Universidad Nacional Autónoma de México

Centro de Ciencias Genómicas

Estudio de la distribución genómica de las variantes asociadas a enfermedades mendelianas humanas

Materia: Genómica Humana

Profesor: Araxi Urrutia

Asistente: Alín Acuña Alonzo

Integrantes:

Daianna González Padilla

José Rodelmar Ocampo Luna

Rodrigo Daniel Hernández Barrera

Fecha: 22/05/2022

Abstract

Las enfermedades mendelianas conforman una parte importante de los padecimientos humanos y por ello resulta relevante examinar factores de interés como la distribución genómica de las variantes asociadas a los genes implicados en estas enfermedades. En este trabajó se encontró que el 89.93% de las variantes asociadas a enfermedades mendelianas humanas residen en regiones exónicas y el 89.2% en regiones codificantes. Por otra parte, no se encontraron variantes en regiones intergénicas ni en el cromosoma Y, y en más del 70% de los genes tampoco fueron encontradas.

Introducción

Observada inicialmente por Gregor Mendel en el siglo XIX, la herencia mendeliana conforma un tipo de herencia biológica en organismos de reproducción sexual que postula tres grandes generalizaciones sobre la herencia de los gametos parentales a la progenie: la dominancia, la segregación y distribución independiente entre ellos. Estas leyes dependen de eventos como la recombinación del genoma de los progenitores del individuo, que deriva en genes que determinan un fenotipo similar al de los padres [6].

La herencia mendeliana de ciertos fenotipos o enfermedades, está dada por una relación de dominancia entre los alelos de organismos diploides, en donde alelos dominantes tienden a expresarse “por encima” de los recesivos. En este sentido, una enfermedad mendeliana puede heredarse de forma dominante o recesiva, lo que puede reflejarse en apariciones características entre los miembros de una misma familia. La herencia de estas enfermedades también puede distinguirse entre aquellas que son autosómicas y aquellas ligadas al sexo del individuo [5].

Pese al componente genético de muchas enfermedades complejas y multifactoriales como la diabetes y el asma, por definición, las enfermedades mendelianas son monogénicas, esto es, que se atribuyen a un solo gen, más no necesariamente a una sola variante. Estas enfermedades resultan de mutaciones que residen en regiones intrónicas, exónicas o intergénicas; cual sea el caso, están asociadas a un gen y a una región cromosómica particular. Estas variantes pueden ser puntuales o abarcar una región más extensa, como en el caso de las inserciones y deleciones (*indels*); además no necesariamente siguen un mismo patrón de expresión, sino que la proteína codificada por un gen mutado puede disminuir o incrementar su actividad, ganar o perder funciones, puede ser inestable, estar truncada o cambiar su estructura, pero todas provocan estragos en el organismo y de ahí que estas variantes se denominen patogénicas.

Los primeros acercamientos a estas mutaciones solían depender de la frecuencia y la fácil distinción de los fenotipos ocasionados por las variantes, esto es, de la penetrancia y la expresividad de los genes, lo que permitía que estos pudieran ser estudiados y secuenciados para identificar las mutaciones en el gen causante [3]. En la actualidad, bases de datos como Ensembl proveen cantidades masivas de información acerca de los elementos génicos humanos y sus anotaciones, además de variantes somáticas, estructurales y puntuales como lo son los *Single Nucleotide Polymorphisms* (SNPs), asociadas a estos genes o *loci* de interés; permite también clasificar a estas variantes por su relevancia clínica, facilitando la distinción de aquellas patogénicas de las benignas, y por su consecuencia y ubicación en el genoma del organismo, proporcionando los datos necesarios para analizar mutaciones relacionadas a enfermedades mendelianas.

Un rasgo de interés sobre estas mutaciones es su distribución en el genoma. Conocer en qué regiones génicas se localizan, así como la proporción en cada una de ellas, revela si existen regiones más propensas a ocasionar enfermedades al ser mutadas, así como la contribución y la relevancia de las regiones intergénicas e intrónicas, que en un inicio y de manera intuitiva, podrían pensarse como menos relevantes al no codificar una proteína en sí mismas [8].

