa, el correcto metabolismo y funcionamiento de nuestro organismo, pero unas pocas pueden ser directamente responsables de causar ciertas enfermedades.

Con el rápido desarrollo de las técnicas de secuenciación masiva (que explicaremos un poco más adelante), surge una disciplina orientada al análisis de las variantes génicas para el diagnóstico genético o molecular, en particular de **enfermedades mendelianas**, también conocidas como monogénicas, que son aquellas que comprenden la mayoría de las **Enfermedades Poco Frecuentes (EPoFs)**. Las enfermedades mendelianas son trastornos genéticos causados usualmente por una única variante, como ser el cambio de un solo nucleótido (SNVs, del inglés Single Nucleotide Variant), y/o por pequeñas inserciones/delecciones, en un gen específico, que siguen un patrón de segregación

¹ (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921.
doi:10.1038/35057062

² The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073 (2010).

mendeliano (autosómico dominante, autosómico recesivo, ligados al X, ligados al cromosoma Y, disomía uniparental y trastornos de imprinting), de ahí el nombre.

Como ya se mencionó, se utiliza el término “variante” para describir cualquier diferencia genética entre el genoma observado en un individuo y el genoma humano de referencia. Se prefiere este término a mutación, dado que si bien las variantes son mutaciones, no siempre son causantes de enfermedad, y el término mutación suele referir a casos donde hay un efecto negativo y significativo en el fenotipo. Por otro lado, las variantes poseen una dada frecuencia poblacional, que representa cuántos alelos de la misma se han encontrado circulando en una muestra de la población. Usualmente se asocia el término mutación con las variantes poco frecuentes (menor 1% y en algunos casos 5%)³ y causales de patologías. A las variantes que poseen alta frecuencia, superior al 5%, se las suele referir como polimorfismos, y no suelen poseer un impacto sobre nuestra salud. Al contrario, son las responsables de la variabilidad genética poblacional, llegando a tener suficiente frecuencia para ser consideradas una variación genética común.

Diagnóstico de EPoFs

Realizar un diagnóstico molecular de una enfermedad mendeliana es un proceso de alta complejidad, con numerosos pasos, como se muestra en la *figura 1*, que consiste en determinar cuál es la (o las) variante responsable del desarrollo de la patología. En primer lugar, un profesional de la salud realiza un examen clínico integral, donde se evalúa el estado físico, antecedentes familiares y pruebas clínicas (en caso de corresponder). A continuación, partiendo de un diagnóstico clínico tentativo (o no)⁴ para una enfermedad genética, se procede a realizar la secuenciación del genoma (o fracción del mismo) del individuo, para luego con esos datos determinar las variantes, que serán analizadas en busca de aquella responsable de causar la patología.

³ Lista de excepciones provista por ClinGen

https://clinicalgenome.org/site/assets/files/3460/ba1_exception_list_07_30_2018.pdf

⁴ Zhang, J., Yao, Y., He, H., & Shen, J. (2020). Clinical interpretation of sequence variants. Current Protocols in Human Genetics, 106,e98. doi: 10.1002/cphg.98

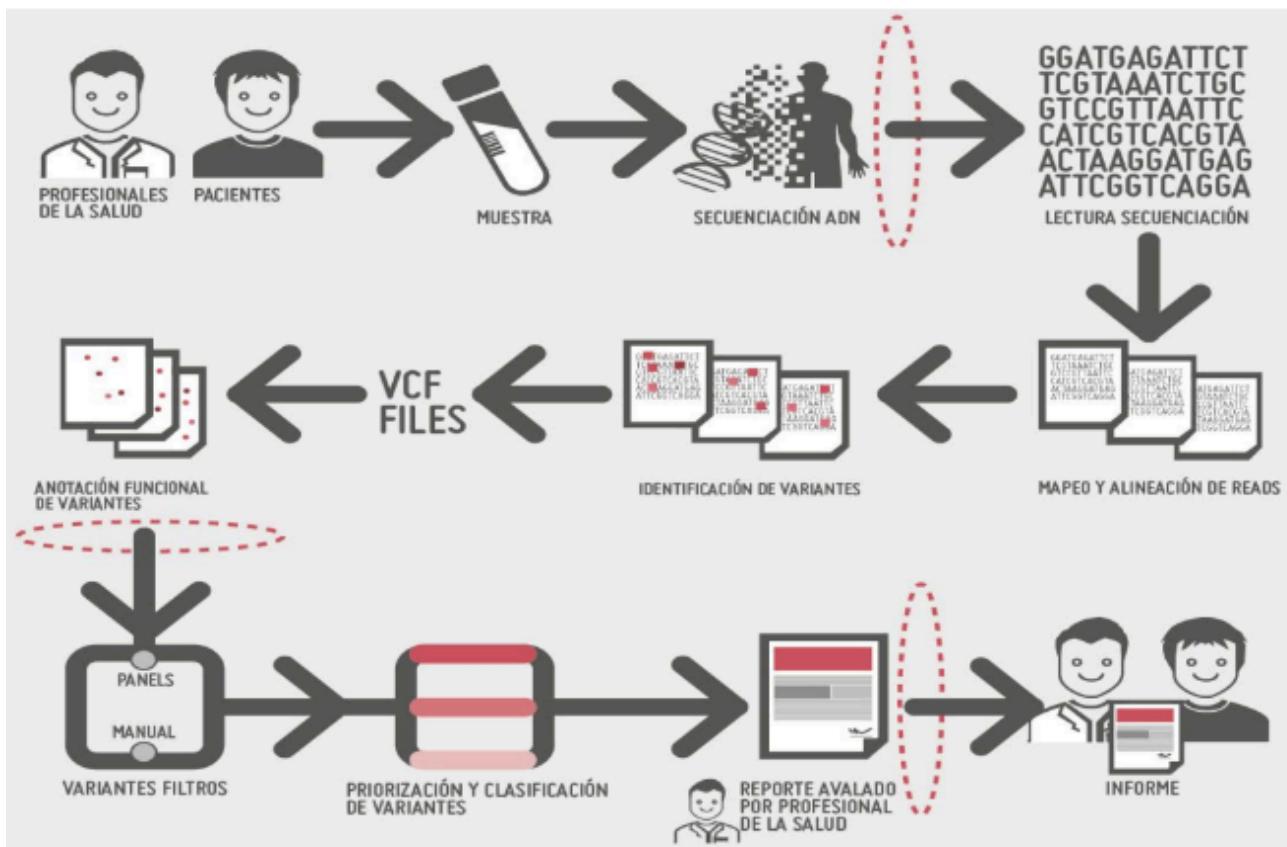


Figura 1: Proceso de diagnóstico de EPoFs. Pasos a seguir desde la sospecha de que la sintomatología del paciente es explicada por una EPoFs, hasta la obtención de un posible diagnóstico. Imagen extraída del curso “escuela de genética clínica” - <http://bioinfo.qb.fcen.uba.ar/egc2024/>

Tecnologías de Secuenciación Masiva.

La secuenciación de nueva generación (Next Generation Sequencing, NGS por sus siglas en inglés)⁵ es una tecnología que ha transformado el estudio de los genomas, al permitir secuenciar ADN (y también ARN) a gran escala. Esto significa que se puede conocer la secuencia de genoma de un individuo de manera mucho más eficiente que con las técnicas tradicionales, como la secuenciación por Sanger. Además, los costos y tiempos se han reducido significativamente, siendo actualmente un método bastante accesible.

⁵ Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H., & Inoue, I. (2012). Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *Journal of Human Genetics*, 57(10), 621–632. doi:10.1038/jhg.2012.91

El proceso de secuenciación, figura 2, comienza con la extracción del ADN de la muestra biológica que se desea analizar. Esta se corta en pequeños fragmentos, a los que se le incorporan adaptadores que permiten su adhesión a una superficie sólida, donde serán amplificados de manera independiente, aumentando la cantidad de muestra. Las técnicas de NGS⁶, más utilizadas, realizan secuenciación por síntesis, donde al duplicar cada hebra, en paralelo, se incorporan por apareamiento nucleótidos marcados con fluorescencia. Estas marcas son leídas pudiendo registrar qué base hay en una determinada posición, de manera similar al método de Sanger. El resultado de un experimento de secuenciación masiva es una serie de fragmentos, del orden de millones, que fueron secuenciados (o leídos) y que de ahí recibe el nombre de lecturas.

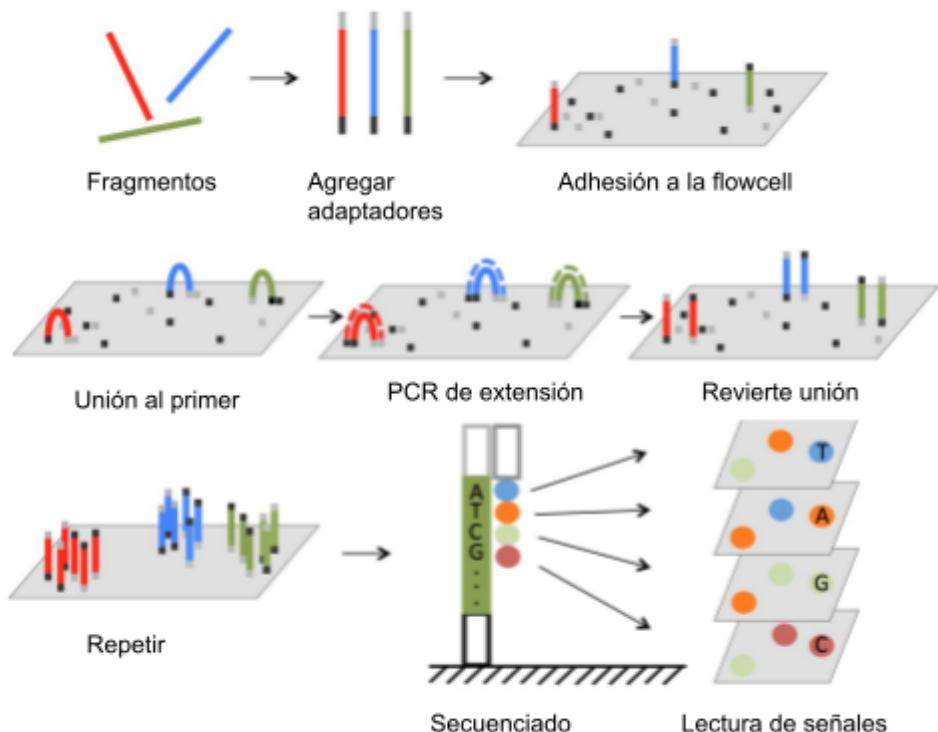


Figura 2: Esquema de pasos seguidos por un equipo de NGS como Illumina⁷.

⁶ Mastorosa, F.K., Miller, D.E. & Eichler, E.E. Applications of long-read sequencing to Mendelian genetics. *Genome Med* 15, 42 (2023). <https://doi.org/10.1186/s13073-023-01194-3>

⁷ Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10), 1932–1941. doi:10.1016/j.bbadi.2014.06.015

En el contexto del diagnóstico de EPOFs usualmente a partir de una muestra de sangre se procede a la **secuenciación del genoma, exoma o genes candidatos**, según lo que sea pertinente para el caso. Secuenciar el genoma se refiere a obtener lecturas del conjunto completo de material genético de un organismo, en este caso el paciente (23 pares de cromosomas, 22 autosomas y los cromosomas sexuales X e Y). Por otro lado, el exoma es la porción del genoma humano que contiene todos los exones, las regiones codificantes de genes y que representan en su conjunto aproximadamente el 1-2% del genoma humano total. Por otro lado, cuando el profesional sospecha que la afección está relacionada a una región, puede enviar a secuenciar unos determinados genes. Esto es lo que se conoce como secuenciación de un “panel” de genes , herramienta que se utiliza para analizar un conjunto específico de genes relacionados con una enfermedad.

Procesamiento de Datos y Llamado de Variantes

El bioinformático comienza su trabajo en el momento que se ha finalizado el proceso de secuenciación y los datos (las lecturas obtenidas) se encuentran *in-silico*. El procesamiento de las mismas (*figura 3*), para obtener las variantes, implica un **control de calidad** del resultado obtenido por el secuenciador, seguido de el **mapeo-alineamiento** de las lecturas contra el genoma de referencia. El Genoma de Referencia Humano es una secuencia de ADN que representa una "versión estándar" del genoma humano. La versión GRCh38 es la más completa y utilizada a partir de su última actualización en enero de 2022. Este actúa como un marco de trabajo para comparar y alinear las secuencias de ADN obtenidas de individuos específicos.

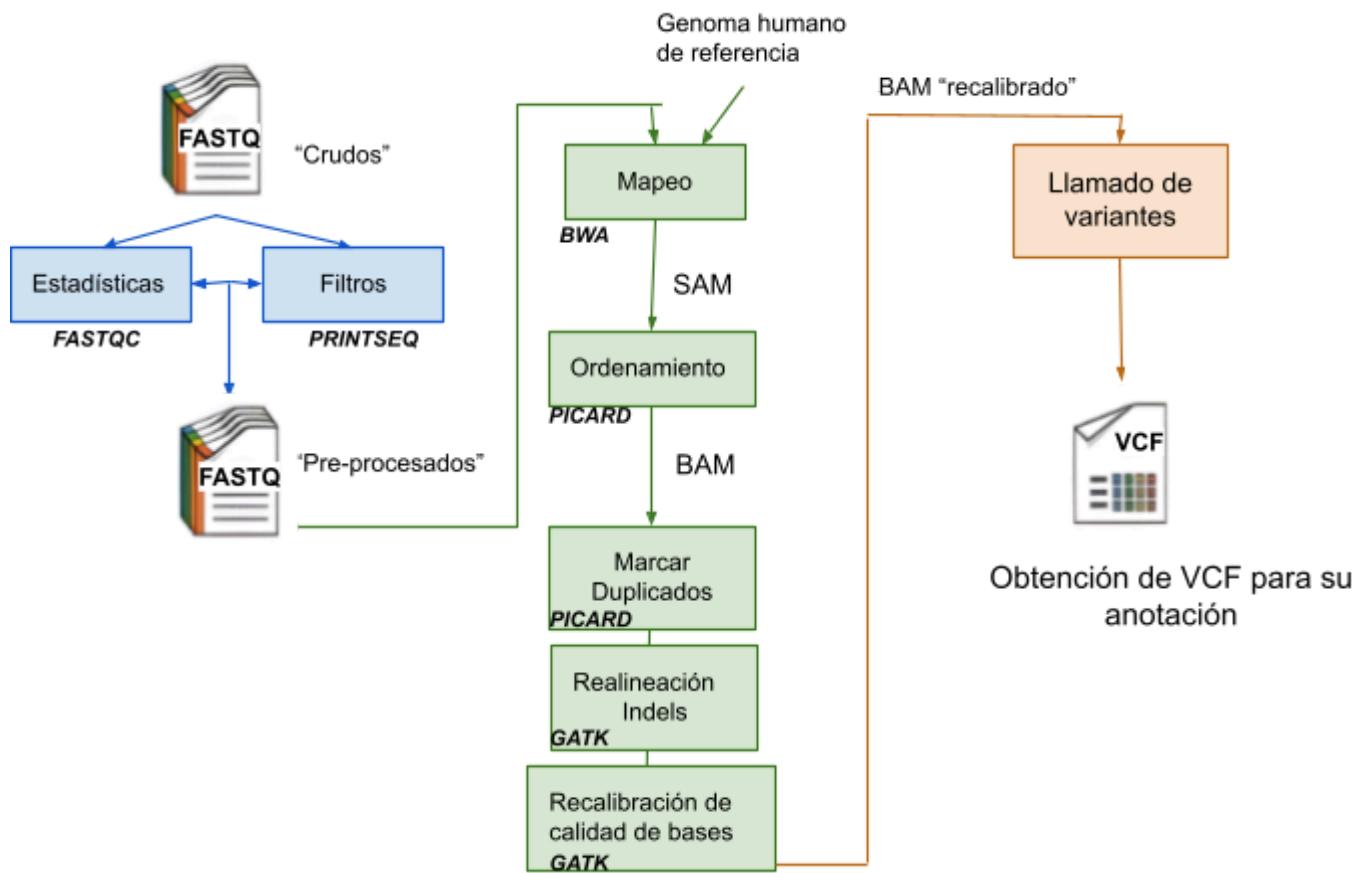


Figura 3: Esquema de pasos bioinformáticos hasta la obtención de variantes. Se comienza a trabajar con los datos crudos obtenidos de la secuenciación. Se continúa con su filtrado, mapeo y alineación (explicado más adelante) hasta llegar a la obtención de variantes en un VCF. El proceso tiene varios pasos y requiere de computadoras especialmente preparadas para el manejo de datos.

Las múltiples lecturas obtenidas de la secuenciación se alinean sobre la referencia. Cada lectura puede presentar pequeñas diferencias con respecto a la referencia, es por eso que se arma (estadísticamente) una secuencia consenso sobre la cual se detectan las variantes. Una vez alineadas las lecturas, figura 4, se realiza, finalmente el **llamado de variantes**, es decir se determinan todas las diferencias que la muestra presentó contra el genoma de referencia. Las variantes quedan dispuestas ordenadas por posición en un archivo de formato estandarizado

conocido como VCF (del inglés Variant Call Format), como se muestra en *la figura 5*, al que se le irá agregando la información que se obtenga en los próximos pasos.

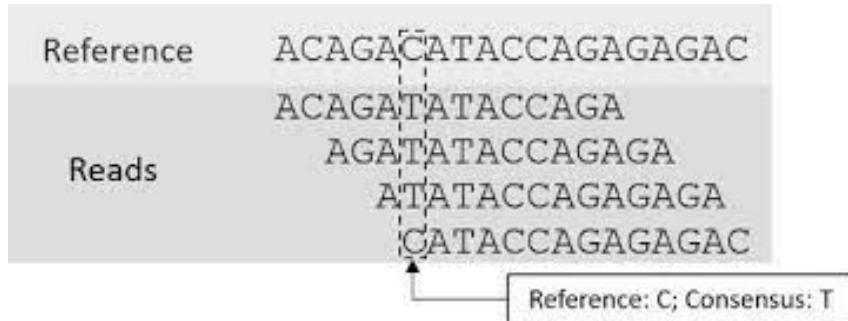


Figura 4: Representación de un mapeo y alineamiento de lecturas sobre el genoma de referencia.

```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT GEN
chr1 82154 rs4477212 a . . . . GT 0/0
chr1 752566 rs3094315 g A . . . . GT 1/1
chr1 752721 rs3131972 A G . . . . GT 1/1
chr1 798959 rs11240777 g . . . . GT 0/0
```

Figura 5: Fragmento de archivo VCF. El archivo VCF contiene un encabezado (header) que nos brinda información acerca del contenido de los campos *INFO* y *FORMAT*. Cada renglón representa una variante indicando su posición en el genoma ("CHROM" y "POS"), cambio de base con el genoma de referencia ("REF" y "ALT") y anotaciones que se tengan hasta el momento, adjuntadas en el campo "INFO"⁸.

⁸ Formato estándar de un VCF versión 4.0 detallado en <https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/VCF%20%28Variant%20Call%20Format%29%20version%204.0/encoding-structural-variants>

Anotación de variantes

La **anotación** consiste en agregar información biológica relevante a cada una de las variantes, buscando en bases de datos biológicas varias, citas bibliográficas, predicciones o análisis bioinformáticos específicos, e información de la clínica del caso.

