

# **Distribución de las mutaciones responsables de enfermedades monogénicas en el genoma humano**

Leal, J.E., Mourra, C.M., Tamayo A., Zagal, A.

Diabetes Mellitus 2, Alzheimer e Hipertensión Arterial, estas enfermedades, entre muchas otras, son consideradas enfermedades complejas por ser resultado de la integración de diversos componentes genéticos y ambientales, generando fenotipos que no siguen patrones estándar de herencia mendeliana (Horikawa, 2018; Escott-Price et al., 2015). Sin embargo, aunque no se puede buscar directamente la etiología de enfermedades complejas en mutaciones de un solo gen, estudiar enfermedades monogénicas ha probado ser de gran utilidad para caracterizar los defectos moleculares que resultan en enfermedades complejas (Peltonen, et al. 2006).

Si bien los defectos moleculares involucrados en muchas enfermedades monogénicas surgen de mutaciones en regiones codificantes, un subconjunto importante puede relacionarse con variantes no codificantes implicadas en la regulación de la transcripción, splicing y la estructura tridimensional del genoma (Valente, E., & Bhatia, K., 2018; Thaventhiran et al., 2020). Lo cual nos permite formular la siguiente interrogante ¿En qué proporción encontramos mutaciones responsables de enfermedades monogénicas en regiones codificantes y regiones no-codificantes?

Por esto, en este trabajo se realizará un análisis bioinformático (*Diagrama 1*) para mapear los variantes relacionados con enfermedades monogénicas, obtenidos de la base de datos OMIM. Esto con el propósito de llegar a una conclusión sobre la distribución genómica de las variantes en cuestión; sospechando que las mutaciones en regiones no codificantes están menos relacionadas con enfermedades monogénicas gracias a su papel regulatorio, el cual posiblemente las relacione más con enfermedades poligénicas.

## **Metodología**

### **Obtención de datos heredables de OMIM**

Mediante la Base de Datos “*Online Mendelian Inheritance in Man*” (OMIM) se obtuvo el archivo *genemap2.txt* (*Figura 1a*). Dicho archivo contiene los genes con las coordenadas correspondientes a las variantes del genoma, con su fenotipo asociado. A partir de este archivo, se extrajeron los datos de mayor interés al discriminar aquellos que si presentaran un campo de herencia. Además de filtrar las enfermedades heredables, se recuperaron campos como el cromosoma, las coordenadas y el número MIM, los cuales se utilizaron más adelante. (“OMIM Gene Map Statistics”, 2021).



Diagrama 1. Flujo de trabajo

### Localización de variantes con Ensembl

Con los datos capturados de OMIM, logramos obtener todas las variantes asociadas a las enfermedades anotadas en el genoma de referencia hg38 mediante los módulos de *python request* y *sys*, con los cuales se estableció un acceso programático a la base de datos *Ensembl*. Para lograr esto, el programa se encargó de recopilar las variantes de interés filtrando únicamente aquellas que estuvieran asociadas a un número *MIM* y que además se encontraran en la región delimitada por las coordenadas cromosómicas obtenidas de OMIM. Una vez realizado este proceso, los campos de interés de las variantes recopiladas se guardaron en una archivo con el nombre *VariantFileOMIM3.txt*. Entre los campos guardados se encontraban el id de la variante, la localización exacta donde ocurrió el cambio de nucleótidos y el tipo de herencia que poseía la variante de acuerdo a la enfermedad a la que se asocia (*Figura 1b*).

### Generación del archivo avinput

Con el archivo *VariantFileOMIM3.txt* obtenido anteriormente, se buscó en la base de datos SNV cual fue el cambio de nucleótidos ocurrido entre el alelo de referencia y el alternativo (*Figura 1c*). Además de capturar el cambio de secuencia entre alelos, se obtuvo la significancia clínica de la variante con el propósito de determinar si esta era causal de la enfermedad al estar etiquetada como patogénica. La información obtenida se guardó en formato *avinput* para poder procesarlo en *annoVar* (*Figura 1d*). Los primeros 5 campos delimitados por espacios de este formato son: Cromosoma, Coordenadas, de inicio y de término, Alelo de Referencia y Alelo Alternativo (Wang, K., 2015).