En este sentido, este trabajo pretende determinar cuál es la proporción de mutaciones asociadas a enfermedades mendelianas humanas que se encuentran en regiones codificantes, exónicas y génicas, así como en sus contrapartes. Para ello, la base de datos *Online Mendelian Inheritance in Man* (OMIM) resulta ser una valiosa herramienta al contener información completa y seleccionada sobre genes del humano, fenotipos genéticos y las relaciones entre ellos. A diferencia de fuentes de datos primarios, OMIM sintetiza y resume información nueva e importante basada en la revisión de expertos de la literatura biomédica. Como consecuencia, OMIM también desempeña un papel de liderazgo en la denominación y clasificación de fenotipos genéticos. *OMIM.org* fue creado para proporcionar un portal fácil de usar y de buscar para una compilación curada de la literatura útil en la investigación de la genética molecular y clínica.

Al 30 de octubre de 2014, OMIM comprendía más de 22,634 entradas que describen 14,831 genes y 7,894 fenotipos. Si bien el contenido de OMIM todavía está indexado y es accesible en el Centro Nacional de Información Biotecnológica (NCBI), el sitio web *OMIM.org* tiene una indexación con mayores capacidades de búsqueda, visualizaciones novedosas de las relaciones entre genes y fenotipos, y enlaces organizados por temas para una amplia variedad de recursos externos dirigidos a la información relacionada específicamente con los datos en la entrada OMIM [7].

Datos

El archivo *genemap2.txt* obtenido del portal de OMIM contiene los datos *input* de este trabajo, conteniendo información de las enfermedades mendelianas en su contexto genómico. Particularmente, contiene entradas de genes involucrados en fenotipos de herencia mendeliana en humano. Los datos presentados provienen tanto de OMIM, como de diferentes bases de datos tales como NCBI o Ensembl, y para facilitar la búsqueda por IDs en estas bases de datos, hay columnas con la información respectiva. Entre otros atributos, los datos de interés directo para este trabajo fueron el cromosoma donde se encuentran los genes y sus coordenadas genómicas en la versión GRCh38 del genoma humano (ver **Figura 1**). Indirectamente, el ID en Ensembl de los genes reveló su anotación en esta base de datos que fue central para efectuar el análisis aquí presentado. Dentro de los datos, existen distintos genes que comprenden una misma región genómica extensa, pero están asociados a distintos fenotipos, por lo que todos ellos fueron considerados.

# Chromosome	Genomic Position Start	Genomic Position End				
chr1	923922	944573	1p36.33	1p36.33	616765	SAMD11, MRS
chr1	944202	959255	1p36.33	1p36.33	610770	NOC2L, NIR
chr1	960583	965718	1p36.33	1p36.33	619262	KLHL17, AF
chr1	975197	982092	1p36.33	1p36.33	615921	PERM1, C1orf170
chr1	998963	1000096	1p36.31	1p36.33	608060	HES4 Hes family

Figura 1: Primeras columnas del archivo de datos de entrada. Se señalan los atributos de interés (en el recuadro rojo), así como datos de 5 genes de ejemplo (en el recuadro amarillo).

Metodología

El trabajo se valió de la generación e implementación de código en el lenguaje de programación **R** versión 4.2.0, así como de la herramienta bioinformática **API REST** (la interfaz de programación de aplicaciones) de **Ensembl** que proporciona acceso a datos genómicos de esta base de datos independientemente del lenguaje usado. Esta última nos permitió acceder y recuperar información de variantes asociadas a regiones génicas en *Homo sapiens*, así como filtrarlas según atributos de interés o conveniencia. El código se encuentra en el archivo *region_variants.ipyn*.