Interpretación de variantes, surgimiento de ACMG

En este contexto, en el año 2015, el Colegio Americano de Genética Médica y Genómica (ACMG, del inglés American College of Medical Genetics)⁹ desarrolló un conjunto de recomendaciones para ponderar la evidencia presentada por cada una de las variantes, y de este modo llegar a una clasificación estandarizada que permita interpretar y reportar sus consecuencias. Para esto, a cada variante se le pueden asignar “etiquetas” de acuerdo a sus propiedades biológicas. Una vez asignadas las etiquetas de cada variante, la misma se clasifica en una de las siguientes categorías: “pathogenic” (P), “likely pathogenic” (LP), “uncertain significance” (VUS), “likely benign” (LB) o “benign” (B). Las variantes Likely benign y likely pathogenic significan que tenemos, estadísticamente hablando, una certeza mayor del 90%, pero menor al 99%, de que la variante sea benigna, o causante de la enfermedad según el caso, y las benign y pathogenic un nivel de confianza del 99%. Las variantes uncertain significance son aquellas en el medio, cuya información no es suficiente para descartar que tengan un posible efecto en la patología, y/o que presentan etiquetas contradictorias.

Existen 26 etiquetas diferentes, las cuáles se las pueden agrupar en 4 categorías.

i) Aquellas relacionadas con el efecto molecular, como ser un tipo de variante: missense, nonsense o frameshift. Pueden también evaluarse en ensayos funcionales o en predicciones in silico.

⁹ Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., ... Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5), 405–423. doi:10.1038/gim.2015.30

Se ha observado que las variantes que afectan las regiones codificantes de los genes suelen provocar efectos distintos en la función de la proteína. El impacto más frecuente está relacionado con la pérdida de función de la proteína mutada. Este fenómeno puede ser el resultado de cambios en un aminoácido que impiden la función normal, o bien, por ejemplo, si una base resulta en un codón STOP que interrumpe la traducción de la proteína. También pueden ocurrir delecciones/insertiones que modifican el marco de lectura o provocan el salto de un exón.

En principio, mediante el análisis computacional, es posible prever las alteraciones en la proteína a partir de cambios en su secuencia de aminoácidos, su ARN mensajero y su secuencia de ADN. Sin embargo, se fomenta la realización de ensayos funcionales. Estos experimentos son fundamentales para dilucidar de manera más precisa las consecuencias de una variante, permitiendo validar las predicciones computacionales y obtener información sobre el impacto real en la función biológica de la proteína.

ii) Aquellas relacionadas con la frecuencia alélica poblacional, debiendo consultar bases de datos de referencia para evaluar su recurrencia.

Desde la publicación del primer genoma humano en 2003, los genetistas han estado interesados también en determinar la variabilidad de los genomas en la población. Han habido múltiples esfuerzos en catalogar todas las variaciones comunes en el genoma, y cada vez más se incluyen aquellas de baja frecuencia poblacional, que como ya mencionamos, son de enorme importancia para el diagnóstico de enfermedades de origen genético. Actualmente, los datos son almacenados y organizados en bases de datos digitales públicas, disponibles gratuitamente, que pueden ser consultadas vía internet.

La población humana tiene una diversidad genética relativamente limitada, aunque la “rareza” es bastante común. Las variantes poco frecuentes constituyen, en muchos casos, las causas primarias de las enfermedades genéticas raras de herencia mendeliana y son limitadas las EPoFs que aparecen en alta frecuencia

poblacional. Sin embargo, el umbral de cuándo algo es “común” o “raro” no es fácil de fijar.

iii) Aquellas que tienen que ver con el fenotipo particular del caso relacionado al gen donde se encuentra la variante, con el fenotipo observado en el paciente y su información clínica. Se puede consultar bibliografía o reportes previos confiables.

Se han identificado y caracterizado las consecuencias de mutaciones germinales en más de 20.000 genes humanos. Las mutaciones asociadas a patologías se relacionan directamente a las variantes presentes en el ADN con la función o dosificación de la proteína alterada. De hecho, mucho de lo que se conoce acerca de la relación entre la función de los genes y los fenotipos humanos se basa en el estudio de variantes raras que subyacen a los fenotipos mendelianos que han sido ampliamente estudiados en el pasado.

A la hora de decidir si una variante está realmente relacionada con el diagnóstico, se estudia si existen variantes previamente reportadas que afecten al mismo gen con fenotipos semejantes al observado. Cobra importancia nuevamente la existencia de una base de datos curada, confiable y accesible que permita analizar estos datos.

iv) Análisis relacionados al patrón de segregación de la variante y su concordancia con el modelo de segregación de la patología que conferirá. Podría ser necesario expandir el análisis al grupo familiar.

La evaluación del patrón de segregación de la variante dentro de la familia proporciona información valiosa. Analizar cómo la variante se hereda a lo largo de las generaciones y su concordancia con el modelo de segregación esperado para la patología (recesivo, dominante etc.) puede fortalecer la evidencia de su asociación con la enfermedad. En algunos casos, puede ser necesario ampliar el análisis al grupo familiar para obtener una comprensión completa de la herencia de la variante y su impacto en la salud de los individuos relacionados.

Las etiquetas tienen códigos, o nombres asociados, que facilitan la referencia al criterio. Cada criterio patogénico se pondera como muy fuerte (PVS1, del inglés “pathogenic very strong”), fuerte (PS1–4, “pathogenic strong”); moderado (PM1–6, “pathogenic moderate”), o de apoyo (PP1–5, “pathogenic supportive”), y cada criterio benigno se pondera como independiente (BA1, “being stand alone”), fuerte (BS1–4, “benign strong”), o de apoyo (BP1–6, “benign supportive”). Por ejemplo, para asignar a una variante la evidencia de tener una alta frecuencia poblacional, se le coloca la etiqueta llamada BA1. Vale aclarar que la numeración dentro de cada categoría de etiquetas no agrega peso (es sólo una enumeración), de modo que BP4 y BP6 tienen el mismo nivel de evidencia. Ver *Tabla 1*.

		Benign		Pathogenic			
		Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4		
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1	
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3		
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1				
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2		
Allelic data		Observed in trans with a dominant variant BP2 Observed in cis with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3			
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5				
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4				

Tabla 1: Entorno de trabajo de cada etiqueta. El siguiente gráfico organiza cada uno de los criterios (etiquetas) según el tipo de evidencia, así como la fuerza de cada uno para una clasificación benigna (lado izquierdo) o patogénica (lado derecho). Abreviaturas: *BS*, benigno fuerte; *BP*, benigno de apoyo; *FH*, antecedentes familiares; *LOF*, pérdida de función; *MAF*, frecuencia alélica menor; *pat.*, patogénico; *PM*, patogénico moderado; *PP*, patogénico de apoyo; *PS*, patogénico fuerte; *PVS*, patogénico muy fuerte. Gráfico extraído de Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., ... Rehm, H. L. (2015). *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. *Genetics in Medicine*, 17(5), 405–423. doi:10.1038/gim.2015.30.

Es importante destacar que cada pieza de evidencia o etiqueta tiene un valor de score asignado según las recomendaciones presentadas por ACMG/AMP, pudiendo ser modeladas en un **marco bayesiano cuantitativo**¹⁰. Luego de anotar una dada variante y analizar qué etiquetas le corresponden, se realiza una ponderación de las mismas, llegando a su clasificación final acorde a la probabilidad estimada de patogenicidad, cómo se mencionó anteriormente. (Ver en ANEXO 1 ejemplo de variantes categorizadas).

Automatización del proceso de clasificación

Considerando que el genoma de un individuo presenta entre 2 a 5 millones de variantes, con unas aproximadamente 100 mil en el exoma, el proceso de anotación de variantes y su clasificación (también referido como interpretación) puede ser muy arduo¹¹. Es por ello que se han desarrollado diversos métodos (semi) automáticos para anotar y clasificar las variantes.

Para ello, se trabaja diseñando protocolos o “**pipelines**” bioinformáticos capaces de recorrer todo el camino a la interpretación de la variante, ver figura 6.

¹⁰Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., & Biesecker, L. G. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *GENETICS in MEDICINE*. doi:10.1038/gim.2017.210

¹¹ Eleanor G. Seaby, Reuben J. Pengelly, Sarah Ennis, Exome sequencing explained: a practical guide to its clinical application, *Briefings in Functional Genomics*, Volume 15, Issue 5, September 2016, Pages 374–384, <https://doi.org/10.1093/bfgp/elv054>

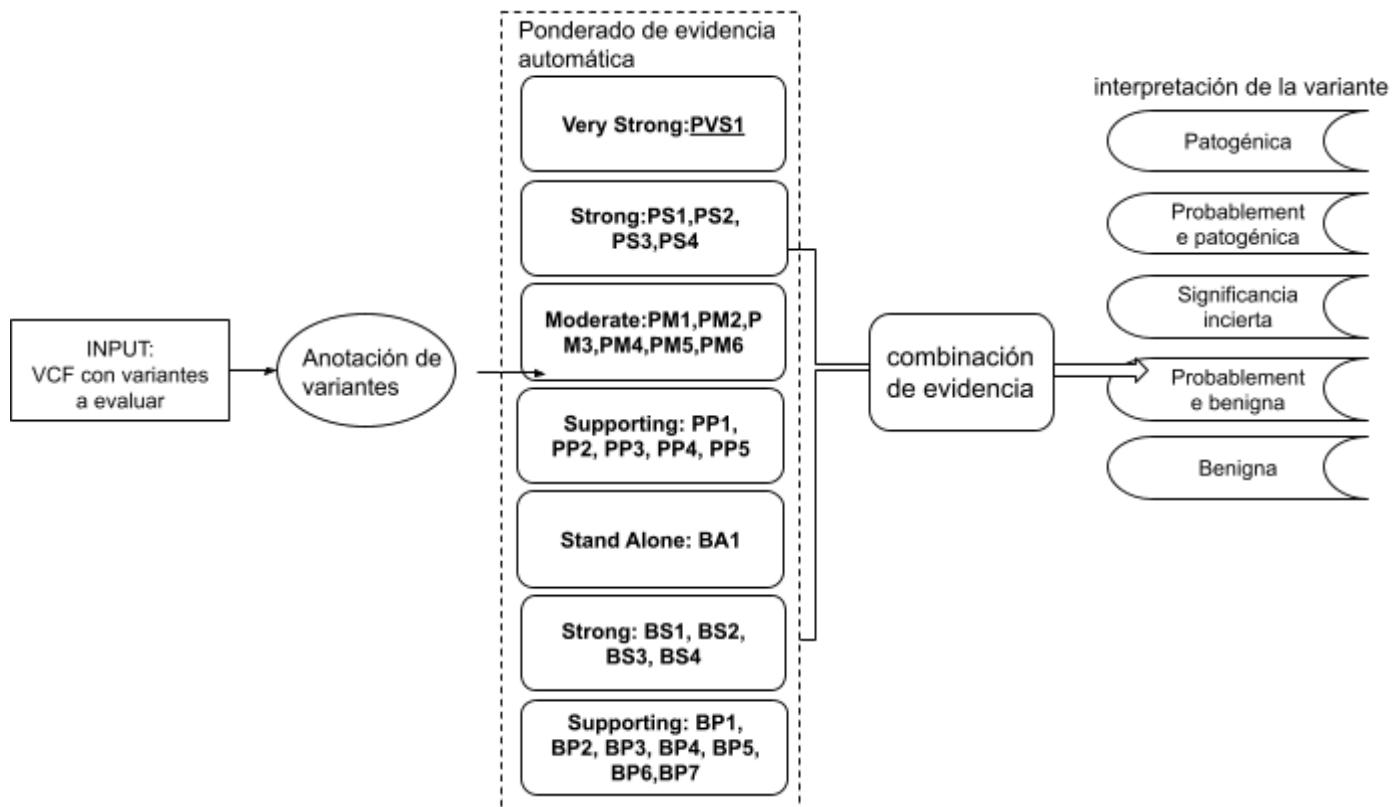


Figura 6: Recorrido de datos en un pipeline para clasificación de variantes: Los datos a analizar ingresan como VCF, para ser seguidamente anotados y etiquetados. Luego esa evidencia es ponderada para dar con una clasificación que permite interpretar a la variante.

Uno de los primeros algoritmos diseñados para clasificar las variantes genéticas en una de las cinco categorías fue InterVar¹². Esta herramienta se basa en los criterios establecidos por el ACMG para la interpretación de variantes genéticas. El objetivo principal de InterVar es usar datos poblacionales, datos funcionales y datos de segregación, para decidir en cada variante qué evidencias ponderar. Para esto usa árboles de decisión, ver figura 7, que recorre los criterios a incorporar.

¹² Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *The American Journal of Human Genetics*, 100(2), 267–280.
doi:10.1016/j.ajhg.2017.01.004

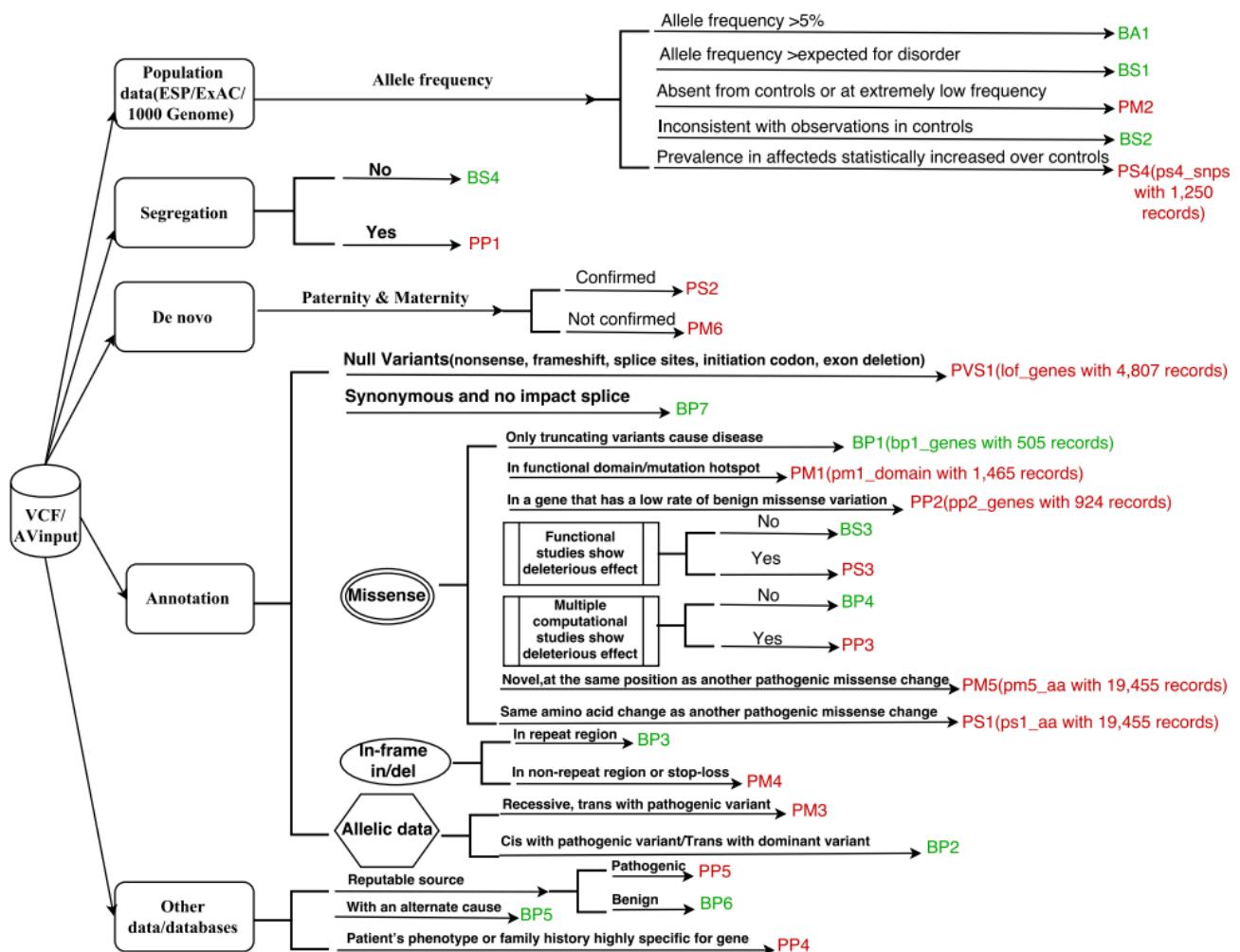


Figura 7: Árbol de decisiones usado por InterVar para la asignación de criterios de ACMG: Se realiza para cada tipo de evidencia una serie de evaluaciones que permiten arribar a una conclusión, colocando una etiqueta. Figura 2 extraída de Li, Q., & Wang, K. (2017). *InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines*. The American Journal of Human Genetics, 100(2), 267–280. doi:10.1016/j.ajhg.2017.01.004

De la anotación al diagnóstico

A partir de la interpretación lograda para cada una de las variantes se deben seleccionar aquellas con mayor probabilidad de ser causantes del fenotipo observado, a este proceso se lo denomina **priorización de variantes**. Es un trabajo manual que busca variantes que han logrado ser clasificadas como P o LP y podrían

estar relacionadas con la sintomatología presentada. En otras palabras variantes (L)P con significancia clínica. En los casos donde no se encuentran variantes relevantes explorando individualmente las patogénicas o probablemente patogénicas, se continúa con aquellas de significado incierto, hasta dar con una que pueda explicar el caso, y en los casos donde no se encuentra ninguna se realiza lo que se denomina un reporte negativo.

Una vez encontradas (o no) las variantes clínicamente relevantes se procede al reporte de las mismas al profesional de la salud, que será el responsable de terminar el diagnóstico y diseñar un tratamiento adecuado.

Actualización de los criterios de ACMG

Unos años después de la publicación de las primeras recomendaciones ACMG-2015, El consorcio Clinical Sequencing Exploratory Research (CSER) analizó estas directrices en 97 variantes que abarcan todas las categorías y que además fueron evaluadas por tres laboratorios independientes¹³.

Los laboratorios clasificaron cada variante utilizando tanto el método propio del laboratorio como los criterios de ACMG-2015. Solo hubo un 34% de concordancia para cualquier sistema de clasificación usado entre los laboratorios. Después de discusiones de consenso y una revisión detallada de los criterios de ACMG, la concordancia aumentó al 71%, ver *figura 8*. A partir de este resultado, se inició en la comunidad un proceso que busca clarificar cómo deben aplicarse los criterios, a fin de trabajar en un marco común y reducir la variabilidad en la clasificación e interpretación de variantes.