## Mapeo de la región genómica

Se utilizó *annoVar* debido a que tiene la ventaja de mapear con distintas bases de datos como referencia (*RefSeq*, *UCSC Known Gene*, *Ensembl Genes*, *GENCODE Genes*). El archivo output provee la región genómica donde se encuentra la variante: región exónica, intrónica, *ncRNA*, *UTR5*, *UTR3*, *splicing*, *upstream*, *downstream* e intergénica (Yang H & Wang K. 2015).

a)

#	Chr	romosome	Genomic Position	Start	Gene
chr1	0	27600000	1p36	607413	AD7CNP
chr1	0	27600000	1p36	612367	ALPQTL2
chr1	0	123400000	1p	606788	ANON1
chr1	0	27600000	1p36	605462	BCC1
chr1	0	27600000	1p36	606928	BMND3

b)

#ID	Chromosome	Start	End	MIM	Inheritance
rs1569231897	X	10213710	10213710	302910	X-linked dominant
rs1569233549	X	10220876	10220876	302910	X-linked dominant
rs1569226551	X	10187602	10187602	302910	X-linked dominant
rs1569230006	X	10206463	10206463	302910	X-linked dominant

c)

□ rs104894737 [*Homo sapiens*]  
1.  
Variant type: SNV  
Alleles: T>C [Show Flanks]  
Chromosome: X:11294790 (GRCh38)  
X:11312910 (GRCh37)  
Canonical SPDI: NC\_00023.11:11294789:T:C  
Gene: AMELX [Varview], ARHGPAP6 [Varview]  
Functional Consequence: initiator\_codon\_variant,intron\_variant,missense  
Clinical significance: pathogenic  
Validated: by frequency,by alfa,by cluster  
MAF: C=0.0 ([ALFA](#))  
HGVS: NC\_00023.11:g.11294790T>C, NC\_00023.10:g.  
NG\_012040.1:g.6378T>C, NM\_001142.2:c.2T>C, |

d)

X 10213710 10213710 G A comments:  
X 10220876 10220876 G C comments:  
X 10187602 10187602 G A comments:  
X 10206463 10206463 C G comments:  
X 11121652 11121652 C T comments:  
X 11120974 11120974 C T comments:

Figura 1. a) Formato de archivo *genemap2.txt* de OMIM junto con los campos de interés. b) Archivo *VariantFileOMIM3.txt* con los campos extraídos de Ensembl. c) Campos capturados de SNV en la base de datos de NCBI. d) Formato del archivo *.avininput* resultante.

## Resultados

Una vez que se filtraron los datos asociados a enfermedades heredables en *genemap2.txt*, se localizaron las secuencias de los alelos alternativos. Tras esta acción obtuvimos como resultado el archivo *Monogenic.avininput* el cual estaba compuesto de un total de 24,570 variantes genéticas asociadas a enfermedades que presentaban un tipo de herencia.

Las variantes arrojadas por *request* y *sys*, coincidían entre las coordenadas anteriormente delimitadas por OMIM, pero algunas, se asociaron a enfermedades con herencia no mendeliana; por esto, una vez creado el *avinput*, se filtraron las variantes con las siguientes especificaciones. Con *Filter Variants.Rmd* y *getPathogenicVariants.ipynb* se filtró el tipo de herencia “*Multifactorial*” y “*Isolated cases*”, para quedarnos con 24530 variantes.

Posteriormente para confirmar que dichas variantes son causantes de la enfermedad nos quedamos con variantes cuya descripción tuvieran “*pathogenic*” y que en la misma descripción no se tuviera “*benign*”, “*uncertain-significance*”, “*conflicting-interpretations of pathogenicity*” y “*protective*”. Al final del filtro, se obtuvieron 20237 variantes.

Se hizo la anotación basada en genes, mediante *annotate\_variation.pl* utilizando la versión del genoma *hg38* y las bases de datos generadas por *annovar* en *humandb*. Se obtuvieron 17582 variantes anotadas en el archivo *variant\_function* y 2655 variantes fueron inválidas (archivo *invalid\_input*). Las variantes inválidas se debieron a *indeles* en los que un mismo nucleótido en el alelo de referencia era sustituido varias veces en los módulos de python anteriormente descritos. Por ejemplo, en la figura 2a, se muestra como en la primer variante con coordenadas 11298779-11298781 su alelo de referencia es C, mientras que en *Ensembl*, la referencia es CCC.