1. Extraer las coordenadas genómicas de los genes implicados en enfermedades mendelianas humanas

La primera etapa del trabajo fue la extracción de las regiones genómicas de los genes del archivo de entrada. Para esto, inicialmente se efectuó la limpieza de los datos, removiendo entradas con datos de interés faltantes. Subsecuentemente los datos fueron parseados y

filtrados de tal forma que fueran accesibles para la etapa siguiente de búsqueda de variantes.

2. Obtener las variantes asociadas a las regiones que cada gen comprende

Una vez filtrados y parseados los datos, las coordenadas genómicas de los 8,551 genes resultantes implicados en enfermedades mendelianas, así como el cromosoma donde se encuentran, fueron usados para extraer las variantes que residen en tales regiones. Esto se logró mediante la generación de solicitudes HTTP a Ensembl; en particular se ejecutaron solicitudes GET para recuperar los registros buscando en el conjunto de datos de variantes *clin_assoc* que contiene las variantes descritas por ***ClinVar*****,** un repositorio público de variantes genéticas relacionadas con fenotipos concretos; este engloba variantes patogénicas, posiblemente patogénicas, de respuesta a fármacos o de histocompatibilidad. La información fue entregada mediante el formato de intercambio de datos JavaScript Object Notation (JSON) (Ver código para más detalles).

3. Extraer variantes SNPs/*indels* patogénicas y con coordenadas genómicas del genoma humano GRCh38

De la fase anterior se obtuvieron 73,233 variantes totales que fueron posteriormente filtradas según su relevancia clínica, la versión del genoma en la que fueron anotadas y el tipo de variante: se retuvieron sólo aquellas variantes catalogadas como patogénicas, pues son aquellas asociadas a las enfermedades y por tanto las de interés; además, dado que las coordenadas genómicas del archivo *input* se basaron en la versión GRCh38 del genoma humano, se filtraron las variantes según este atributo; finalmente, solo se consideraron variantes provenientes del recurso ***dbSNP*** que contiene variantes de nucleótido único e *indels*, donde estos últimos no presentaron ambigüedad al ser todas las variantes únicas y atribuirse a un solo tipo de región genómica. Si bien existieron variantes repetidas, estas se encontraron asociadas a diferentes fenotipos (explicadas por efectos pleiotrópicos), y por tanto, se consideraron como diferentes variantes. Como *output* solo se tomó el tipo de consecuencia de cada variante resultante, pues ese atributo es justamente el que describe en qué tipo de región genómica se encuentra.

4. Separar variantes por tipo de región genómica en donde residen

Se obtuvieron 32,148 variantes pertenecientes a 2,094 genes, las cuales fueron separadas por tipo de consecuencia. La **Tabla 1** muestra las clasificaciones de las consecuencias de variantes en tipos de regiones genómicas mutuamente excluyentes entre sí, basados en la **Figura 2**.