¹³ Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., ... Rehm, H. L. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *The American Journal of Human Genetics*, 99(1), 247. doi:10.1016/j.ajhg.2016.03.024

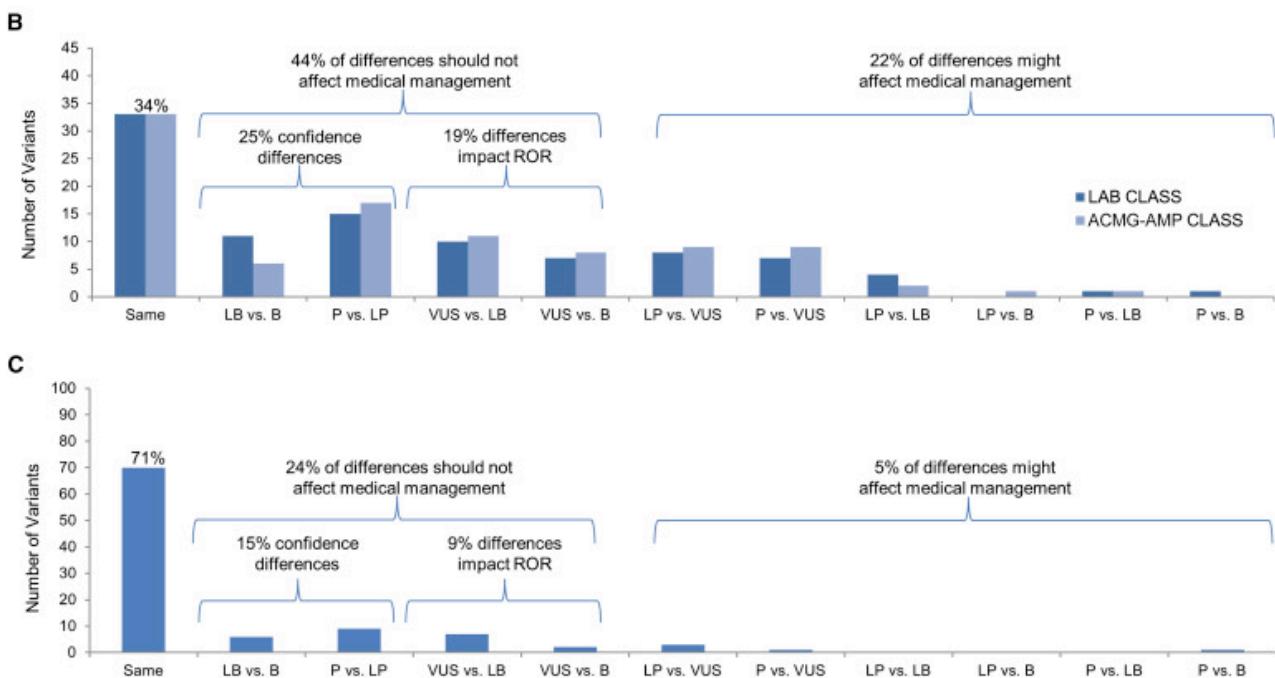


Figura 8: distribución de la clasificación de variantes comparadas por tres laboratorios: En B, se muestran los resultados de analizar 97 variantes tanto con las reglas propias de cada laboratorio como con los criterios de ACMG. En C, se muestran los resultados después de un consenso entre laboratorios. Las columnas se nombran con las categorías abreviadas: B (benign), LB (likely benign), VUS (uncertain significance), LP (likely pathogenic) y P (pathogenic). Figura 1 extraída de Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., ... Rehm, H. L. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *The American Journal of Human Genetics*, 99(1), 247. doi:10.1016/j.ajhg.2016.06.001

En los últimos 5 años, The NIH-funded Clinical Genome Resource (ClinGen)¹⁴, se ocupó de la revisión de estos criterios, particularmente los requerimientos y el peso de cada etiqueta, con el objetivo de refinar la clasificación. Por ejemplo, se buscó atenuar la cantidad de variantes clasificadas como “uncertain significance”, es decir, aquellas cuya evidencia disponible no alcanza para la toma de decisiones

¹⁴ Harrison, S. M., Biesecker, L. G., & Rehm, H. L. (2019). Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines. *Current Protocols in Human Genetics*, 103(1). doi:10.1002/cphg.93

clínicas. Para esto algunas etiquetas fueron redefinidas, y otras desdobladas en subetiquetas. A estas modificaciones se las conoce como modificaciones de ClinGen-Rev.

En resumen, los criterios del ACMG generaron las bases necesarias para empezar a clasificar variantes de manera estandarizada. Sin embargo, muchas decisiones quedaban a la libre interpretación del investigador, pudiendo resultar en una falla sistemática de criterio cuando el etiquetado es automatizado. Por otro lado, los avances en el conocimiento y la evaluación de los criterios sugieren una continua revisión de los mismos. En este contexto, las recomendaciones de ClinGen buscan proporcionar un enfoque más estructurado para la evaluación creando marcos de trabajos cuantitativos, y fijando umbrales de manera de convertir los criterios en herramientas cada vez más objetivas y precisas.

Objetivos

En el grupo de trabajo, donde se propone realizar la presente tesis, se han desarrollado herramientas computacionales para la anotación y clasificación de variantes de acuerdo a los criterios de ACMG-2015, y más recientemente se han incorporado las modificaciones de ClinGen-Rev. La metodología es utilizada regularmente en el grupo de trabajo para el análisis de pacientes con EPoFs, en el marco de otros proyectos de investigación, habiendo estudiado en los últimos 5 a 10 años miles de casos. **En este contexto, en la presente tesis de licenciatura, se propone utilizar los métodos y datos mencionados para analizar de manera profunda el impacto de las modificaciones ClinGen-Rev sobre la clasificación de variantes, comparando con las clasificación original ACMG-2015, y evaluar su efecto en el proceso de priorización de variantes en el contexto del diagnóstico molecular genómico de EPoFs.**

Específicamente, en la presente tesis se busca responder la siguiente pregunta **“Cuál es el impacto, en términos cuantitativos, de las modificaciones de ClinGen-Rev, en la anotación y clasificación de variantes, y cómo ello impacta en el**

proceso de priorización y selección de variantes en el diagnóstico molecular de enfermedades poco frecuentes”?.

MATERIALES Y MÉTODOS

En un primer paso, se propone la implementación de un código de etiquetado automático de variantes de acuerdo a los criterios originales de ACMG-2015, y un segundo paso, para los criterios ClinGen-Rev. Todos los programas serán realizados en Python y/o scripts de bash, y se tomará como punto de partida los programas/scripts ya utilizados en el grupo de trabajo. Se trabaja con un set de variantes curadas por el consorcio ClinGen, adquiridas a través de su repositorio de datos públicos. Además se utilizan VCFs de variantes obtenidas de muestras humanas ya secuenciadas y mapeadas, también obtenidas por el grupo de trabajo.

Anotación de variantes: buscando evidencias

Antes de la atribución de cualquier tipo de evidencia, las variantes son anotadas. En este proceso se recorre computacionalmente cada línea del VCF incorporando sistemáticamente información sobre el gen al que pertenece la variante, la frecuencia con la que la base cambia y por cuál es reemplazada. A su vez se realiza una predicción de los efectos moleculares de la variante para saber qué impacto podría tener en la proteína o en la expresión. También se anota la cigocidad con la que se encuentra, si la variante ya fue registrada con anterioridad y otras observaciones pertinentes. Para ello se utilizan diversas bases de datos biológicas algunas de las cuales describiremos a continuación

dbSNP

La base de datos de polimorfismos de un solo nucleótido (dbSNP del inglés *Database of Single Nucleotide Polymorphisms*¹⁵) es una base de datos pública y libre para el almacenamiento de variantes genéticas en diferentes especies. Fue desarrollada por el *National Center of Biotechnology Information* (NCBI) en colaboración con el *National Human Genome Research Institute* (NHGRI).

¹⁵ Smigelski EM, Sirotnik K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000 Jan 1;28(1):352-5. doi: 10.1093/nar/28.1.352. PMID: 10592272; PMCID: PMC102496.

dbSNP asigna un número de referencia a cada variante¹⁶. El catálogo RefSNP es una colección no redundante de variantes que luego de su registro fueron agrupadas. Existen dos tipos de categorías de identificación: “ss” y “rs”. La categoría “ss” refiere a single submitter que se asigna a cada caso reportado y la categoría “rs” refiere a “reference submissions” que es el conjunto de todos los ss reportados para esa variante. dbSNP fue una de las primeras bases de datos de variantes y es por ello que su identificador único, el rs#, es ampliamente utilizado en la comunidad como identificador de la variante para agrupar información de diferentes bases de datos.

gnomAD

gnomAD (*Genome Aggregation Database*)¹⁷ es un recurso desarrollado por una coalición internacional de investigadores, con el objetivo de unificar la información de experimentos de secuenciación. Es una base de datos con información de la frecuencia poblacional de cada variante. Además, proporciona parámetros asociados a la predicción de la tolerancia del gen a la pérdida de función (LOF, por sus siglas en inglés). Estos parámetros son importantes para evaluar el impacto funcional y su potencial rol en una enfermedad.

Es importante aclarar que dentro de gnomAD podemos encontrar versiones (llamadas releases, principalmente v2 y v3). En este trabajo se utiliza gnomAD v3.1.2, que abarca 76,156 genomas ya que está asociado al genoma de referencia con el que anotamos (GRCh38).

OMIM

¹⁶ Bruford, E. A., Braschi, B., Denny, P., Jones, T. E. M., Seal, R. L., & Tweedie, S. (2020). Guidelines for human gene nomenclature. *Nature Genetics*, 52(8), 754–758. doi:10.1038/s41588-020-0669-3

¹⁷ Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M; Genome Aggregation Database Consortium; Rehm HL, MacArthur DG, O'Donnell-Luria A. Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat*. 2022 Aug;43(8):1012-1030. doi: 10.1002/humu.24309. Epub 2021 Dec 16. PMID: 34859531; PMCID: PMC9160216.

Online Mendelian Inheritance in Man (OMIM)¹⁸ es una base de datos que cuenta con un amplio compendio de genes humanos y sus fenotipos asociados. A los genes de secuencia conocida se los puede encontrar también por posición en el genoma, permitiendo relacionar directamente a la variante con un gen. Además a cada gen se lo relaciona con las enfermedades que puede ocasionar, la cigosidad correspondiente y genera relaciones bibliográficas con otras bases de datos genómicas. También posee para cada patología las principales características fenotípicas asociadas a la misma.

Clinical Variants

Esta plataforma fue desarrollada para la interpretación de la relación entre variantes y fenotipos médicaamente importantes¹⁹. Para facilitar la evaluación de la relevancia médica de cada variante reportada, ClinVar agrega información de otras bases de datos de NCBI e informa si hay interpretaciones clínicas contradictorias. Los datos disponibles incluyen detalles del fenotipo, interpretación de significado funcional y clínico, metodología utilizada para capturar las variantes y evidencia de respaldo. Las entradas se clasifican también según la forma de obtención de datos (por ejemplo, pruebas clínicas, evaluación e investigación de la literatura) y el nivel o estado de revisión (presentación única, panel de expertos, etc.).

UNIPROT

UniProt es un repositorio de datos que proporciona información detallada sobre proteínas²⁰. Está compuesta por tres bases de datos principales (UniProtKB, UniParc y UniRef) en particular UniProtKB provee información sobre secuencias de proteínas y sus anotaciones funcionales, como ser su participación en vías metabólicas, modificaciones post traduccionales, relevancia de residuos

¹⁸ Hamosh, A. (2004). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research, 33(Database issue), D514–D517.

doi:10.1093/nar/gki033

¹⁹ <https://www.clinicalgenome.org/data-sharing/clinvar/>

²⁰ <https://www.uniprot.org/>

particulares (i.e sitio activo, sitios de interacción etc.), estructura de la misma, presencia de variantes naturales, y su asociación con enfermedades.

También registra variantes genéticas conocidas que han sido asociadas con fenotipos específicos o condiciones patológicas. Esto es útil para comparar las variantes encontradas en el análisis genómico con las variantes previamente caracterizadas.

REVEL

REVEL es un método para predecir la patogenicidad de las variantes missense basado en una combinación de puntuaciones de 13 herramientas individuales: MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP++, SiPhy, phyloP y phastCons²¹.

Las herramientas como Polyphen-2, son predictores del posible impacto de una sustitución aminoacídica en la estabilidad, estructura y función de una proteína humana. La predicción se basa en una serie de características estructurales, filogenéticas y de secuencia que caracterizan la sustitución. A continuación, estima la probabilidad de que la mutación sea perjudicial, en función de una combinación de propiedades estudiadas.

REVEL se entrenó utilizando variantes missense patogénicas, excluyendo aquellas utilizadas previamente para entrenar las herramientas que lo componen. La puntuación para una variante individual puede oscilar entre 0 y 1, siendo las puntuaciones más altas indicativas de una mayor probabilidad de enfermedad. En la presente tesis utilizaremos REVEL como nuestro predictor bioinformático para las variantes missense.

SpliceAI

²¹ Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., ... Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*, 99(4), 877–885. doi:10.1016/j.ajhg.2016.08.016

Es un algoritmo de predicción de cambios en el splicing para variantes sinónimas e intrónicas, es de código abierto y se basa en el aprendizaje profundo²². El empalme de los pre-ARNm en transcritos maduros es notable por su precisión, un cambio de base puede afectar el sitio de reconocimiento de la enzima de corte, causando que una región no deseada llegue a la traducción y a la proteína. Por otro lado, un cambio en una región intrónica puede causar que un exón necesario para la proteína no se transcriba al no ser reconocido correctamente por la maquinaria de splicing, resultando en la exclusión del exón durante el procesamiento del ARN mensajero (ARNm), y dando como resultado una proteína alterada o truncada que puede no funcionar correctamente. Se estima que entre el 9% y el 11% de las variantes patogénicas en pacientes con trastornos genéticos raros son causadas por esta clase mutación. SpliceAI será utilizado como predictor bioinformático para las variantes asociadas a alteraciones de splicing, ya sean intrónicas y/o en las regiones frontera de los exones.

Etiquetado

Cada uno de los códigos de etiquetado que se desea evaluar (ACMG-2015 y ClinGen-Rev), llamados “pipelines”, recomienda un conjunto distinto de bases de datos para la anotación. Esto se debe a varios motivos: razones históricas, la necesidad de registrar más variantes anotadas y curadas, y el continuo desarrollo de bases de datos más grandes y confiables. A fin de comparar en su totalidad las recomendaciones propuestas por ClinGen, se decide seguir también las principales preferencias en bases de datos. Así se utiliza para el pipeline fiel a ACMG-2015 las bases de datos allí sugeridas, y para el pipeline ClinGen-Rev las recomendadas en la revisión.

Una vez anotado el gen, el siguiente paso del pipeline bioinformático es el etiquetado. Este proceso también se realiza computacionalmente, y consiste en recorrer el archivo VCF con las variantes anotadas, evaluando qué etiquetas

²² Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., ... Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. doi:10.1016/j.cell.2018.12.015

corresponde colocar según la evidencia obtenida. *En la tabla 2 se muestran las principales diferencias de criterio de etiquetado de cada algoritmo.*

	Pipeline fiel a ACMG-2015	Pipeline fiel a CLINGEN-REV
B A 1	La frecuencia del alelo es mayor al 5% en proyecto de secuenciado de exoma, proyecto 1000 genomas o ExAC ²³ .	La frecuencia del alelo es mayor al 0.15% en cualquier base de datos de frecuencia poblacional con más de dos mil alelos testeados y variante presente en más de 5 alelos. Hay una lista de excepciones ²⁴ .
B S 1	La frecuencia del alelo es mayor de lo esperado para el trastorno.	Se define la frecuencia alélica máxima esperada para cada enfermedad: $(\text{prevalencia} \times \text{heterogeneidad})/\text{penetrancia}$ (bases de datos específicas (ExAC)) Se desdobra la categoría en distintos niveles de evidencia (Strong, supportive, moderate).

²³ Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissono D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016 Aug 18;536(7616):285-91. doi: 10.1038/nature19057. PMID: 27535533; PMCID: PMC5018207.

²⁴ Ghosh, R., Harrison, S. M., Rehm, H. L., Plon, S. E., & Biesecker, L. G. (2018). Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Human Mutation*, 39(11), 1525–1530. doi:10.1002/humu.23642

B S 2	La variante se observó en un individuo adulto sano con la misma <i>cigosidad</i> (mismo modelo de herencia evaluado) para una enfermedad totalmente penetrante a una edad temprana. ACMG-2015 no está programado para reconocer este nivel de evidencia, no coloca la etiqueta.	Sin cambios provistos por ClinGen-Rev. Comienza a colocarse.
B S 3	Estudios in vivo o in vitro funcionales que muestran que no hay efecto dañino en la función de la proteína o splicing.	Se dan recomendaciones para los evaluadores con respecto a la evaluación de la validez clínica de los datos funcionales y un marco provisional de cuatro pasos para determinar la fuerza de evidencia adecuada ²⁵ .
B S 4	Ausencia de segregación en miembros afectados de la familia.	Se incorpora un sistema de puntos para evaluar el fenotipo y la co-segregación como tipos de evidencia para respaldar o refutar un locus.

²⁵ Brnich, S. E., Abou Tayoun, A. N., Couch, F. J., Cutting, G. R., Greenblatt, M. S., ... Berg, J. S. (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Medicine*, 12(1).
doi:10.1186/s13073-019-0690-2

B P 1	Variante <i>missense</i> en un gen para el cual la enfermedad es producida principalmente por variantes que causan truncamientos en la proteína.	Sin cambios.
B P 2	Observada en <i>trans</i> con una variante conocida patogénica para un trastorno dominante de penetrancia completa u observado en <i>cis</i> con una variante patogénica conocida en cualquier patrón de herencia.	Sin cambios.
B P 3	Insertiones o delecciones dentro del marco de lectura (<i>in-frame</i>) en una región repetitiva sin función conocida (<u>podría programarse</u>).	Sin cambios.

B P 4	Múltiples líneas/predictores de evidencia computacional sugieren que no hay impacto en el producto genético (criterios de conservación, impacto evolutivo, plegado, etc.).	Se dan recomendación de qué algoritmos usar. Se recomienda el uso de la categoría Supportive y qué umbrales usa según gen/enfermedad ²⁶ .
B P 5	Variante encontrada en un caso con una base molecular alternativa para la enfermedad.	Sin cambios.
B P 6	Una fuente confiable reporta la variante como benigna, pero la evidencia no está disponible para el laboratorio para una evaluación individual.	Pese a la desrecomendación del uso de este criterio por parte de ClinGen, en la presente tesis nuestra versión ClinGen-Rev continúa con su uso ²⁷ .

²⁶ Vikas Pejaver, Alicia B. Byrne, Bing-Jian Feng. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. CELL. <https://doi.org/10.1016/j.ajhg.2022.10.013>

²⁷ Biesecker, L. G., & Harrison, S. M. (2018). The ACMG/AMP reputable source criteria for the interpretation of sequence variants. GENETICS in MEDICINE. doi:10.1038/gim.2018.42

B P 7	Una variante sinónima que computacionalmente no se predicen cambios en el splicing y el nucleótido no está altamente conservado.	Se dan recomendación de qué algoritmos usar. Se recomienda el uso de la categoría Supportive y qué umbrales usar según gen/enfermedad. Se aplica a variantes sinónimas e intrónicas utilizando spliceAI.
P V S 1	Variante (nonsense, frameshift, codón de inicio, deleterio única o múltiple) en un gen donde su falta de función es mecanismo conocido de enfermedad. -Precaución donde la pérdida de función no es un mecanismo conocido. precaución interpretando variantes en la región 3'. -precaución con variantes de splicing que evitan un exón pero la proteína no cambia. -precaución con la presencia de múltiples transcritos.	Se realizan consideraciones específicas para los distintos tipos de pérdida de función ²⁸ . VER ANEXO 2.
P S 1	Cambia el mismo aminoácido que en una variante patogénica previamente establecida, sin importar que sea el mismo cambio de nucleótido.	Sin cambios.