Para afrontar eso, se creó la función *FilterInvalid.Rmd* en donde se parsearon las variantes de un formato *avinput* a formato *simple\_format* (Figura 2b) el cual contiene el cromosoma con la coordenada de inicio separados por “:”, seguido de las coordenadas de término separados por “-”. Creando el archivo *.simple\_format*

<b>a)</b>	<table border="1"> <tbody> <tr> <td>1</td><td>X</td><td>11298779</td><td>11298781</td><td>C</td><td>-</td><td>comments:</td><td>rs387906491,</td><td>Deletion,</td><td>X-linked</td></tr> <tr> <td>2</td><td>X</td><td>11298833</td><td>11298834</td><td>C</td><td>-</td><td>comments:</td><td>rs387906489,</td><td>Deletion,</td><td>X-linked</td></tr> <tr> <td>3</td><td>X</td><td>11298899</td><td>11298902</td><td>C</td><td>-</td><td>comments:</td><td>rs387906490,</td><td>Deletion,</td><td>X-linked</td></tr> </tbody> </table>	1	X	11298779	11298781	C	-	comments:	rs387906491,	Deletion,	X-linked	2	X	11298833	11298834	C	-	comments:	rs387906489,	Deletion,	X-linked	3	X	11298899	11298902	C	-	comments:	rs387906490,	Deletion,	X-linked
1	X	11298779	11298781	C	-	comments:	rs387906491,	Deletion,	X-linked																						
2	X	11298833	11298834	C	-	comments:	rs387906489,	Deletion,	X-linked																						
3	X	11298899	11298902	C	-	comments:	rs387906490,	Deletion,	X-linked																						
<b>b)</b>	<pre>chr10:4000000-4000100 chr10:7000000-8000000</pre>																														
<b>c)</b>	<pre>&gt;chrX:11298779-11298781 Comment: this sequence (leftmost exon at chrX:11298778) is generated by ANNOVAR on Thu May 27 23:43:02 2021, based on regions specified in Monogenic3.simple_region and sequence files stored at ../annovar/humandb/hg18seq/chroms. CCC &gt;chrX:11298833-11298834 Comment: this sequence (leftmost exon at chrX:11298832) is generated by ANNOVAR on Thu May 27 23:43:02 2021, based on regions specified in Monogenic3.simple_region and sequence files stored at ../annovar/humandb/hg18seq/chroms. CC &gt;chrX:11298899-11298902 Comment: this sequence (leftmost exon at chrX:11298898) is generated by ANNOVAR on Thu May 27 23:43:02 2021, based on regions specified in Monogenic3.simple_region and sequence files stored at ../annovar/humandb/hg18seq/chroms. CCCC &gt;chrX:11298243-11298246 Comment: this sequence (leftmost exon at chrX:11298242) is generated by ANNOVAR on Thu May 27 23:43:02 2021, based on regions specified in Monogenic3.simple_region and sequence files stored at ../annovar/humandb/hg18seq/chroms. CCCC &gt;chrX:11772247-11772255 Comment: this sequence (leftmost exon at chrX:11772246) is generated by ANNOVAR on Thu May 27 23:43:02 2021, based on regions specified in Monogenic3.simple_region and sequence files stored at ../annovar/humandb/hg18seq/chroms. GATTGTTG</pre>																														

**Figura 2.** a) Variantes inválidas. b) Formato *simple\_format*. c) Archivo *.simple\_region.fa* con las secuencias de los alelos de referencia de las variantes inválidas ya corregidas en formato *fasta*.

Utilizando *retrieve\_seq\_from\_fasta.pl* de annovar, se recuperaron las variantes de invalid format con formato fasta donde se recuperaron 2565 variantes de 2655. Una vez teniendo las secuencias correctas de referencia, se escribieron en formato *avinput* y se anotaron con las mismas especificaciones.

Finalmente se filtraron las variantes por *loci*, es decir, se eliminaron variantes sinónimas y nos quedamos con sólo una variante por coordenada, esto para estudiar cuántos *loci* hay asociados a cada enfermedad y determinar si los resultados son distintos a los del análisis hecho con las variantes. Así de 20237 variantes nos quedamos con 16781 *loci*.