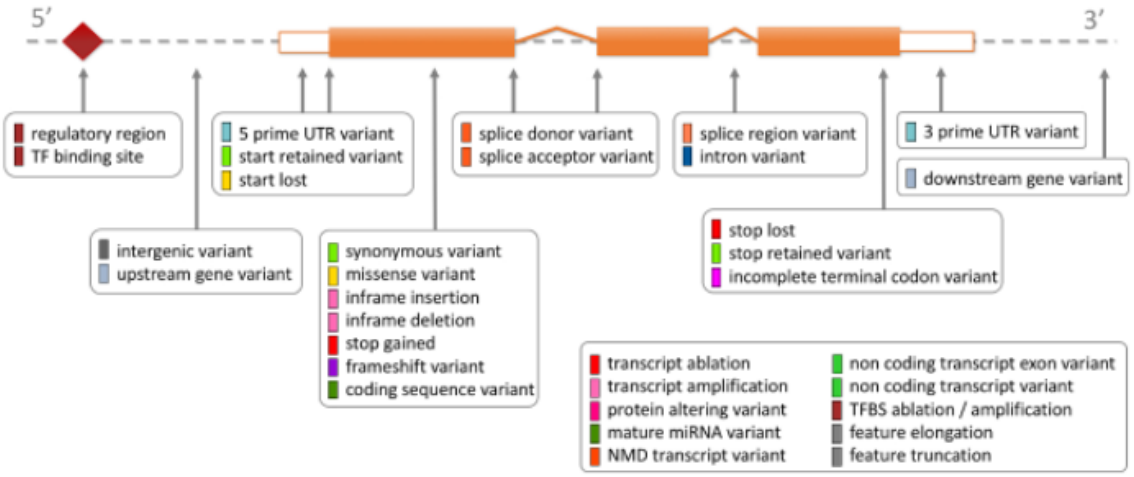


Figura 2: Tipo de localización genómica de las variantes según el tipo de consecuencia que ocasionan. Tomado de *Calculated consequences (ensembl.org)*.

Tabla 1: Tipos de región genómica de las variantes según sus consecuencias

Por región exónica/intrónica

Exónicas

- Variante en 5'-UTR
- Variante retenedoras del inicio de la transcripción
- Variante perdedoras del inicio de la transcripción
- Variante sinónimas
- Variante *missense*
- Inserción en marco de lectura
- Deleción en marco de lectura
- Ganancia codón de paro prematuro
- Variante de cambio en el marco de lectura
- Variante en región codificante
- Pérdida del codón de paro
- Variante retenedora del codón de paro
- Variante de codón de paro incompleto
- Variante en 3'-UTR
- Variantes exónicas no codificantes

Intrónicas

- Variantes en región de splicing
- Variantes intrónicas
- Variante en región donante de *splicing*
- Variante en región aceptora de *splicing*

Por región codificante/no codificante (dentro de genes)

Codificante

- Variante retenedoras del inicio de la transcripción
- Variante perdedoras del inicio de la transcripción
- Variante sinónimas
- Variante *missense*
- Inserción en marco de lectura
- Deleción en marco de lectura
- Ganancia codón de paro prematuro
- Variante de cambio en el marco de lectura
- Variante en región codificante

No codificante

- Variantes en región de splicing
- Variantes intrónicas
- Variante en región donante de *splicing*
- Variante en región aceptora de *splicing*
- Pérdida del codón de paro
- Variante retenedora del codón de paro
- Variante de codón de paro incompleto
- Variante en 3'-UTR
- Variante en 5'-UTR

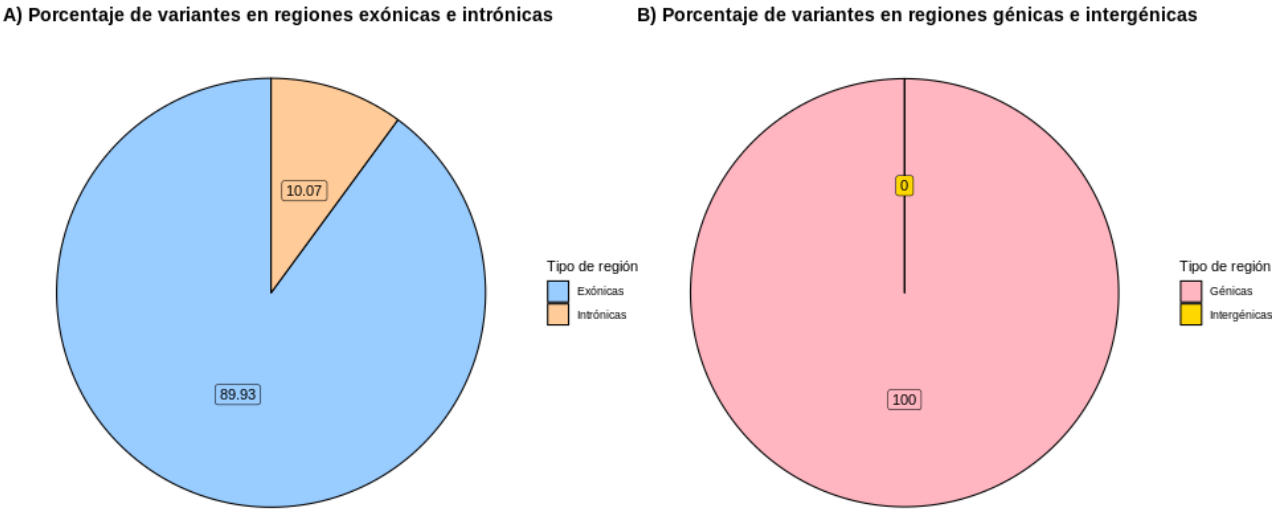
Por región Génica/Intergénica	
Génica	
• Variante en 5'-UTR • Variante retenedoras del inicio de la transcripción • Variante perdedoras del inicio de la transcripción • Variante sinónimas • Variante *missense* • Inserción en marco de lectura • Deleción en marco de lectura	• Variantes en región de splicing • Variantes intrónicas • Variante en región donante de *splicing* • Variante en región aceptora de *splicing* • Ganancia codón de paro prematuro • Variante de cambio en el marco de lectura • Variante en región codificante
Intergénica	
• Regiones regulatorias • Sitios de unión a TFs • Variantes intergénicas • Variantes río arriba del gen • Variantes río abajo del gen	