²⁸ Abou Tayoun, A. N., Pesaran, T., DiStefano, M. T., Oza, A., Rehm, H. L., ... Biesecker, L. G. (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Human Mutation. doi:10.1002/humu.23626

P S 2	De novo (padre y madre confirmados) en un paciente con la enfermedad sin historia familiar.	Se establece un enfoque cuantitativo según un valor de punto basado en la consistencia fenotípica y el estado de novo confirmado o asumido, determinando un nivel de confianza aplicable.
P S 3	Estudios funcionales in vivo o in vitro que muestran que hay efecto dañino en el gen.	Se dan recomendaciones para los evaluadores con respecto a la evaluación de la validez clínica de los datos funcionales y un marco provisional de cuatro pasos para determinar la fuerza de evidencia adecuada.
P S 4	La prevalencia de la variante en individuos afectados se incrementa significativamente comparado con la prevalencia en controles (<u>podría programarse</u>).	Se establecen umbrales de recuento de probandos que permiten alcanzar distintos niveles de evidencia. en supporting, Moderate, Strong, Very Strong.
P M 1	Se trata de un hot spot de mutaciones o un dominio funcional bien estudiado sin variante benigna.	Se propone usar bases de datos curadas y subdividir la etiqueta en 3 categorías basadas en un score bayesiano. Se intenta llevar la decisión hacia un esquema más cuantitativo que cualitativo.

P M 2	Ausencia de controles o frecuencia extremadamente baja si es recesivo en Exome Sequencing Project, 1000Genomes Project, o Exome Aggregation Consortium.	Se propone que la rareza es bastante común, se reduce el valor de la etiqueta a PM2_Supporting. Se expanden las bases de datos recomendadas.
P M 3	Para desórdenes recesivos, detectado en trans con una variante patogénica.	Para ayudar en la consistencia en la aplicación de PM3, ClinGen SVI propuso un enfoque cuantitativo. Según la fase de las dos variantes en cuestión (confirmadas en trans versus desconocidas) y la clasificación de la variante en el otro alelo, se otorga una calificación acorde, desdoblando la categoría en Supporting, Moderate, Strong, y Very Strong.
P M 4	La proteína cambia su longitud como resultado de una delección/inserción dentro del marco de lectura en una región no repetitiva ni una variante stop-loss.	Sin cambios.
P M 5	Cambio novel missense en un aminoácido donde otro cambio missense se determinó como patogénico.	Sin cambios.

P M 6	Asumida de novo pero sin confirmación de padre y madre.	Se establece un enfoque cuantitativo según un valor de puntaje basado en la consistencia fenotípica y el estado de novo confirmado o asumido, determinando un nivel de confianza aplicable.
P P 1	Cosegrega con la enfermedad en múltiples afectados de la familia en un gen que se sabe que causa la enfermedad.	Se incorpora un sistema de puntos para evaluar el fenotipo y la co-segregación como tipos de evidencia para respaldar o refutar un locus.
P P 2	Variante missense en un gen con baja tasa de variantes missense benignas, y en los genes en que las variantes missense son un mecanismo conocido de desarrollo de la enfermedad.	Se propone usar bases de datos curadas y subdividir la etiqueta en 3 categorías basadas en un score bayesiano. Se intenta llevar la decisión hacia algo más cuantitativo que cualitativo.
P P 3	Múltiples líneas de evidencia computacional que apoyan el efecto deletéreo en el gen o su producto.	Se da recomendación de qué algoritmos usar. Se recomienda el uso de la categoría Supportive y una escala continua de benign a pathogenic de score. Los umbrales además dependen del gen o la enfermedad.

P P 4	El fenotipo del paciente o la historia familiar es altamente específica para la enfermedad con una misma etiología genética.	Se incorpora un sistema de puntos para evaluar el fenotipo y la co-segregación como tipos de evidencia para respaldar o refutar un locus ²⁹ .
P P 5	Una fuente confiable reporta la variante como patogénica, pero la evidencia no está disponible para el laboratorio para una evaluación individual.	Pese a la desrecomendación del uso de este criterio por parte de ClinGen, en la presente tesis nuestra versión ClinGen-Rev continúa con su uso.

Tabla 2: Diferencias en los criterios en la etiquetación: A la izquierda se observa el pipeline fiel a los criterios recomendados por ACMG-2015, a la derecha los criterios con las actualizaciones de ClinGen-Rev. En rojo se muestran aquellas etiquetas no programables, en amarillo se destacan aquellas que podrían ser programadas pero no lo fueron. El resto de las etiquetas, fondo blanco, se colocan regularmente.

Es importante destacar que varias de las etiquetas propuestas poseen impedimentos a la hora de programarse. Estas son por ejemplo la PS2 o PM6 que involucran fenotipos paternos propios del caso particular. Para aplicar esta etiqueta tendría que secuenciarse el gen sospechoso en madre o padre, o tener un conocimiento profundo de la historia clínica familiar. Otro ejemplo podría ser BS3/PS3 referidas a estudios funcionales que revelan la naturaleza de la variante, que ante la ausencia de una base de datos curada al respecto se basan en la lectura e interpretación de trabajos de literatura por parte del investigador.

²⁹ Leslie G. Biesecker, Alicia B. Byrne, Steven M. Harrison. (2023). ClinGen guidance for use of the PP1/BS4 co-segregation and PP4 phenotype specificity criteria for sequence variant pathogenicity classification. Cell, VOLUME 111, ISSUE 1, P24-38. DOI:<https://doi.org/10.1016/j.ajhg.2023.11.009>

Modelo Bayesiano de Clasificación de Variantes

Como ya se mencionó, cada pieza de evidencia (etiqueta) posicionada sobre una variante es tratada de modo individual. El modelo original de ACMG/ AMP propone un modo de combinar las etiquetas para llegar al veredicto de clasificación (*figura 9 derecha*).

En el 2018 se desarrolló un modelo cuantitativo bayesiano que permite asociar a cada etiqueta (y a un conjunto de ellas) y por ende a cada variante una “chance” de patogenicidad (Odds of pathogenicity, OP). Sabiendo que las EPoFs son justamente poco frecuentes, todas las variantes parten de un score de 0.100 de OP. Al terminar el etiquetado entonces se realiza un cálculo combinando las etiquetas asignadas, llegando a un score bayesiano representativo de la OP y se establecen umbrales de score compartidos que permiten categorizar la variante en las 5 categorías mencionadas (ver *figura 9 izquierda*: pathogenic > 0.99, likely pathogenic > 0.90, 0.90 < uncertain significance < 0.10 , likely benign < 0.10 y benign < 0.01).

En la presente tesis luego de la anotación, asignación de etiquetas con ambos criterios (ACMG-2015 y ClinGen-Rev), se realiza para ambos el cálculo bayesiano y la clasificación del mismo modo. En función de la clasificación resultante y las etiquetas subyacentes es que analizamos los resultados presentados en el próximo capítulo.

Patogénica	<ul style="list-style-type: none"> i. 1 Muy fuerte (PVS1) Y <ul style="list-style-type: none"> a. ≥ 1 Fuerte (PS1–PS4) ó b. ≥ 2 Moderadas (PM1–PM6) ó c. 1 Moderada (PM1–PM6) Y 1 de Apoyo (PP1–PP5) ó d. ≥ 2 de Apoyo (PP1–PP5) ii. ≥ 2 Fuertes (PS1–PS4) iii. 1 Fuerte (PS1–PS4) Y <ul style="list-style-type: none"> a. ≥ 3 Moderadas (PM1–PM6) ó b. 2 Moderadas (PM1–PM6) Y ≥ 2 de Apoyo (PP1–PP5) ó c. 1 Moderada (PM1–PM6) Y ≥ 4 de Apoyo (PP1–PP5)
Probablemente patogénica	<ul style="list-style-type: none"> i. 1 Muy fuerte (PVS1) Y 1 Moderada (PM1–PM6) ii. 1 Fuerte (PS1–PS4) Y 1-2 Moderadas (PM1–PM6) iii. 1 Fuerte (PS1–PS4) Y ≥ 2 de Apoyo (PP1–PP5) iv. ≥ 3 Moderadas (PM1–PM6) v. 2 Moderadas (PM1–PM6) Y ≥ 2 de Apoyo (PP1–PP5) vi. 1 Moderada (PM1–PM6) Y ≥ 4 de Apoyo (PP1–PP5)
Benigna	<ul style="list-style-type: none"> i. 1 Independiente/Única (BA1) ii. ≥ 2 Fuertes (BS1–BS4)
Probablemente benigna	<ul style="list-style-type: none"> i. 1 Fuerte (BS1–BS4) and 1 de Apoyo (BP1–BP7) ii ≥ 2 de Apoyo (BP1–BP7)
Significado incierto	<ul style="list-style-type: none"> i. No cumple con los criterios mencionados anteriormente. ii. Se contradice la evidencia de patogenicidad con reportes de carácter benigno de la variante.

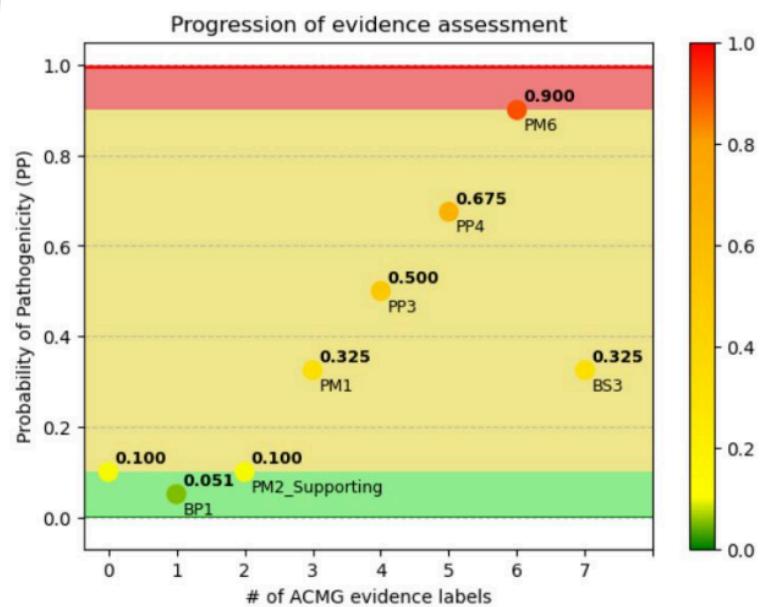


Figura 9: Paso de etiquetas a clasificación: A la izquierda combinaciones de etiquetas que permiten llegar a una clasificación determinada. A la derecha, cómo se distribuye cada etiqueta en la escala de probabilidad de patogenicidad.

Comparación de criterios

Se testea el desempeño de los algoritmos para una muestra de procedencia humana y para una muestra de variantes (también humanas) ya curadas manualmente por el consorcio ClinGen³⁰. Luego se indaga en las causas que generan las diferencias observadas.

I) Primeras cuantificaciones

Partiendo de un VCF de variantes y una vez recorrido ambos pipelines bioinformáticos, se obtienen dos VCFs que en la categoría “INFO” poseen todos los datos asociados a la anotación, cada uno de ellos etiquetado según el criterio correspondiente (ACMG-2015 o ClinGen-Rev), y con la consecuente categorización de sus variantes.

³⁰ Variantes curadas por ClinGen extraídas de su repositorio digital : <https://erepo.clinicalgenome.org/evrepo/>

En un primer acercamiento a la comparación de resultados se procede a la cuantificación diferencial. Diseñando programas específicos en Python y bash se busca responder cuantas variantes resultan clasificadas en cada categoría.

VCFpy

Es una biblioteca de herramientas diseñada para usar en Python para la escritura y lectura de archivos VCF³¹. Un VCF tiene un formato interno difícil de ver a simple vista. VCFpy posee funciones pensadas para facilitar la búsqueda de campos en el archivo, pues al procesar el formato son capaces de reducir la complejidad de la búsqueda y por lo tanto el tiempo de dedicación del usuario para realizar dicha tarea. Así es posible en segundos recorrer el archivo VCF guardando datos que puedan resultar interesantes en otro archivo de fácil manejo manual.

Además, VCFpy proporciona una amplia gama de funciones para realizar tareas avanzadas en estos archivos, como filtrar variantes, calcular estadísticas genéticas y realizar análisis específicos de variantes. Al ser una biblioteca activamente mantenida y desarrollada por la comunidad de bioinformática de Python es sencillo resolver problemas posteriores que puedan aparecer.

PANDAS

Pandas es una biblioteca de Python ampliamente utilizada en el ámbito de la ciencia de datos y análisis de datos³². Está construida sobre la base de otras bibliotecas importantes de Python como NumPy, y proporciona estructuras de datos de alto nivel y herramientas para manipular y analizar datos de manera eficiente. Un DataFrame es una estructura de datos bidimensional similar a una tabla en una base de datos relacional, donde cada fila y columna puede contener datos de diferentes tipos. Esto hace que sea conveniente para almacenar y manipular datos tabulares, como las variantes genéticas encontradas en un archivo VCF.

³¹ Documentación en <https://vcfpy.readthedocs.io/en/stable/>

³² Documentación en: <https://pandas.pydata.org/docs/>

Pandas tiene herramientas que permiten tanto la limpieza de datos, como la selección de subconjuntos y unión de los mismos, entre otras funciones. Estas operaciones se pueden realizar de manera eficiente, incluso en grandes conjuntos de datos, agilizando las operaciones, por ejemplo, a la hora de querer obtener una estadística con un alto número de variantes (de más de una muestra a la vez).

MATPLOT

Matplotlib es una biblioteca de visualización de datos en Python que proporciona herramientas para crear gráficos estáticos, interactivos y de alta calidad³³. Ofrece una variedad de tipos de gráficos, incluidos histogramas, diagramas de dispersión, diagramas de caja, gráficos de barras y gráficos de líneas (entre otras funciones), que son útiles para visualizar datos genéticos y sus características. Permite crear gráficos personalizables para comprender mejor los patrones y las relaciones en los conjuntos de datos, para comunicar sus resultados de manera efectiva.

II) Exactitud

Con el fin de evaluar si las calificaciones alcanzadas por cada uno de los pipelines (criterios) son correctas, se propone la comparación de los resultados de cada pipeline con una base de datos de variantes extraídas del repositorio del ClinGen. Estas variantes, al estar curadas, proveen un buen conjunto de datos testigo para la evaluación de la presión de cada algoritmo. Además, resultan un buen control positivo y negativo, pues intentan ser una muestra representativa cubriendo ejemplos de diferentes tipos de variantes como ser: missense, frameshift, nonsense, alta y baja frecuencia poblacional y distintas cigociidades.

La base de datos de variantes se descarga en formato .TSV y debe ser transformada a formato VCF para poder evaluarse en cada uno de los pipelines. Para esto se realizó un código combinando herramientas de Pandas y VCFpy que permitiera recorrer el .TSV e ir copiando registro a registro en un nuevo VCF.

³³ Documentación en: <https://matplotlib.org/>

Los identificadores utilizados en la anotación hasta el momento fueron del tipo *dbSNP*, pero en la base de datos curada se utiliza la nomenclatura HGVS. Fue necesario entonces entender las particularidades de la misma, incluirla en el programa y obtener el VCF con los IDs esperados.

Nomenclatura Human Genome Variation Society o HGVS

La nomenclatura HGVS busca denominar de manera inequívoca a cada una de las posibles variantes de nuestro genoma. Siempre posee una primera parte que indica la secuencia de referencia utilizada (NC_ si es una secuencia de referencia basada en un cromosoma, o NM_ referencia basada en un ARN que codifica una proteína). Hay una segunda parte del identificador conformada por prefijos, pueden ser varios según el tipo de secuencia al que se refiere la posición que le sigue ("g." para una secuencia de referencia genómica, "c." para una secuencia de referencia de ADN codificante). A continuación se indica la posición en el genoma de referencia, la base de referencia y su sustitución.

Por ejemplo, para la variante *NM_004006.2:c.5234G>A*; *NM_004006.2* es el identificador de la secuencia de referencia, con el ".2" indicando la versión de la misma y *NM_* indicando una secuencia de referencia basada en un ARN que codifica una proteína. La "c." indica que el tipo de secuencia de referencia es ADN codificante. 5234 es la posición de la variante en la numeración dada por la secuencia de referencia que estemos utilizando y G>A es el cambio de la base en la referencia al encontrado.

Una vez familiarizados con el formato se construyó una clave para traducir de una nomenclatura a la otra, pudiendo armar el VCF listo para introducir a cada uno de los pipelines. Con los VCFs obtenidos se realizó un análisis cuantitativo, similar al descrito previamente y se calcularon las respectivas coincidencias con las clasificaciones provistas por el consorcio ClinGen.

III) Comparación del modo de evaluar evidencia

Las diferencias en la clasificación obtenidas en los pasos anteriores son el resultado de un etiquetado diferencial. Se quiere estudiar cuáles son esas etiquetas que generan un mayor impacto. Como cada variante tiene, en general, más de una etiqueta, el análisis se torna complejo. Para reducir el problema a uno univariado, analizando las etiquetas de una a la vez, se realiza una comparación a través de un programa que permite visualizar los datos realizando filtrados sucesivos.

Power BI

Power BI es una herramienta versátil que puede ser utilizada para analizar y compartir datos³⁴. Ofrece una amplia gama de visualizaciones, como gráficos de dispersión, diagramas de barras, gráficos circulares (entre otros). Se pueden filtrar y segmentar datos fácilmente desde el gráfico, para ver subconjuntos específicos de variantes genéticas o etiquetas. Estas visualizaciones ayudan a explorar y comprender patrones en los datos de variantes genéticas.

Power BI puede integrar datos de diversas fuentes, como archivos .TCV y archivos .CSV. Estos datos pueden combinarse para ver distintos resultados en una misma planilla, por ejemplo distintas categorizaciones por distintos pipelines para un mismo ID de variante.

³⁴ Referencia del programa: <https://www.microsoft.com/es-es/power-platform/products/power-bi>

RESULTADOS

En esta sección se presentan los resultados obtenidos del análisis del impacto de las modificaciones ClinGen-Rev sobre la clasificación de variantes, comparándola con la clasificación original ACMG-2015. En una primera parte se trabaja con un conjunto de etiquetas de categorización conocidas, pudiendo estudiar la exactitud y principales diferencias entre ambos métodos.

En una segunda instancia, a partir de los resultados obtenidos previamente, se exploran las etiquetas cuya modificación explica buena parte de las diferencias obtenidas en el veredicto final de cada variante.

Por último, se analizan por ambos métodos VCFs de muestras reales, buscando entender cómo se ve afectado el proceso de priorización de variantes en un escenario de trabajo real en genómica clínica tras la incorporación de las modificaciones de ClinGen.

I) Comparación de la exactitud obtenida trabajando con variantes resueltas

Para evaluar la exactitud de ambos criterios, se analizaron un conjunto de 4764 variantes curadas extraído del repositorio de ClinGen. Para simplificar los análisis cuantitativos de la mayoría de los gráficos, a continuación se muestran agrupando en benign las benignas y probablemente benignas y en patogénico las probablemente patológicas y patogénicas. En *la figura 10* se observa en cada fila la clasificación provista por el consorcio ClinGen (a la que consideramos “idónea”) y en colores la alcanzada por cada uno de los protocolos evaluados: ACMG-2015 o ClinGen-Rev. Los resultados se analizan por separado para cada una de las 3 categorías.

Como puede observarse de la figura, de las 700/4,764 variantes benignas analizadas ACMG-2015 solo clasifica correctamente el el 55.71%, marcando el resto como de significado incierto; En cambio ClinGen-Rev parece más exacto, ya que clasifica correctamente el 99.5% de las variantes benignas.

Clasificación diferencial alcanzada por cada criterio de etiquetado

Clasificación según etiquetado: ● Benignas ● VUS ● Patogénicas

Clasificación según consorcio ClinGen

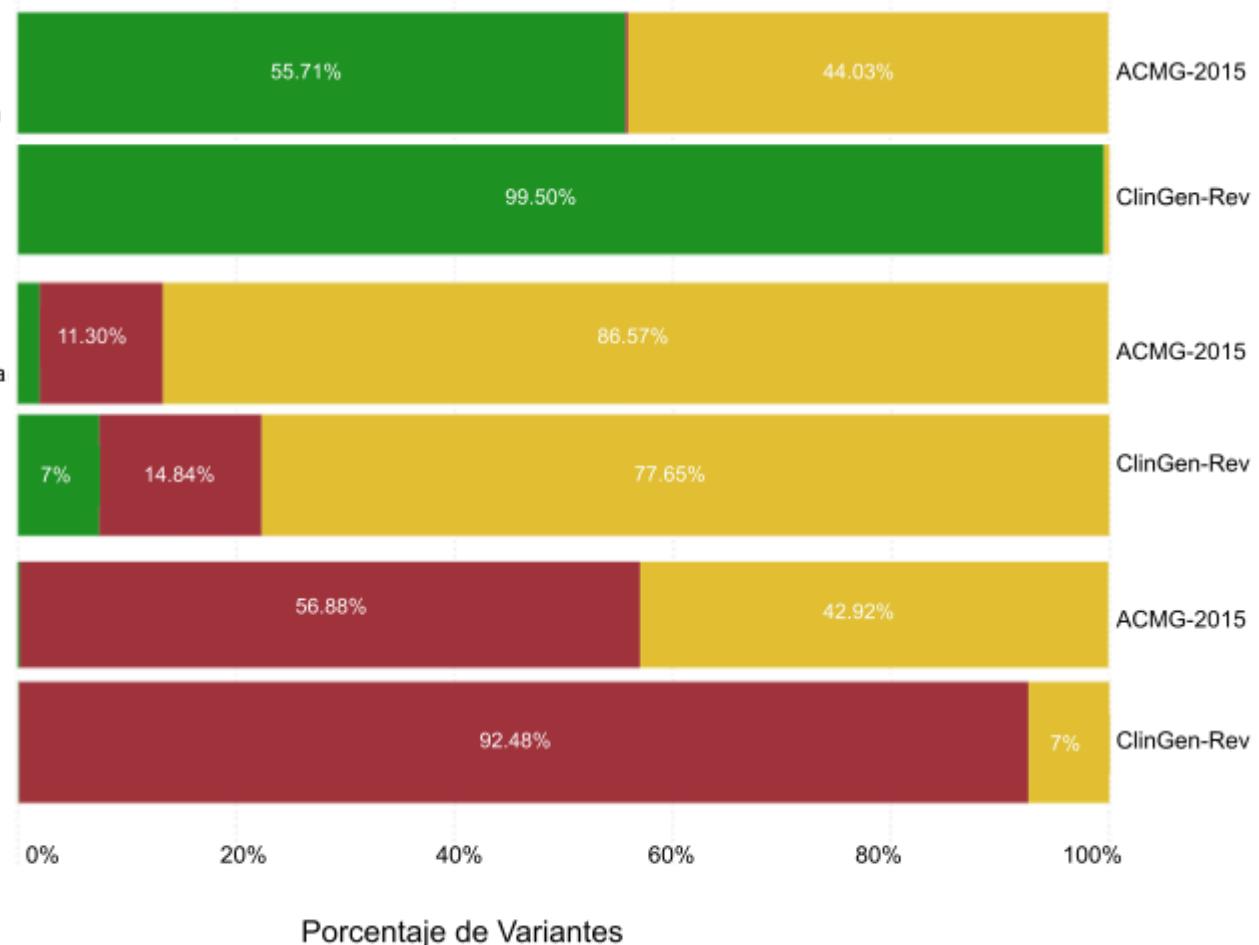


Figura 10, Diferencia porcentual en la clasificación final alcanzada por variantes analizadas provenientes de un repositorio de ClinGen. Cada fila representa las variantes que alcanzan una clasificación según los expertos de ClinGen. En colores se ve cómo se clasifican esas mismas variantes según ACMG-2015 o ClinGen-Rev.

Si nos vamos al otro extremo, para las 2,007/4,764 variantes patogénicas, también se observa una notoria mejora en la exactitud, ya que mientras ACMG-2015 consigue una coincidencia con el consorcio de 56.88%, ClinGen-Rev logra clasificar correctamente un 92.48%, de las patogénicas. Es importante destacar, que tanto para las B como para las P las variantes “mal” clasificadas, pasan en ambos casos y ambos criterios a variantes VUS, no encontrándose saltos de probable benigna a

probable patogénica y vice-versa. Esto es importante ya que muestra que los errores de ambos criterios son moderados.

Cuando finalmente analizamos las variantes de significado incierto, vemos que el comportamiento de ambos criterios es similar, siendo la gran mayoría de las variantes correctamente asignadas a esta categoría. Es interesante notar que ACMG-2015 suele clasificar erróneamente como VUS tanto variantes B y P, mientras que ClinGen-Rev tiende a equivocarse, aunque en menor medida, marcando como VUS variantes P.

<i>Coincidencia</i>	ACMG-2015	ClinGen-Rev	
<i>Consorcio ClinGen</i>	67%	90%	
<i>Cuantificación</i>	<i>Benignas</i>	<i>Inciertas</i>	<i>Patogénicas</i>
Consorcio ClinGen	1,221	1,536	2,007
ACMG 2015	7,01	2,769	1,294
ClinGen-Rev	1,297	1,488	1,979

Tabla 3: datos extraídos de la clasificación diferencia de las 4,764 variantes curadas.

Analizando los datos en conjunto (*Tabla 3*) vemos que ClinGen Rev posee una exactitud del 90% mientras que ACMG de sólo el 67%, lo que evidencia la relevancia clínica y potencial impacto de utilizar las recomendaciones de ClinGen. Sabiendo la exactitud de cada pipeline, dada la clasificación de una variante cualquiera, se puede decir que el veredicto provisto por ClinGen-Rev tiene mayor probabilidad de estar en lo correcto.

Tomando los resultados en conjunto, da la impresión de que las modificaciones provistas por ClinGen-Rev logran hacer a las variantes un poco más “benignas”. Para profundizar en el análisis de esta observación realizamos un análisis cuantitativo de la clasificación de las variantes utilizando el esquema bayesiano. En *la figura 11*, se muestran 40 variantes que obtuvieron un score de 0.5 en el pipeline ACMG-2015 (o sea son clasificadas como VUS) y los scores

resultantes de estas variantes para ClinGen-Rev. Como puede observarse, sólo 1 variante mantiene el mismo score de 0.5, y todas las demás bajaron su puntaje de manera significativa. Como consecuencia de este cambio en el puntaje, de las 40 variantes de significancia incierta, solo 14 mantienen la categoría VUS y las restantes resultan probablemente benignas o incluso benignas. En la próxima sección analizaremos qué etiquetas son específicamente las principales



responsables de este comportamiento.

Figura 11: Cambio en los scores intermedios alcanzado por cada pipeline: Se analizan 40 variantes categorizadas por el criterio de etiquetado de ACMG-2015 como significancia incierta y con un score bayesiano de 0.50.

Es importante destacar, que en líneas generales los resultados son esperados, pues uno de los conceptos claves de las modificaciones de ClinGen-Rev fue reducir el potencial patogénico de diversas etiquetas. Así, ClinGen-Rev es muy bueno identificando variantes benignas, separándolas rápidamente de las VUS o LP que podrían tener significancia clínica para el caso. Como se verá más adelante, esto facilitará enormemente el trabajo de análisis, ya que las variantes benignas son

usualmente descartadas del análisis de manera automática al no poseer relevancia clínica. Por otro lado, la observación de una fuerte disminución de las variantes clasificadas como patogénicas, implica la reducción de falsos positivos (variantes reportadas como patogénicas que no lo son) (VER ANEXO 3, matriz de confusión).

II) Comparación del modo de evaluar evidencia

A Continuación se realiza el análisis de algunas etiquetas específicas que permiten explicar gran parte de los cambios observados en las clasificaciones provistas por ClinGen-rev.

a) Clasificación diferencial en evidencias relacionadas con la frecuencia alélica poblacional (BA1/BS1).

BA1 y BS1 son etiquetas relacionadas a la frecuencia alélica poblacional de la variante y son las evidencias más fuertes hacia la benignidad. Vale la pena aclarar que estas BA1/BS1 no pueden ser colocadas juntas en ClinGen-Rev, entendiendo que responden al mismo tipo de evidencia. ClinGen-Rev utiliza bases de datos menos restrictivas y un umbral diferencial para cada gen de lo que representa una “alta frecuencia poblacional” o “frecuencia creíble para el trastorno”. A su vez, se mantiene en BA1 la categoría Stand Alone que convierte a la variante automáticamente en benigna (sin importar que otras etiquetas tenga), y se subdivide a BS1 en dos niveles de evidencia. Estas modificaciones se reflejan en la figura 12, donde se analizan variantes provenientes de la base de datos curada de ClinGen. Estas variantes son clasificadas correctamente por ClinGen-Rev como probablemente benignas (derecha) o benignas (izquierda). ClinGen-Rev asigna las etiquetas BA1 (frecuencia alélica poblacional alta) o BS1 (frecuencia del alelo mayor a lo esperado para el trastorno). Las variantes están ordenadas en columnas según la clasificación de ACMG-2015.

Clasificación de variantes que tienen asignadas BA1 o BS1 en ClinGen-Rev

BA1 o BS1 (según ClinGen-Rev): ● BA1:StandAlone ● BS1:Strong ● BS1:Supporting

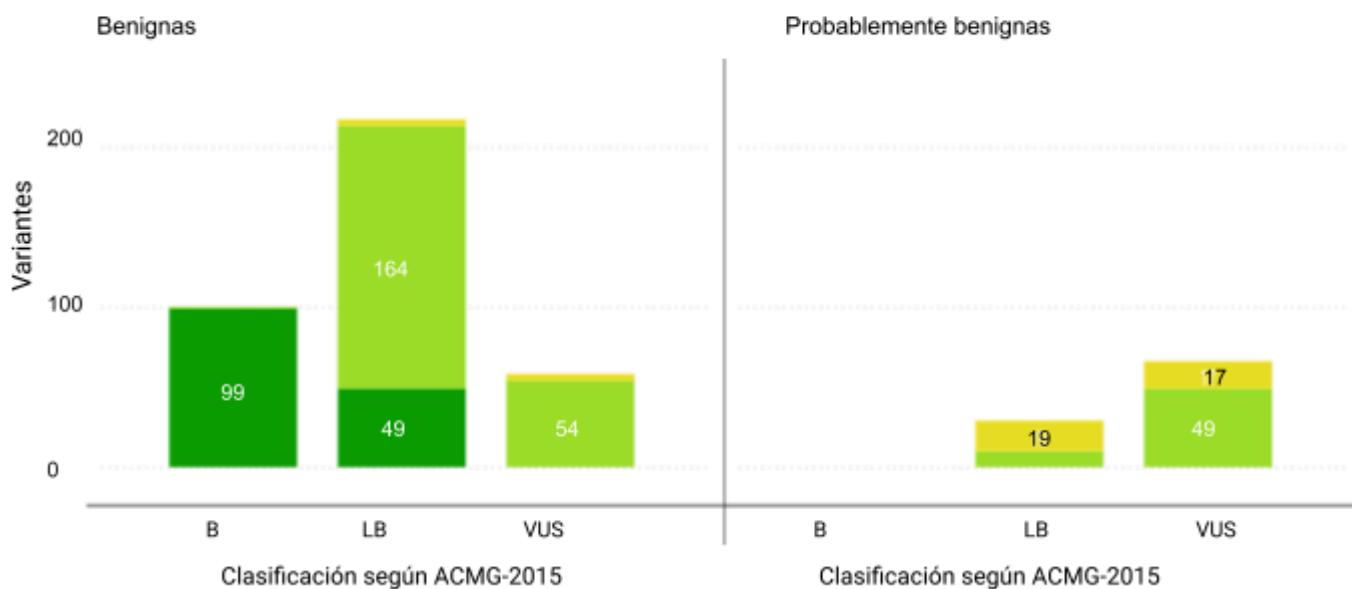


Figura 12: Colocación y distribución diferencial de BA1 y BS1: Se analiza a qué categoría llegan en el pipeline ClinGen-Rev, las variantes curadas a las que atribuye BS1 o BA1 (aunque tengan otras etiquetas además). En cada recuadro se ve la categoría alcanzada según ClinGen-Rev y en cada columna la categoría atribuida por ACMG-2015.

La primera observación es que como era de esperarse la mayoría de las variantes clasificadas por ClinGen-Rev como B son etiquetadas con un BA1 o un BS1 Strong, mientras que aquellas que terminan como VUS poseen BS1 strong o supporting. Cuando analizamos cómo eran clasificadas las mismas por ACMG-2015 lo que vemos es que un 25% de las posiblemente benignas, adquieren etiqueta BA1 por lo que pasan a ser benignas, y que el 75% de las VUS resultan etiquetadas como BS1 strong, lo que explica que terminen clasificadas como (probablemente) benignas.

Así, una de las principales mejoras de ClinGen-Rev es la utilización de umbrales diferenciales, bases de datos más actualizadas y una consecuente mayor precisión para la asignación de BS1. La modificación de BA1 junto con BS1 logran categorizar correctamente 118 variantes como LB/B que ACMG-2015 marca como VUS, reduciendo un 22% las falsas VUS que en realidad son actualmente benignas.

b) Clasificación diferencial en etiquetas relacionadas a un efecto molecular patogénico (PVS1, PM1)

La *figura 13*, muestra un gráfico análogo al mostrado en (a) pero con foco en las etiquetas asociadas al efecto molecular de la variante. PVS1(variante nula en gen con un mecanismo conocido de enfermedad) y PM1 (variante en un hot spot mutacional o dominio funcional). Nuevamente, PVS1 y PM1 no se colocan juntas en ClinGen-Rev, pues responden al mismo tipo de evidencia. Además en ClinGen-Rev estas etiquetas fueron ajustadas para tener umbrales en función del gen y desdobladas en categorías que permitan un mejor peso de la evidencia. Es interesante notar que, a diferencia de lo visto en la *figura 12* donde más variantes eran clasificadas como benignas, en la *figura 13* la clasificación de las variantes patogénicas se mantiene consistente entre ambas anotaciones.

La mayor parte de los cambios se producen en “the twilight zone” entre las VUS y probablemente patogénicas, que es crítica en casos donde determinar una variante de relevancia clínica es difícil. Estas variantes son unas pocas (11), pero ClinGen-Rev logra categorizarlas correctamente como patogénicas gracias a la asignación de etiquetas como PVS1:supportive o PM1, que confieren menor score de patogenicidad , y que no pueden ser asignadas por ACMG-2015.

Clasificación diferencial de variantes que tienen asignadas PVS1 o PM1 por ClinGen-Rev

PVS1 o PM1 (Según ClinGen-Rev): ● PVS1:Very Strong ● PVS1:Strong ● PVS1:Moderate ● PVS1:Supporting ● PM1:Supporting

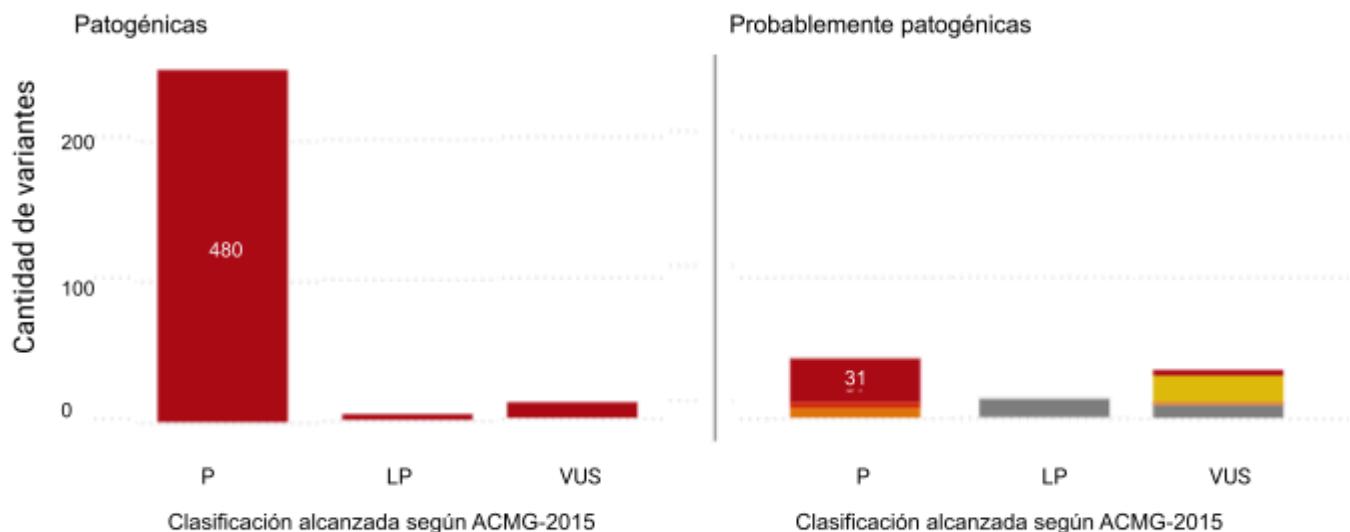


Figura 13: Colocación y distribución diferencial de PVS1 y PM1: Se analiza a qué categoría llegan en el pipeline ClinGen-Rev, las variantes a las que atribuye PVS1 o PM1 (aunque tengan otras etiquetas además). En cada recuadro se ve la categoría alcanzada según ClinGen-Rev y en cada columna la categoría atribuida por ACMG-2015.

En ambos criterios PVS1 y PM1 es una etiqueta sumamente relevante a la hora de identificar variantes como patogénicas o probablemente patogénicas. La granularización de PVS1 en ClinGen-Rev (permitiendo moderar según la evidencia el nivel como very strong, strong, moderate o supportive) permite categorizar de manera más exacta cada variante, por ejemplo rescantando variantes con un efecto molecular de patogenicidad intermedia, muy suave para el PVS1 utilizado en ACMG-2015, y muy fuerte para PM1. En estas variantes, ClinGen-Rev permite sumar para la clasificación final evidencia que si bien no es tan fuerte en su patogenicidad para ser un PVS, la misma, con menor valor (i.e PM) sumada al conjunto de las otras etiquetas permite clasificarlas como probablemente patogénicas.

c) Breve revisión sobre el modo de combinar criterios y etiquetado excesivo

Un problema frecuente, en la anotación de variantes, sobre todo la automática, es el de la doble contabilización. Esto ocurre cuando se cuenta un mismo nivel de evidencia dos veces para un mismo lado, como ser el caso de BA1 y BS1 en relación a la frecuencia alélica poblacional. Colocar etiquetas que se superponen puede llevar a una sobrevaloración de la evidencia, y la consecuente clasificación inexacta de la variante.