UTR: *untranslated region.* TF: factor de transcripción.
*Para más detalles consultar la referencia [4].

Adicionalmente, se calculó el número de variantes por gen y cromosoma con la finalidad de analizar la distribución genómica de las variantes a un nivel más amplio del genoma.

Resultados

Como es posible apreciar en la **Figura 3**, se encontró que el 89.93% de las variantes fueron exónicas y el restante 10.07% intrónicas, mientras que el 89.2% resultaron codificantes y el 10.8% no codificantes; dado que las regiones génicas comprenden a las exónicas e intrónicas, el 100% de las variantes resultaron ser génicas, sin variantes intergénicas en regiones río arriba o abajo de genes, ni en regiones reguladoras o sitios de unión a factores de transcripción. Además, dado que todas las variantes codificantes son forzosamente exónicas, en conjunto estas representan el 89.93% de las variantes.



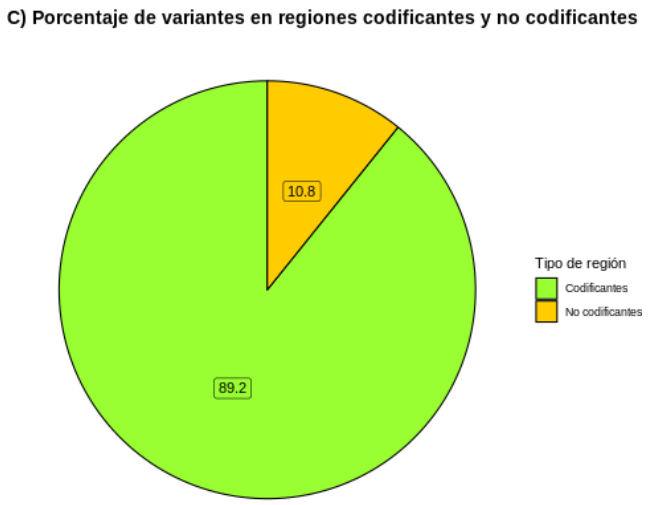


Figura 3: Proporción de variantes patogénicas asociadas a enfermedades mendelianas humanas en **A)** regiones exónicas e intrónicas, **B)** regiones génicas e intergénicas y **C)** regiones codificantes y no codificantes.

Cabe destacar que en conjunto las variantes codificantes, intrónicas e intergénicas no abarcan al 100% de la variación puesto que existen variantes exónicas que no son codificantes y no se están contemplando (ver **Figura 4**).