En este aspecto, el pipeline ClinGen-Rev toma medidas que ACMG-2015 no. ACMG-2015 recorre todos los niveles de evidencia y decide si corresponde colocarlo o no. En cambio, ClinGen-Rev es cuidadoso a la hora de evaluar cada evidencia y una vez colocada una etiqueta, esto elimina la posibilidad de colocar otras que serían consideradas “repetidas” en cuanto al fenómeno biológico subyacente. Un ejemplo estadístico de esto es que ACMG-2015 pone en promedio 3.5 etiquetas por variante, mientras que ClinGen-Rev coloca sólo 2.8. Estos valores son ilustrativos, y está claro, dependen mucho de la naturaleza de las variantes analizadas y las etiquetas. Pero resulta interesante pensar que ClinGen-Rev está logrando una clasificación más precisa en parte mediante la reducción del “sobreetiquetado” y la combinación inteligente de criterios.

Un buen ejemplo del etiquetado excesivo, es representado por el caso de PM2 (variante de novo o frecuencia extremadamente baja en gen recesivo). ACMG-2015 coloca esta etiqueta 3,973 veces trabajando sobre las 4,764 variantes extraídas de ClinGen, mientras que ClinGen-Rev la coloca solo 3,326 veces. Esto se ve en la *figura 14*, que muestra las variantes clasificadas por el consorcio de ClinGen, que tienen asignada la etiqueta PM2 en ACMG-2015 o ClinGen-Rev, y en colores la categorización a la que llegan en cada caso. Pueden observarse 429 variantes que cuentan con todas las propiedades para ser benigna según ClinGen, pero ACMG-2015 termina clasificándolas como de significado incierto únicamente por tener baja frecuencia poblacional.

Clasificación diferencial de variantes con PM2

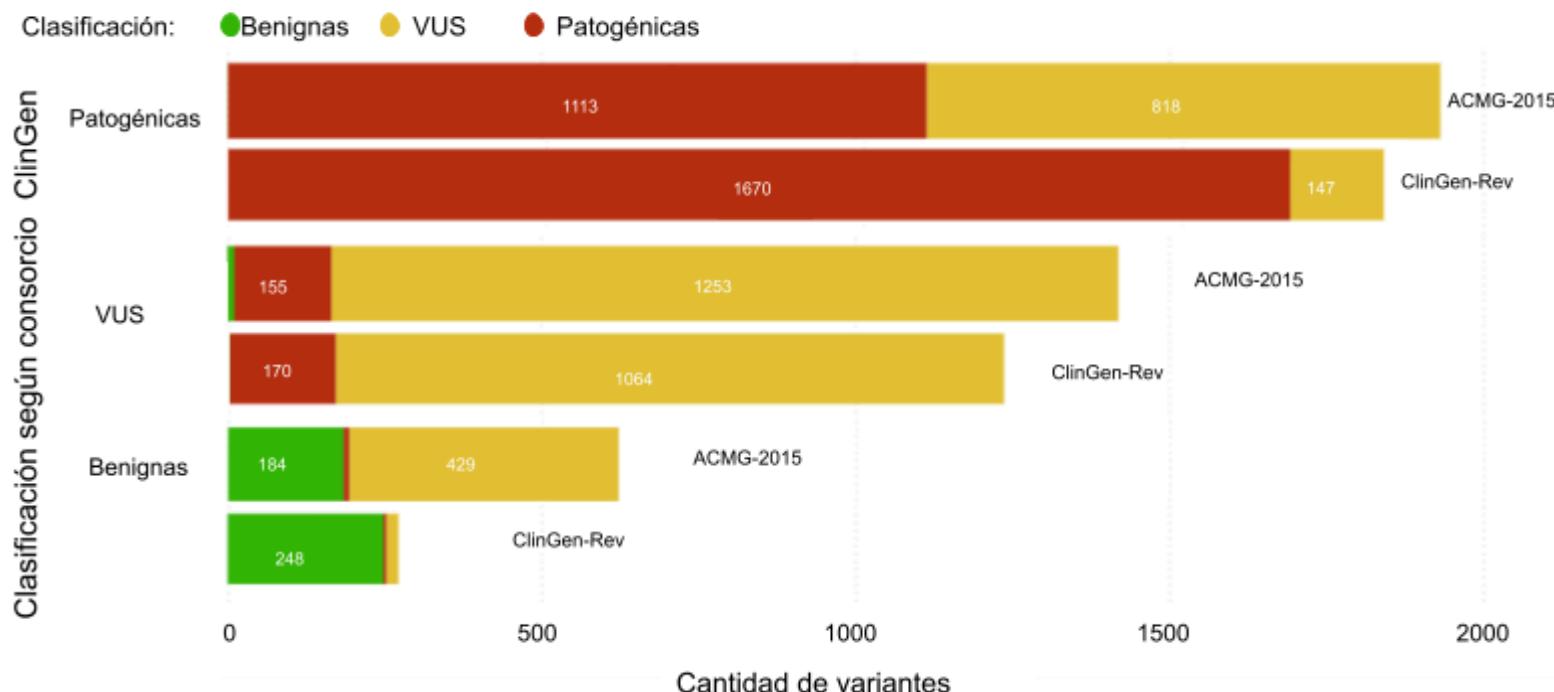


Figura 14: Clasificación de variantes con la etiqueta PM2 en ACMG-2015 y ClinGen-Rev: Se analiza a qué categoría llegan en el pipeline ACMG-2015, las variantes a las que atribuye PM2 ACMG-2015. Del mismo modo con ClinGen-Rev.

ClinGen-Rev entiende que una baja frecuencia o aparición de novo no conlleva automáticamente a que una variable sea patógenica, pues interpreta la rareza como algo común. De este modo, se reduce la fuerza de patogenicidad que confiere a la etiqueta (PM2:Supportive), y se fijan umbrales más estrictos de frecuencia alélica poblacional para poder asignarla, logrando disminuir la cantidad de veces que se aplica un criterio de patogenicidad moderado (PM2). Puede observarse en la figura cómo las variantes marcadas por ClinGen-Rev como B son prácticamente todas reconocidas como tal. Este ajuste a PM2, es otra de las modificaciones que esclarece cómo ClinGen-Rev logra corregir variantes de falso significado incierto a través de reducir el sobreetiquetado.

d) Impacto de etiquetas relacionadas al reporte de variantes (BP6/PP5)

Las etiquetas BP6 y PP5 están relacionadas a la búsqueda de evidencia de una variante en fuentes confiables. Surge el debate de qué es una fuente de datos confiable y si una variante puede ser juzgada del mismo modo, pese a estar en otro contexto genético (o sea otro individuo). Así mismo, se cree que colocar estas etiquetas genera un problema en la doble contabilización de criterios, ya que se asume que la fuente de clasificación confiable utilizó las etiquetas basadas en la evidencia disponible, que es la misma que se usa para su clasificación. Es por esto que BP6 y PP5 dejaron de utilizarse para la clasificación luego de la revisión de criterios de ACMG/AMP por parte del consorcio ClinGen.

Teniendo experiencia con el pipeline ACMG-2015, que coloca estas etiquetas de manera efectiva, se tomó la decisión de mantenerlas en el ClinGen-Rev, pese a no ser aconsejables. La razón parte de pensar que si bien hay etiquetas que se podrían estar colocando dos veces, BP6/PP5 podrían darle a una variante clasificada como incierta el empujón necesario para llegar a patogénica o benigna. Este salto podría ser relevante si la variante tuviera evidencias, que si bien no pueden ser observadas por un anotador automático ya que dependen del caso particular (o sea del paciente y su familia), sí son posibles de ser asignadas debido al análisis de un caso/familia específico, y que si fuera el caso, esto quede registrado en una base de datos confiable en una clasificación particular. En otras palabras, la presencia de la variante clasificada como (L)P o (L)B en una base de datos confiable, cuando el análisis automático resulta en una VUS cercana a la frontera correspondiente, sugiere que la clasificación se logró mediante la observación de evidencia adicional del caso particular inaccessible al método automático, y que por ende la incorporación indirecta de esta información (con la etiqueta correspondiente según ACMG-2015) mejoraría la clasificación de la misma.

En este contexto, el impacto de utilizar este etiquetado adicional se da principalmente en los siguientes casos. Por un lado es muy común dar con variantes

que están muy cerca de ser clasificadas como benignas, pero quedan como inciertas con un score muy bajo. Por otro lado si la variante ha sido confirmada como incierta previamente, que no se ponga BP6/PP5 lo reafirma como tal. En caso contrario si la variante ha sido confirmada como benigna el agregado de BP6 puede llevar a su correcta clasificación. Del mismo modo, por ejemplo, para una variante VUS con un score alto, PP5 podría permitir rescatar una pieza de evidencia clave para marcarla como LP/P, marcando su significancia clínica.

También por este motivo, se decide desdoblar las etiquetas BP6 y PP5 en supportive, strong y moderate, para poder juzgar cada caso de un modo más adecuado, la fuente utilizada, influyendo con distintos valores en el score final de la variante.

Por otro lado, se entiende que BP6 y PP5 pueden tomar en cuenta niveles de evidencia que escapan al análisis computacional, pero hacen a la correcta categorización. Estos son, por ejemplo BS3 y PS3, asociados a ensayos funcionales que pueden haber sido reportados en la literatura.

La *figura 15* muestra en sus barras la clasificación que alcanzan las 3,133 variantes curadas que se les asigna BP6/PP5 según nuestra versión de ClinGen-Rev y son correctamente clasificadas por el mismo (coinciden con ClinGen). Los colores muestran la clasificación que alcanzaría si se modifica ClinGen-Rev para que no coloque BP6 ni PP5 (lo llamamos ClinGen-Rev- $\{\text{BP6/PP5}\}$). En primer lugar se observa que hay muy pocas variantes (159) en categoría incierta que tengan estas etiquetas, en línea con lo esperado, y estas no se modifican al excluir la información de BP6/PP5.

Clasificación diferencial de variantes por ClinGen-Rev con y sin la contabilización de BP6 y PP5

Clasificación sin sumar BP6 o PP5: ● Benignas ● VUS ● Patogénicas

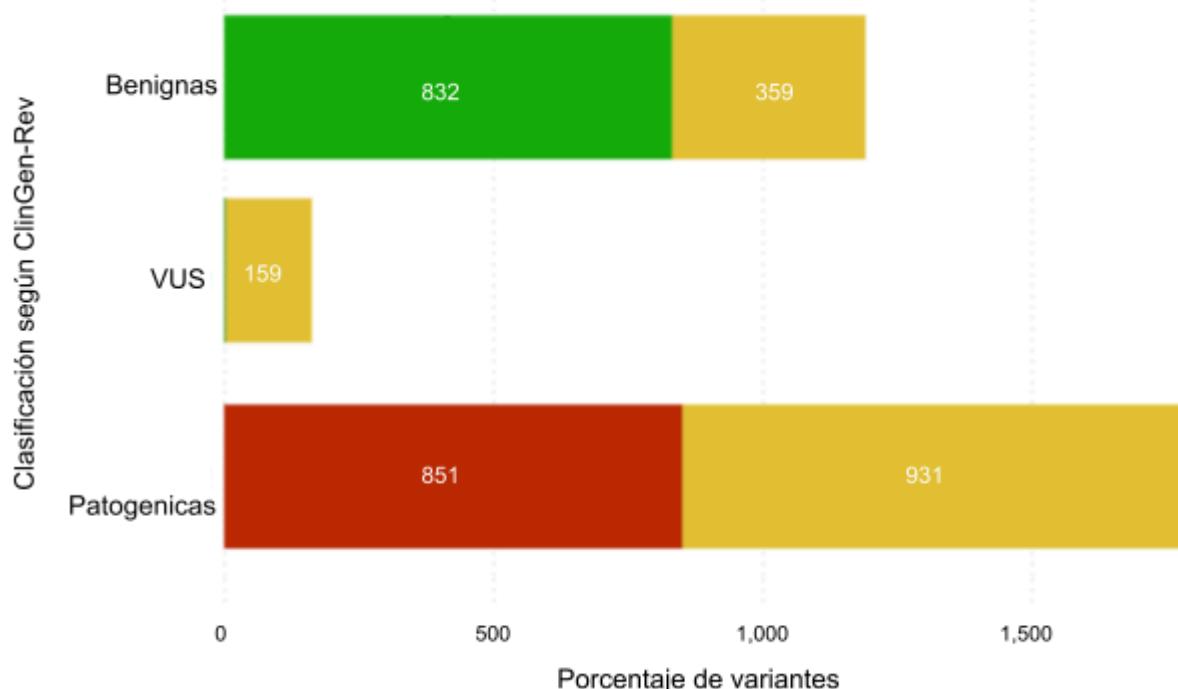


Figura 15: clasificación diferencial de variantes con y sin BP6 o PP5: Se analizan las variantes a las que ClinGen-Rev les coloca BP6 o PP5. A estas, se las clasifica nuevamente por ClinGen-Rev modificado para no sumar BP6/PP5 en el score. En las filas se muestran la categoría alcanzada en ClinGen-Rev y en colores aquellas de ClinGen-Rev-{BP6/PP5}.

Lo relevante de nuestro argumento, es que, en la figura se observa que son 359 las variantes que pierden su clasificación como benignas, lo que representa el 27% de las mismas, y 931 las que dejan de ser patogénicas, casi un 47% del total de patogénicas marcadas por ClinGen-Rev (1,979 variantes). En estas variantes, BP6/PP5 añade el puntaje necesario para cambiar de categoría de clasificación, logrando así la clasificación correcta. Esto se sustenta con la evidencia extraída de bases de datos específicas.

En resumen, los algoritmos de anotación automática tienen muchas limitaciones a la hora de colocar determinadas etiquetas y tomar decisiones críticas por fuera del programa. En este contexto, son muchas las variantes que ya han sido analizadas, curadas manualmente y registradas en repositorios, por lo que es posible asignar BP6/PP5. Los resultados obtenidos en esta sección demuestran que

estas etiquetas están ayudando a un aumento significativo en la exactitud de la clasificación, en línea con el 90% de concordancia con la curación manual del consorcio ClinGen.

III) Impacto de las modificaciones en la priorización de variantes

Tras analizar por ambos pipelines un conjunto de 6 VCFs reales (422,786 variantes) obtenidos de proyectos de secuenciación de pacientes³⁵, se comparan cuantitativamente los resultados obtenidos. Con ambos criterios la mayor parte de las variantes resultan benignas o probablemente benignas. Este es un resultado esperado, dado que atribuimos estas a una variabilidad poblacional, y solo unas pocas variantes LP o P se encuentran presentes en cada muestra. Además, como ya se ha mencionado, hay variantes que son comunes en una población. Es por eso que se espera que entre pacientes de una misma comunidad existan variantes repetidas. En la *tabla 4* se muestran los resultados obtenidos para las variantes analizadas **quitando repetidas** (201,340 variantes) con ambos criterios y el porcentaje de coincidencia entre ambos para cada categoría.

Cuantificación	Benignas	Probablemente Benignas	Incias	Probablemente Patogénicas	Patogénicas
ACMG 2015	172967	24831	3498	33	11
ClinGen-Rev	186061	13279	1986	13	1
% Coincidencia		92.03%			

Tabla 4: datos extraídos de la clasificación diferencia de las variantes

La figura 16 muestra en cada barra horizontal el total de variantes clasificadas por ACMG-2015 en cada una de las tres categorías, y luego en color el porcentaje de estas que resulta de la clasificación de ClinGen-Rev. La figura permite observar que prácticamente el total (99.74%) de las variantes clasificadas como Benignas por ACMG-2015, también lo son según ClinGen-Rev. Lo que resulta ya más interesante es que, del total de las variantes clasificadas como de significado incierto en

³⁵ Muestras elegidas al azar de base de datos de pacientes históricos analizados por el grupo de investigación donde se desarrolló esta tesis.

ACMG-2015, un 58.98% son pasadas a benignas por ClinGen-Rev. En el caso de las clasificadas como patogénicas por ACMG-2015, también un 68.18% de variantes son “bajadas” de categoría a la de VUS.

Clasificación diferencial alcanzada por cada criterio de etiquetado ACMG-2015 vs ClinGen-Rev

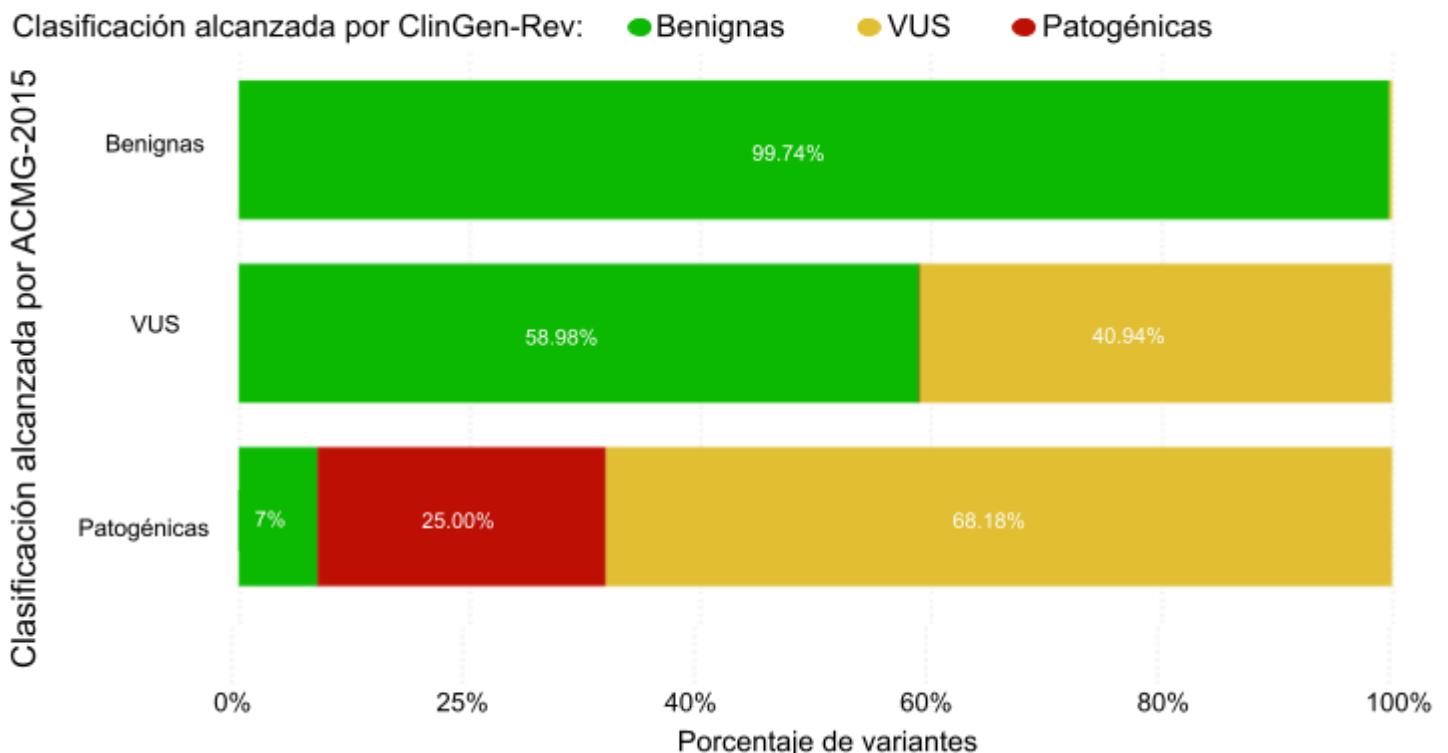


Figura 16, Diferencias en la clasificación final alcanzada por variantes analizadas provenientes de una muestra de un paciente.

En la figura llaman la atención 3 variantes (el 7%) que son clasificadas como P o LP según ACMG-2015, y son pasadas a la categoría B o LB en ClinGen-Rev. En la tabla 5 se muestran las diferentes etiquetas que coloca cada protocolo, junto a datos de frecuencia poblacional y reportes previos de las variantes correspondientes en ClinVar (si los hubiera). Se observa, en estas variantes, que el cambio de la categoría está relacionado a la adición de las etiquetas de frecuencia poblacional BS1 y BA1 por parte de ClinGen-Rev, que como ya mencionamos, aunque la evidencia de la frecuencia poblacional no ha cambiado, la toma de umbrales

diferenciales para cada gen le permite colocar la etiqueta en estos casos. Es así como estas variantes obtienen el score necesario para dar el salto de (L)P a (L)B.

<i>ID</i>	<i>Gen y tipo de variante</i>	<i>ACMG-2015</i>	<i>ClinGen-Rev</i>	<i>ClinVar</i>	<i>Frecuencia poblacional</i>
<i>rs112966887</i>	<i>Gen : PABPC1 Frameshift Variant</i>	<i>PVS1</i>	<i>PVS1:moderate BS1:strong</i>	<i>No reportada</i>	<i>delT=0.154114 (ExAC)</i>
<i>rs142985461</i>	<i>Gen : PABPC1 Stop Gained</i>	<i>PVS1 PP3</i>	<i>PVS1:moderate BA1:standalone</i>	<i>No reportada</i>	<i>A=0.06805 (ALFA)</i>
<i>rs768161285</i>	<i>Gen : EVL Frameshift Variant</i>	<i>PVS1</i>	<i>PVS1:Moderate, BS1:Strong</i>	<i>No reportada</i>	<i>delA=0.00888 (14KJPN)</i>

Tabla 5: Datos relevantes de variantes clasificadas como Pathogenic por ACMG-2015 y Benign por ClinGen-Rev.

En resumen, con el conocimiento de las principales etiquetas modificadas y sabiendo el incremento de la presión en la clasificación del nuevo protocolo, se puede decir que la categorización de las variantes de la tabla 5 es más certera con ClinGen-Rev. Nuevamente da la impresión de que, ClinGen-Rev vuelve las variantes un poco más benignas, reduciendo notoriamente la cantidad de variantes que son marcadas como patogénicas y VUS, en contraste con ACMG-2015. Es así como la diferencia del 7.97% entre ambos protocolos debe ser interpretado como el impacto en la priorización de variantes a realizar al finalizar con la clasificación (semi) automatizada.

DISCUSIONES Y CONCLUSIONES

Este trabajo se propuso analizar de manera integral el impacto de las modificaciones de ClinGen-Rev sobre la clasificación de variantes, en comparación con la clasificación original de ACMG-2015, evaluando el efecto de las mismas en el proceso de priorización de variantes en el contexto del diagnóstico molecular genómico de EPoFs. Para esto, se desarrollaron dos protocolos bioinformáticos, uno acorde a cada criterio, a través del cual se pudieron analizar muestras y variantes genéticas, resolviendo sus respectivas significancias clínicas.

Primero se clasificaron con ambos métodos variantes curadas extraídas del repositorio de ClinGen, lo que permitió tener una muestra más diversa de variantes y tener una referencia contra la que verificar la exactitud de los resultados obtenidos. ACMG-2015 obtiene una coincidencia del 67% contra los veredictos provistos por el consorcio ClinGen, encontrándose nuevamente la mayor parte de la diferencia compuesta de variantes categorizadas erróneamente como VUS por ACMG-2015. En cambio, ClinGen-Rev logra una remarcable mejora llegando a un 90% de concordancia, pudiendo clasificar correctamente el 99.5% de las variantes benignas, y marcando correctamente el 92% de las variantes patogénicas, es decir, sin aumentar la cantidad de falsos positivos.

Hasta acá hemos podido ver cómo las modificaciones propuestas por ClinGen-Rev logran reducir de manera significativa el potencial patogénico de las variantes. Es interesante notar, que muchas de las decisiones que cambian se toman frente a una misma evidencia, siendo más estrictos en cuanto a cómo atribuir cada criterio, y al peso que se le otorga en la clasificación final. Pensándolo desde un principio de parsimonia, las EPoFs son de naturaleza poco frecuente, por lo que si uno elige al azar una variante cualquiera es altamente probable que no sea responsable de la enfermedad, y por ende patogénica. ClinGen-Rev usa este razonamiento para permitirle fácilmente a una variante llegar a la categoría de benigna, y exigir en cambio, a las patogénicas tener las condiciones necesarias para ser categorizadas como tales.

A continuación, se resumen algunas de las modificaciones de ClinGen-Rev que causan más impacto, obtenidos trabajando con la base de datos curada de variantes (ClinGen).

- Evidencias relacionadas con la frecuencia alélica poblacional (BA1/BS1)

Se realizó un análisis de la clasificación diferencial de variantes entre AMG-2015 y ClinGen-Rev a las que ClinGen-Rev asigna BA1 o BS1. ClinGen-Rev introduce nuevas bases de datos más actualizadas y fija diferencialmente para cada gen los umbrales sobre los cuales la variante se considera de alta frecuencia poblacional para la asignación de la etiqueta. Además, BS1 es dividida en supportive y strong, lo que permite una evaluación más individualizada de los casos. Como resultado, con la modificación de estas dos etiquetas, se logra categorizar correctamente el 22% de las variantes que ACMG-2015 marca como VUS pese a ser de naturaleza benigna.

Tal es el caso de la variante NM_000329.3(RPE65):c.496-4G>A (ID: rs138146176) variante en región no codificante, clasificada como VUS según ACMG-2015 y como LB en ClinGen-Rev. La variante tiene una frecuencia alélica poblacional de 0.0014 según la base de datos para una población africana (gnomAD v.2.1.1), por lo que ACMG-2015, que usa un umbral estándar de 0.005 no coloca la etiqueta BS1. En cambio ClinGen-Rev tiene un umbral específico para el gen RPE65 que considera que si la frecuencia es mayor a 0.0008 puede colocarse BS1. Es así como la misma variante utilizando la misma información puede ser categorizada correctamente utilizando los criterios provistos por ClinGen-Rev.

- Revisión de etiquetas relacionadas con un efecto molecular patogénico (PVS1/PM1)

Similar a lo realizado para las etiquetas de frecuencia poblacional, se exploran los criterios PVS1 y PM1, etiquetas sumamente relevantes a la hora de identificar

variantes como patogénicas o probablemente patogénicas. Se observa primero cómo ambos criterios tienen una alta coincidencia en aquellas variantes marcadas como P/LP. Sin embargo, lo notable es que ClinGen-Rev logra rescatar 11 variantes, erróneamente categorizadas como VUS en ACMG-2015, mediante la granularización de PVS1 (marcando Very strong, strong, moderate y supportive), categorizando de manera más exacta cada variante.

A modo de ejemplo, la variante NM_004360.5(CDH1):c.2T>G (p.Met1Arg) (ID: rs1555509623) altera el codón de inicio de la secuencia codificante del gen CDH1, en un gen donde la pérdida de función es un mecanismo conocido de desarrollo de la enfermedad. Ambos criterios colocan las mismas etiquetas a excepción de PVS1; acorde a los criterios de ACMG-2015 no corresponde colocar esta etiqueta, por lo que termina teniendo un veredicto incierto. En cambio, ClinGen-Rev tiene incorporado colocar la variante PVS1:Supportive cuando se predice que el cambio de codón de inicio dará una interrupción en el producto génico. Gracias a esta modificación, ClinGen-Rev clasificó correctamente a la variante como probablemente patogénica.

- Problemáticas asociadas al modo de en qué orden asignar y combinar evidencia

La mayoría de las veces las variantes no tienen una única etiqueta que define de manera inequívoca su probabilidad de patogenicidad, o sea su clasificación definitiva, y lo más frecuente es llegar al veredicto a través de la combinación de varias etiquetas. Es por eso que al momento de pensar el modo de evaluar la evidencia los protocolos prestan atención a cómo se ordena la asignación de etiquetas para hacer el proceso más eficiente y preciso. Una de las primeras consideraciones es que no tiene sentido colocar dos etiquetas que representan a una misma evidencia como benigna y patogénica. Tal sería el caso de colocarle a una misma variante, por ejemplo, BP4 y PP3; ya que la predicción bioinformática no puede ser interpretada como benigna y patogénica en simultáneo. En la práctica, esto hace que el algoritmo o coloque una, o la otra (o ninguna) pero nunca las dos.

Las complicaciones que trae el sobre etiquetado pueden observarse al analizar la variante NM_001110792.2(MECP2):c.638C>T (p.Ala213Val), a la cual según el pipeline de ACMG-2015 le corresponden 6 etiquetas [PM1, PP3, BS1, BP1, BP6] categorizándose como LB. En cambio, al utilizar ClinGen-Rev, se colocan solamente dos etiquetas [BP6:Strong, BS1:Strong] quedando rápidamente clasificada como B. Esto se explica dado que ACMG-2015 revisa todas las etiquetas para ver si colocarlas o no, mientras que ClinGen-Rev intenta tener un enfoque de etiquetado minimal.

Dada una variante, ACMG-2015, asigna evidencias de un modo que, si bien no es que esté necesariamente mal la aplicación del criterio, muchas veces al ponderar las diversas etiquetas se termina por “neutralizar” la variante (al colocar etiquetas que por un lado restan pero por otro suman evidencia a favor de su patogenicidad), terminando en muchos casos la variante categorizada como VUS. ClinGen-Rev hace un esfuerzo que ACMG-2015 no hace en la definición de qué “pares” (o grupos) de etiquetas se contrastan juntas, poniendo énfasis en evitar etiquetas que usen una misma evidencia, esquivando su duplicación y buscando reducir el consecuente sobreetiquetado como se vio con el caso de PM2. Este protocolo intenta imitar una curación manual, en el que el analista observa las evidencias que presenta cada variante y coloca las etiquetas necesarias para llegar al diagnóstico.

- Impacto de etiquetas relacionadas al reporte de variantes (BP6/PP5)

Pese a que las recomendaciones de ClinGen desaconsejan su uso, nuestra versión de ClinGen-Rev propone continuar con la colocación de los criterios BP6 y PP5, relacionados a variantes ya reportadas y almacenadas en repositorios. La construcción de un tercer pipeline que es exactamente igual a ClinGen-Rev, a excepción de que no coloca BP6 y PP5 (en estricto cumplimiento con las recomendaciones), confirma que la decisión nuestra fue acertada. Los resultados de clasificar variantes genéticas extraídas del repositorio de ClinGen con y sin estas etiquetas muestran que un 27% de las variantes B o LB y un 47% de las P o LP pierden su categoría cuando no se aplican BP6/PP5, pasando a VUS.

A modo de ejemplo la variante NM_004360.5(CDH1):c.731A>G (p.Asp244Gly) (ID: rs1064794231), se categoriza como VUS usando el pipeline ClinGen-Rev-{BP6/PP5} y como LB usando ClinGen-Rev. Esta variante missense es responsable del cambio de aminoácidos de ácido aspartico a glicina en la proteína codificada por el gen CDH1. Ambos pipelines le atribuyen la etiqueta PM2:supporting, dado que tiene una frecuencia poblacional extremadamente baja de 0.000007 (según GnomAD). Los predictores bioinformáticos como REVEL no dan resultados relevantes, no hay predicciones sobre el efecto molecular, ni se sabe cómo afecta al fenotipo, y tampoco hay información relevante sobre la cigosidad. Sin embargo, numerosos ensayos funcionales fueron hechos sobre la variante reportándola como LB y se la encuentra en la base de datos curadas del consorcio ClinGen como B. De este modo, nuestra versión de ClinGen-Rev toma una decisión “inteligente” al asignarle BP6, entendiendo que aunque computacionalmente no se tenga más evidencia, hay información que justifica la categorización de la variante como LB.

Los resultados obtenidos en esta sección demuestran que BP6 y PP5 están ayudando a la correcta clasificación de las variantes. Sin embargo, sigue abierta la disputa entre quienes creen que esto fomenta la doble contabilización de etiquetas y quienes plantean que los contextos genéticos no son extrapolables, desestimando la validez del criterio. Respecto al primer argumento, este trabajo entiende la gran ventaja que conlleva incorporar criterios no evaluables computacionalmente. Además se sabe que si bien se podría estar considerando el mismo criterio dos veces, la etiqueta en realidad engloba un montón de criterios, por lo que si hay duplicación su peso es relativamente bajo al estar “diluido”. Respecto al argumento de extrapolación de contextos genéticos, no hay que perder en vista el objetivo final del proceso de clasificación automática: resolver un diagnóstico sumamente complejo, de manera rápida y precisa. Claro está que cada caso debe ser analizado de acuerdo a sus particularidades, pero resultaría imposible realizar, por ejemplo, un ensayo funcional de cada una de las variantes de categoría incierta, por lo que se piensa que proveer una clasificación automatizada basada en bases de datos

confiables, que reportan este tipos de experimentos, y luego realizar la priorización manual, es un camino más cercano al óptimo para la resolución del caso.

Por último se comparan 6 VCFs provenientes de muestras reales, para terminar de comprender el efecto que los criterios de ClinGen-Rev tienen sobre el trabajo real en el contexto del uso de genómica para el diagnóstico de EPoFs. Se obtuvieron por ambos pipelines resultados esperados para conjuntos de variantes biológicas, con más del 95% de las mismas clasificadas como B o LB y un par (de 1 a 5) variantes P o LP por muestra. Los métodos coinciden en la categorización asignada a las variantes en un 92.03% de las veces. Además, se ve que la mayor parte de variantes pertenecientes al 7.97% de diferencia radica en variantes que ACMG-2015 clasifica como VUS, mientras que ClinGen-Rev las categoriza como LB o B, siguiendo uno de los conceptos claves de las modificaciones planteadas por ClinGen-Rev.

Hablar de un impacto del 7.97% de diferencia entre las categorizaciones alcanzadas por cada criterio puede parecer quizá poco. Sin embargo, implica una reducción de un 40% en variantes VUS y de un 68% de patogénicas, resultado muy valorable a la hora de tener que realizar la priorización, agilizando el proceso manual de curado de las mismas, y facilitando la llegada al diagnóstico final.

Palabras finales y perspectivas a futuro

Hasta acá, los cambios observados en ClinGen-Rev refieren a criterios que se logran perfeccionar gracias a la experiencia de utilizar ACMG-2015. Pese a que la exactitud que este último obtiene contra el consorcio ClinGen (67%) parece baja, desde su construcción, se han resuelto satisfactoriamente numerosos casos pudiendo arribar al tan deseado diagnóstico. Las modificaciones propuestas surgen precisamente de observaciones realizadas durante la priorización de variantes que permiten comprender mejor los patrones propios detrás de una variante patogénica. ClinGen-Rev es el resultado de un proceso iterativo de revisión de decisiones, ajuste de umbrales y cambios en el puntaje que aporta cada etiqueta.

En paralelo, desde el establecimiento de ACMG/AMP se ha fomentado el reporte de variantes y estudios realizados sobre las mismas. Conforme la cantidad

de datos de secuenciación e información sobre las variantes crecen, las bases de datos se perfeccionan, permitiendo una mejor toma de decisiones. Es así como se construye el círculo virtuoso del conocimiento sobre las variantes, donde la información es sucesivamente validada y actualizada, y “resolver” un caso (es decir, identificar la variante clasificada como (probablemente) patogénica responsable) permitirá en el futuro dilucidar otro.

Una interpretación incorrecta de los resultados presentados podría sugerir que en el futuro los protocolos no marcarían más variantes con categoría de significancia incierta, pues que se llamen “inciertas” no implica que no exista una certeza sobre sus características. Esta categoría debe entenderse conformada por dos grupos de variantes. Por un lado, aquel compuesto por variantes que se asignan a esta categoría por no poseerse la evidencia necesaria. Por otro lado, el segundo, aquellas variantes en las que se han realizado numeroso ensayos, se tiene información y se decide determinarla incierta, por ser de compleja interpretación. Por ejemplo aquellas variantes que se comportan distinto en fenotipos sanos y enfermos. Viendo la evolución entre ACMG-2015 y ClinGen-Rev, es cierto que ha sido posible clasificar menos variantes como VUS (empujándolas a B o P), estas deben entenderse como variantes del primer grupo, que se recategorizan gracias a una mejor interpretación de la evidencia disponible.

Finalmente, como comentario personal quiero volcar una experiencia o anécdota relacionada con el desarrollo de esta tesis. Mientras la desarrollaba y escribía no fueron pocas las personas que me preguntaban si no se podía resolver la clasificación de variantes , y el diagnóstico de EPoFs con inteligencia artificial “y listo”. Si bien es posible el desarrollo de modelos de esta índole para predecir el veredicto de las variantes, en primer lugar hay que destacar que la cantidad de datos requeridos para el entrenamiento que serían necesarios aún no se encuentra disponible, dado que las muestras que se manejan están sesgadas a personas con patologías. Además, la mayor parte de las secuencias humanas disponibles refieren únicamente a regiones del exoma. El genoma completo con sus regiones no codificantes sigue siendo en gran parte un enigma para los investigadores, pudiendo esconder la clave para la correcta comprensión de muchas variantes responsables de enfermedades.

Futuros protocolos buscarán sucesivamente converger hacia un modo más preciso y eficiente de priorizar variantes, en función de los constantes avances en las tecnologías de secuenciación y manejo de datos. Si bien se debe apuntar a seguir desarrollando metodologías acordes a las herramientas disponibles, no se debe perder de vista el objetivo diagnóstico final; después de todo, el trabajo realizado es en busca del bienestar de los pacientes. Deben remarcarse la importancia de las decisiones realizadas por profesionales de la salud y bioinformáticos capaces de entender la totalidad de cada caso con una empatía y criterio que un algoritmo jamás podrá.

- **Anexo 1: Ejemplo clasificación de variantes**

Pathogenic Variant: NM_000277.2(PAH):c.1A>G (p.Met1Val) ³⁶

- PS3:
<3% de actividad enzimática de PAH como porcentaje del wild type. ³⁷

- PM2:
gnomAD MAF=0.00002

- PM3:
Detectado en trans con c.1315+1G>A (conocido como patogénico) y también en estado homocigoto.

Un paciente probando era homocigoto para esta mutación (transición de A a G (met-val) en el codón 1). En otros probandos, la mutación en el codón 1 se heredó una vez con la mutación de la unión de empalme en el exón 12 (c.1315+1G>A) y se heredó dos veces con una mutación en el haplotipo 1.³⁸

- PP 4 Moderate:
Observado en pacientes con PKU. Se descartaron los trastornos de BH4.

El análisis de haplotipos RFLP y mutaciones reveló una nueva mutación, una transición de A a G (met-val) en el codón 1 (el codón de inicio de la traducción). Ocurrió en 5 de los 18 cromosomas mutantes. Un paciente probando homocigoto para esta mutación tenía el fenotipo PKU. Se realizaron las pruebas bioquímicas adecuadas para descartar trastornos de la homeostasis de tetrahidrobiopteterina.³⁹

No llega a PVS1 por ser una variante en un codón de iniciación

Esta variante cumple con los criterios para ser clasificada como patogénica para fenilcetonuria (PKU) de manera autosómica recesiva según los criterios ACMG/AMP aplicados según lo especificado por el Panel de Expertos de PAH: (PM2, PM3, PM4 Moderate, PS3).