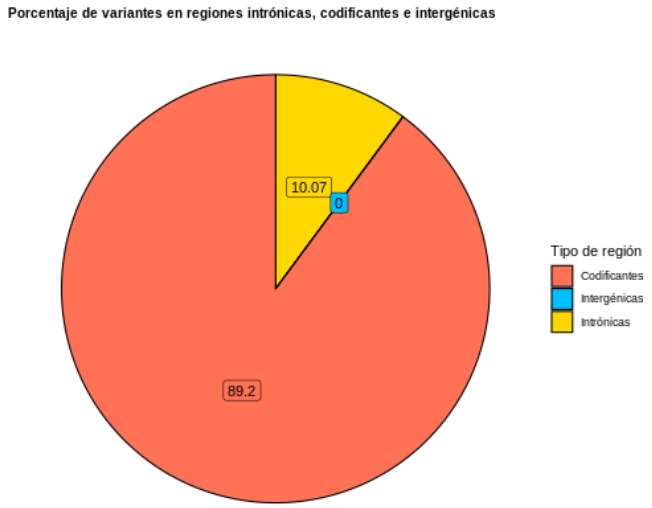


Figura 4: Proporción de variantes patogénicas asociadas a enfermedades mendelianas humanas en regiones codificantes, intrónicas e intergénicas.

Por otro lado, en la **Figura 5** se muestra el número de variantes encontradas en cada cromosoma, así como la proporción que representan del total. Como es posible notar, los cromosomas 1, 13, 15 y 17 fueron los que mayor número de variantes poseyeron, con más del 7% del total cada uno; en contraste, los cromosomas 8, 20 y 22 no alcanzaron ni el 1% de las variantes totales, y para el cromosoma Y no se encontraron.

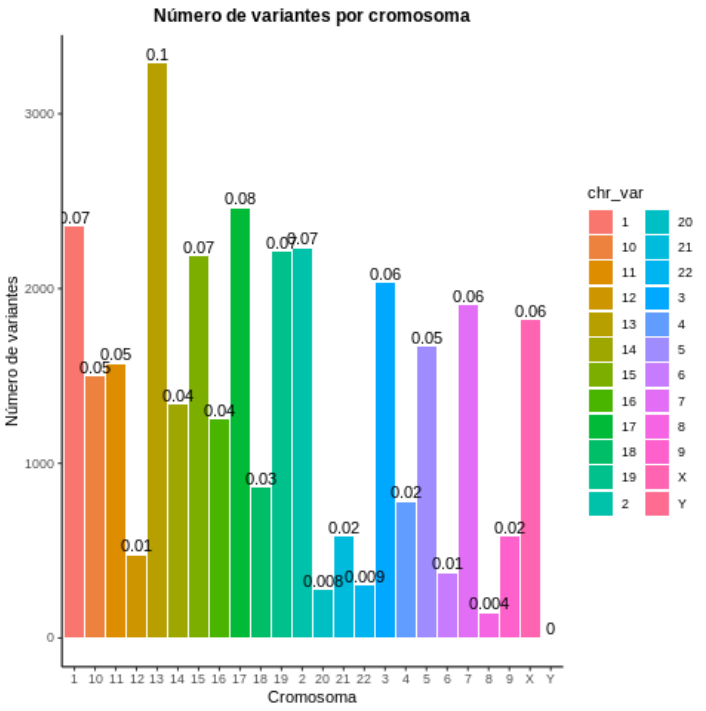


Figura 5: Número de variantes patogénicas asociadas a enfermedades mendelianas humanas en cada cromosoma. Arriba de cada barra se muestra el porcentaje de variantes de cada cromosoma con respecto al total de las encontradas.

Finalmente, en la **Figura 6** se puede apreciar la distribución del número de variantes en los genes, y notablemente, que más del 70% de los 8,551 genes considerados, no tuvieron una variante patogénica asociada. Inesperadamente, el gen BRCA2 tuvo 2,418 variantes asociadas (ver Figura 6B, no se muestra en la Figura 6A).