Likely pathogenic Variant: NM_000277.2(PAH):c.581T>C (p.Leu194Pro) ⁴⁰

- PP4 Moderate:
Detectado en 3 pacientes (1 HPA, 1 PKU). La deficiencia de BH4 fue excluida en 2 pacientes.

L194P detectada en 2 cromosomas de pacientes con PKU de 4 centros en Gran Bretaña.

⁴¹El Paciente 28 fue detectado con L194L (PKU Leve). Se excluyó un defecto en la síntesis o reciclaje de tetrahidrobiopteterina mediante el análisis de pterinas urinarias y la actividad de dihidropteridina reductasa en eritrocitos.⁴²

- PP3
Predicho como perjudicial en SIFT, Polyphen-2, MutationTaster. REVEL=0.899

³⁶ <https://erepo.genome.network/evrepo/ui/classification/89f04437-ed5d-4735-8c4a-a9b1d91d10ea>

³⁷ <https://pubmed.ncbi.nlm.nih.gov/9450897/>

³⁸ <https://www.ncbi.nlm.nih.gov/pubmed/?term=2574002>

³⁹ <https://www.ncbi.nlm.nih.gov/pubmed/?term=2574002>

⁴⁰ <https://erepo.genome.network/evrepo/ui/classification/434f1539-4967-4ff7-abc2-e2ff3ca9ecbe>

⁴¹ <https://pubmed.ncbi.nlm.nih.gov/9012412/>

⁴² <https://pubmed.ncbi.nlm.nih.gov/16601866/>

- PM3 strong

Detectado en trans con V245A y R261X, ambos patogénicos.

Genotipo del Paciente 28: L194P/R261X (VarID610, Patogénico). Todas las mutaciones identificadas fueron confirmadas mediante el análisis del ADN parental, lo que también nos permitió seguir la segregación de las mutaciones.⁴³

- PM2

Ausente en 1000G, ESP. gnomAD MAF: 0.00004.

En resumen, esta variante cumple con los criterios para ser clasificada como probablemente patogénica para fenilcetonuria de manera autosómica recesiva según los criterios ACMG/AMP aplicados según lo especificado por el Panel de Expertos de PAH: (PM2, PP3, PM3_strong, PP4 Moderate).

VUS : Variant: NM_001754.4:c.1395C>A (p.Asn465Lys)⁴⁴

- PM2

The variant is absent from all population databases.

- BP4 Moderate

REVEL: 0.108 < 0.15

Esta variante cumple con los criterios para ser clasificada como de significado incierto para fenilcetonuria (PKU) de manera autosómica recesiva según los criterios ACMG/AMP aplicados según lo especificado por el Panel de Expertos de PAH: (PM2, BP4 Moderate).

Likely Benign Variant: NM_000277.1:c.772C>T⁴⁵

- BP7:

No se predice un efecto perjudicial.

- BS1:

Mayor que las guías específicas de PAH de AF-0.0002 (0.02%).

Esta variante cumple con los criterios para ser clasificada como probablemente benigna para fenilcetonuria (PKU) de manera autosómica recesiva según los criterios ACMG/AMP aplicados según lo especificado por el Panel de Expertos de PAH: (BS1, BP7).

Benign Variant: NM_000277.2(PAH):c.707-7A>T⁴⁶

- BA1:

Highest MAF=0.10514 in 1000G. 35 homozygotes in ExAC

- BP4:

HSF: No significant splicing motif alteration detected. This mutation has probably no impact on splicing. CADD=1.163344

Esta variante cumple con los criterios para ser clasificada como benigna para fenilcetonuria (PKU) de manera autosómica recesiva según los criterios ACMG/AMP aplicados según lo especificado por el Panel de Expertos de PAH: (BA1, BP4).

⁴³ <https://pubmed.ncbi.nlm.nih.gov/16601866/>

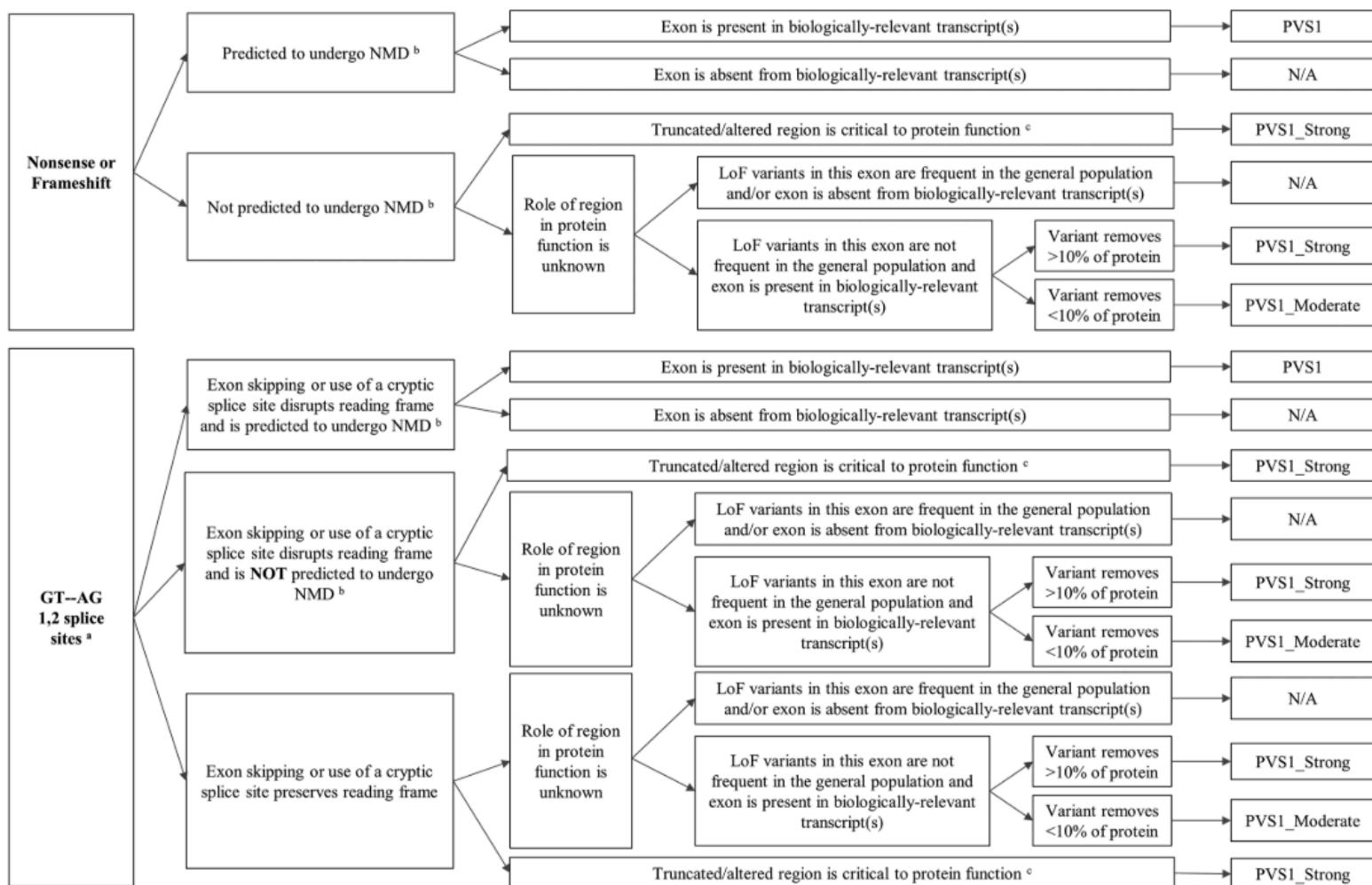
⁴⁴ <https://erepo.genome.network/evrepo/ui/classification/49879cb4-10b0-44b7-a695-0af82d9e37d5>

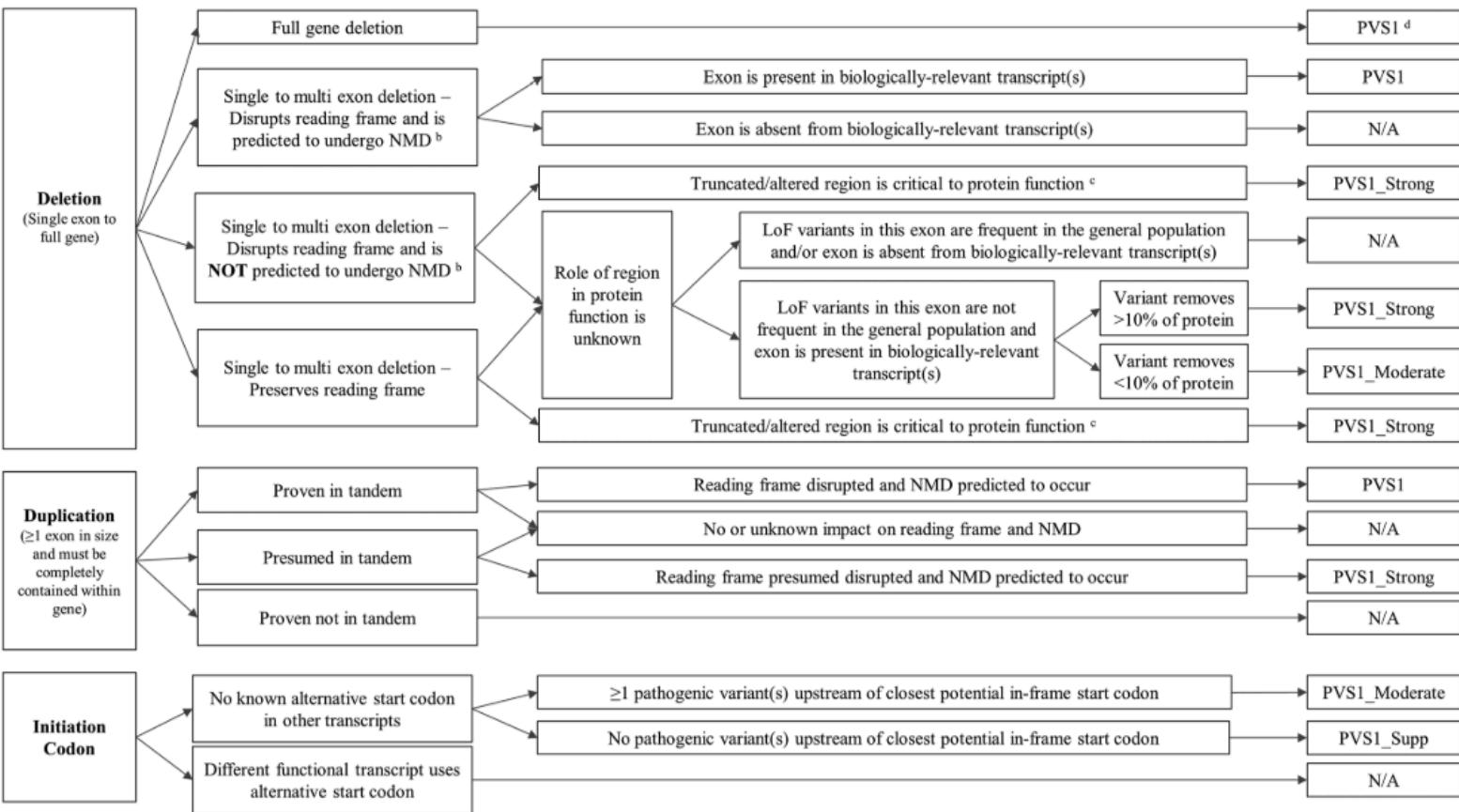
⁴⁵ <https://erepo.genome.network/evrepo/ui/classification/254ed141-a219-404a-8c94-1fd2aceb8103>

⁴⁶ <https://erepo.clinicalgenome.org/evrepo/ui/classification/ef426544-d234-401a-81e4-5e2f6a9ded0a>

● Anexo 2: recomendaciones de ACMG/ClinGen para PVS1

Figura 1 extraída de - Abou Tayoun, A. N., Pesaran, T., DiStefano, M. T., Oza, A., Rehm, H. L., ... Biesecker, L. G. (2018). Recomendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Human Mutation. doi:10.1002/humu.23626





- **Anexo 3:** Gráficos de confusión

Gráfico de confusión por cada categoría de clasificación para ClinGen-Rev

Clasificación ClinGen		Veredicto ClinGen-Rev (predicho)				Total de aciertos		
		Benignas		VUS		Patogénicas		
V	6	1184	363	1258	144	1795	527	4237
	3453	121	2993	150	2584	241		
F	V	F	V	F	V	F	Errores	Aciertos

Gráfico de confusión por cada categoría de clasificación para ACMG-2015

		Clasificación ClinGen			
		V	V	V	V
F	V	524	663	216	1404
	F	3243	334	1787	1357
		F	V	F	V
Veredicto ACMG-2015 (predicho)		Benignas	VUS	Patogénicas	Total de aciertos
Errores		1593	3171		

BIBLIOGRAFÍA

- (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. doi:10.1038/35057062
- Abou Tayoun, A. N., Pesaran, T., DiStefano, M. T., Oza, A., Rehm, H. L., ... Biesecker, L. G. (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Human Mutation*. doi:10.1002/humu.23626
- Amendola, L. M., Dorschner, M. O., Robertson, P. D., Salama, J. S., Hart, R., Shirts, B. H., ... Kim, D. S. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Research*, 25(3), 305–315. doi:10.1101/gr.183483.114
- Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., ... Rehm, H. L. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *The American Journal of Human Genetics*, 99(1), 247. doi:10.1016/j.ajhg.2016.03.024
- Biesecker, L. G., & Harrison, S. M. (2018). The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *GENETICS in MEDICINE*. doi:10.1038/gim.2018.42
- Brnich, S. E., Abou Tayoun, A. N., Couch, F. J., Cutting, G. R., Greenblatt, M. S., ... Berg, J. S. (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Medicine*, 12(1). doi:10.1186/s13073-019-0690-2
- Bruford, E. A., Braschi, B., Denny, P., Jones, T. E. M., Seal, R. L., & Tweedie, S. (2020). Guidelines for human gene nomenclature. *Nature Genetics*, 52(8), 754–758. doi:10.1038/s41588-020-0669-3
- Buda, Guadalupe. (2017). Aplicación de tecnologías de secuenciación masiva como método de detección de mutaciones para el diagnóstico de enfermedades poco frecuentes. (Tesis de Grado. Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales.). Recuperado de http://hdl.handle.net/20.500.12110/seminario_nBIO001602_Buda

- Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10), 1932–1941. doi:10.1016/j.bbadi.2014.06.015
- Documentación en: <https://matplotlib.org/>
- Documentación en <https://vcfpy.readthedocs.io/en/stable/>
- Documentación en: <https://pandas.pydata.org/docs/>
- Eleanor G. Seaby, Reuben J. Pengelly, Sarah Ennis, Exome sequencing explained: a practical guide to its clinical application, *Briefings in Functional Genomics*, Volume 15, Issue 5, September 2016, Pages 374–384, <https://doi.org/10.1093/bfgp/elv054>
- Ghosh, R., Harrison, S. M., Rehm, H. L., Plon, S. E., & Biesecker, L. G. (2018). Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Human Mutation*, 39(11), 1525–1530. doi:10.1002/humu.23642
- Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., ... Reid, J. G. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. doi:10.1038/nature15393
- Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M; Genome Aggregation Database Consortium; Rehm HL, MacArthur DG, O'Donnell-Luria A. Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat*. 2022 Aug;43(8):1012-1030. doi: 10.1002/humu.24309. Epub 2021 Dec 16. PMID: 34859531; PMCID: PMC9160216.
- Hamosh, A. (2004). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue), D514–D517. doi:10.1093/nar/gki033
- Harrison, S. M., Biesecker, L. G., & Rehm, H. L. (2019). Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines. *Current Protocols in Human Genetics*, 103(1). doi:10.1002/cphg.93
- Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., ... Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the

Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*, 99(4), 877–885. doi:10.1016/j.ajhg.2016.08.016

- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., ... Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. doi:10.1016/j.cell.2018.12.015
- Jarvik, G. P., & Browning, B. L. (2016). Consideration of Cosegregation in the Pathogenicity Classification of Genomic Variants. *The American Journal of Human Genetics*, 98(6), 1077–1081. doi:10.1016/j.ajhg.2016.04.003
- Kanavy, D.M., McNulty, S.M., Jairath, M.K. et al. Comparative analysis of functional assay evidence use by ClinGen Variant Curation Expert Panels. *Genome Med* 11, 77 (2019). <https://doi.org/10.1186/s13073-019-0683-1>
- Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Albarca Aguilera, M., Meyer, R., & Massouras, A. (2018). VarSome: The Human Genomic Variant Search Engine. *Bioinformatics*. doi:10.1093/bioinformatics/bty
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., ... Maglott, D. R. (2015). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1), D862–D868. doi:10.1093/nar/gkv1222
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardiissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; Exome Aggregation Consortium. Analysis of protein-coding

genetic variation in 60,706 humans. *Nature*. 2016 Aug 18;536(7616):285-91. doi: 10.1038/nature19057. PMID: 27535533; PMCID: PMC5018207.

- Leslie G. Biesecker, Alicia B. Byrne, Steven M. Harrison. (2023). ClinGen guidance for use of the PP1/BS4 co-segregation and PP4 phenotype specificity criteria for sequence variant pathogenicity classification. *Cell*, VOLUME 111, ISSUE 1, P24-38. DOI:<https://doi.org/10.1016/j.ajhg.2023.11.009>
- Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *The American Journal of Human Genetics*, 100(2), 267–280. doi:10.1016/j.ajhg.2017.01.004
- Mastrorosa, F.K., Miller, D.E. & Eichler, E.E. Applications of long-read sequencing to Mendelian genetics. *Genome Med* 15, 42 (2023). <https://doi.org/10.1186/s13073-023-01194-3>
- Preston, C.G., Wright, M.W., Madhavrao, R. et al. ClinGen Variant Curation Interface: a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines. *Genome Med* 14, 6 (2022). <https://doi.org/10.1186/s13073-021-01004-8>
- Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H., & Inoue, I. (2012). Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *Journal of Human Genetics*, 57(10), 621–632. doi:10.1038/jhg.2012.91
- Referencia del programa: <https://www.microsoft.com/es-es/power-platform/products/power-bi>
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., ... Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5), 405–423. doi:10.1038/gim.2015.30
- Smigielski EM, Sirotnik K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res*. 2000 Jan 1;28(1):352-5. doi: 10.1093/nar/28.1.352. PMID: 10592272; PMCID: PMC102496.

- Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., & Biesecker, L. G. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *GENETICS in MEDICINE*. doi:10.1038/gim.2017.210
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073 (2010).
- Variantes curadas por ClinGen extraídas de su repositorio digital : <https://erepo.clinicalgenome.org/evrepo/>
- Vikas Pejaver, Alicia B. Byrne, Bing-Jian Feng. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *CELL*. <https://doi.org/10.1016/j.ajhg.2022.10.013>
- Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16), e164–e164. doi:10.1093/nar/gkq603
- Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K., Ing, A. Y., ... Ware, J. S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine*, 19(10), 1151–1158. doi:10.1038/gim.2017.26
- Zhang, J., Yao, Y., He, H., & Shen, J. (2020). Clinical interpretation of sequence variants. *Current Protocols in Human Genetics*, 106,e98. doi: 10.1002/cphg.98