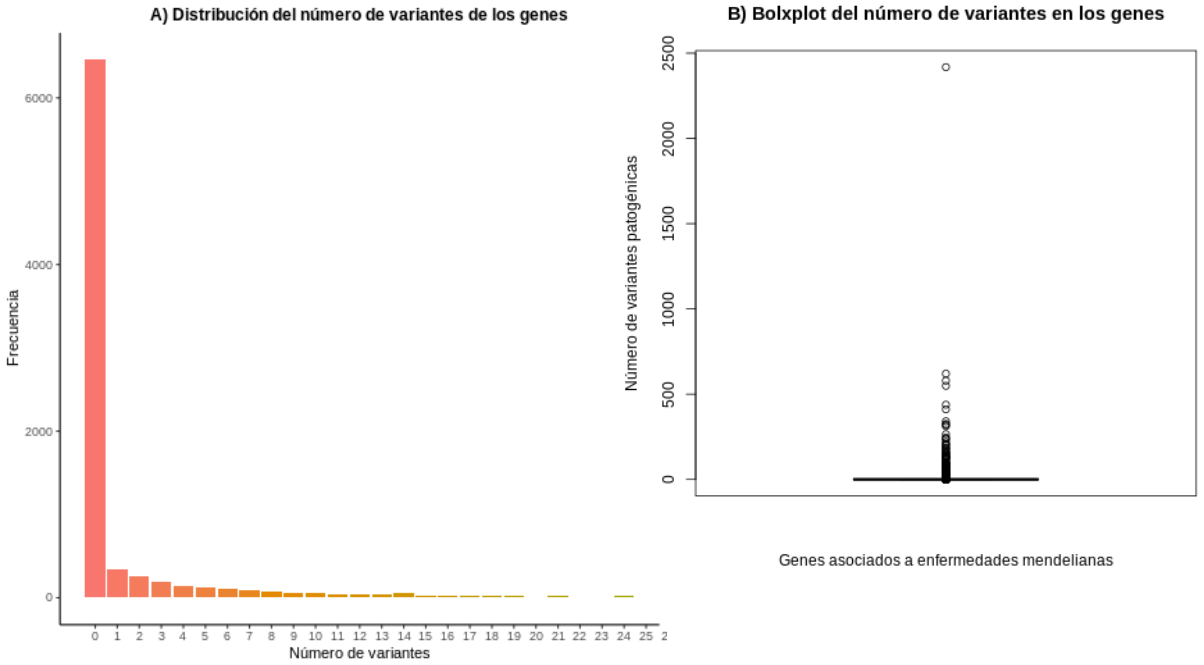


Figura 6: A) Distribución del número de variantes de los genes asociados a enfermedades mendelianas en humano. **B)** Boxplot del número de variantes de los genes. Nótese el *outlier* en la parte superior correspondiente al gen BRCA2.

Congruentemente con esto último encontrado, se han reportado 1730 mutaciones y variantes en el gen BRCA2, un supresor de tumores con un rol en la reparación en el ADN. Mutaciones en este gen están asociadas a una forma de cáncer de mama que es heredable de manera autosómica [1]. Por tanto, haberlo encontrado asociado a tantas variables en este estudio reafirma y se sustenta de tales datos.

Discusión

Al realizar este análisis se obtuvieron resultados relevantes sobre la distribución de las variantes en distintas partes del genoma. Encontramos que la mayoría de las variantes encontradas en pacientes que presentaban enfermedades mendelianas se acumulan en regiones exónicas o codificantes del genoma en comparación con aquellas encontradas en zonas intrónicas o no codificantes. Este resultado es el esperado, ya que estas zonas del genoma son las que como producto final generan a las proteínas que dan lugar al fenotipo que podemos apreciar en las enfermedades cuando se tienen variantes patogénicas. Sin embargo, es importante aclarar que la fuente primaria de datos OMIM, se concentra principalmente en enfermedades ligadas a la fracción codificante o exónica del genoma [2], ocasionando que los datos iniciales de este trabajo estén sesgados hacia estas regiones, y por ende, que las variantes encontradas también lo estén.

Otro punto a destacar es la distribución de variantes en los genes, pues la gran mayoría de los genes no poseyeron variantes patogénicas. Esto podría deberse a que, al analizar las variantes para un enfermedad en particular, solo se registraron variantes perjudiciales en aquellos genes cuya función afecta de forma directa o indirecta al fenotipo; también es posible que existan variantes patogénicas en genes cuyas funciones no poseen una relación aparente con cierto fenotipo de herencia mendeliana, pero que pueden interactuar con aquellos que sí tienen un efecto directo. Asimismo, otra posible explicación para esto, es que el archivo de entrada de OMIM involucra tanto a genes implicados en enfermedades cuya base molecular ha sido descrita y mutaciones han sido encontradas, como a genes que han sido efectivamente asociados a la enfermedad pero cuyos efectos subyacentes y/o mutaciones no han sido descubiertos, esto es, sin variantes asociadas.

De igual manera es importante mencionar que el cromosoma que menos variaciones presentó fue el Y. Se sabe que la pérdida de uno de los segmentos del cromosoma X dio lugar al cromosoma Y. De esta manera, la dimensión de este cromosoma es más reducida y, dado su menor tamaño, incluye también menos información genética que el resto de los cromosomas, por lo que se puede deducir que el tamaño y la cantidad de información

contenida tienen una correlación con la probabilidad de encontrar variaciones nocivas en este cromosoma.

Finalmente es importante resaltar que los patrones de herencia mendelianos no son universales en las enfermedades de carácter genético. Este fenómeno ha ocasionado que se busquen mecanismos alternativos para explicar el origen de dichas enfermedades que no siguen un patrón mendeliano de herencia, principalmente en vías reguladoras independientes de mutaciones en genes codificantes. Dado que estas enfermedades son multifactoriales, requieren otro tipo de acercamientos o enfoques que permitan asociar un conjunto o *cluster* de genes asociados a una condición o enfermedad [2].

Referencias

1. Abu-Helalah, M., Azab, B., Mubaidin, R. et al. BRCA1 and BRCA2 genes mutations among high risk breast cancer patients in Jordan. Sci Rep 10, 17573 (2020). https://doi.org/10.1038/s41598-020-74250-2

2. Brownlee, C. (2017, mayo). OMIM Turns 50: A Genetic Database's Past, Present, and Future. Hopkinsmedicine.Org. https://www.hopkinsmedicine.org/research/advancements-in-research/fundamentals/in-depth/omim-turns-50-a-genetic-databases-past-present-and-future

3. Carnevale, A. (2014, 1 junio). El nuevo abordaje de las enfermedades mendelianas. revista.unam.mx. http://www.revista.unam.mx/vol.15/num6/art46/

4. Ensembl. (s.f). Ensembl Variation - Calculated variant consequences. Sitio web: Calculated consequences (ensembl.org)

5. González-Lamuño, D., & García Fuentes, M.. (2008). Enfermedades de base genética. Anales del Sistema Sanitario de Navarra, 31(Supl. 2), 105-126. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S113766272008000400008&lng=es&tlng=es.

6. Hindorff, L. (2022, 11 mayo). Herencia mendeliana | NHGRI. Genome.gov. https://www.genome.gov/es/genetics-glossary/Herencia-mendeliana

7. Joanna S. A., Carol A. B., François S., Alan F. S., Ada H. (2015, 28 enero). OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders, Nucleic Acids Research, Volume 43, Issue D1, Pages D789–D798, https://doi.org/10.1093/nar/gku1205

8. Supriya, N. (2021, 24 junio). What is Mendelian Inheritance? Definition, Traits & Laws. Biology Reader. https://biologyreader.com/mendelian-inheritance.html