Siguiendo el mismo *workflow*, estos *loci* filtrados se anotaron con *annotate\_variant.pl* de annovar y los *loci* con alelos de referencia inválidos corregidos con *retrieve\_seq\_from\_fasta.pl*.

Al final se anotaron 20147 variantes de 20237, y 15208 *loci* de 16781.

El output de *annotate\_variation.pl* de annovar contiene dos archivos, *variant\_function* y *exonic\_variant\_function*. Los valores posibles de la anotación en *variant\_function* son:

- **Exonic:** Cuando la variante sobrelapa con la porción exónica codificante, excluyendo a las porciones *UTR*
- **Splicing:** Cuando la variante se encuentra a 2 nucleótidos de distancia de un límite exón/intrón.
- **ncRNA:** La variante se superpone a un transcripto sin anotación codificante
- **UTR5 y UTR3:** Cuando la variante sobrelapa una región 5' o 3' no traducida. Puede ser el caso que la región coincida en ambas *UTR5* y *UTR3*, ambas regiones de distintos genes, entonces el valor será “*UTR5,UTR3*”
- **Intronic:** Cuando la variante sobrelapa con una región de transcripción primaria que se elimina al generar el ARN maduro.
- **Upstream, Downstream:** Cuando la variante se encuentra a 1 kb de distancia del sitio de inicio de la transcripción o del sitio final de la transcripción, respectivamente, teniendo en cuenta la hebra del ARNm. Se puede dar el caso que una variante se encuentre en ambas regiones, entre el inicio de la transcripción y el final de la transcripción de dos distintos genes, en ese caso el valor será “*upstream,downstream*”
- **Intergenic:** Cuando la variante sobrelape con una región no codificante que este ubicada entre genes, la cual podría contener elementos regulatorios (esta distancia es mayor a 1kb).

Es posible que existan anotaciones que sobrelapan en varias categorías, un exón, puede ser clasificado en la parte de splicing, es por esto que se siguen los siguientes niveles de procedencia:

exonic = splicing > ncRNA > UTR5/UTR3 > intron > upstream/downstream > intergenic.

**Figura 3a**

```
X 10213710 10213710 G A comments: rs1569231897, Mutation, X-linked dominant, 302910, pathogenic
X 10220876 10220876 G C comments: rs1569233549, Mutation, X-linked dominant, 302910, pathogenic
X 10187602 10187602 G A comments: rs1569226551, Mutation, X-linked dominant, 302910, pathogenic
X 10206463 10206463 C G comments: rs1569230006, Mutation, X-linked dominant, 302910, pathogenic
X 11121652 11121652 C T comments: rs121917889, Mutation, X-linked dominant, 300056, pathogenic
X 11120974 11120974 T C comments: rs121917888, Mutation, X-linked dominant, 300056, pathogenic
X 11294802 11294802 TTTTATTG - comments: rs387906488, Deletion, X-linked dominant, 300391, pathogenic
X 11294790 11294790 T C comments: rs104894737, Mutation, X-linked dominant, 300391, pathogenic
X 11298779 11298779 C - comments: rs387906491, Deletion, X-linked dominant, 300391, pathogenic
X 11298833 11298833 C - comments: rs387906489, Deletion, X-linked dominant, 300391, pathogenic
```

Primeras líneas del archivo .avinput

**Figura 3b**

```
exonic CLCN4 X 10213710 10213710 G A comments: rs1569231897, Mutation, X-linked dominant, 302910, pathogenic
exonic CLCN4 X 10220876 10220876 G C comments: rs1569233549, Mutation, X-linked dominant, 302910, pathogenic
exonic CLCN4 X 10187602 10187602 G A comments: rs1569226551, Mutation, X-linked dominant, 302910, pathogenic
exonic CLCN4 X 10206463 10206463 C G comments: rs1569230006, Mutation, X-linked dominant, 302910, pathogenic
exonic HCCS X 11121652 11121652 C T comments: rs121917889, Mutation, X-linked dominant, 300056, pathogenic
exonic HCCS X 11120974 11120974 C T comments: rs121917888, Mutation, X-linked dominant, 300056, pathogenic
exonic AMELX X 11294802 11294802 TTTTATTG - comments: rs387906488, Deletion, X-linked dominant, 300391, pathogenic
exonic AMELX X 11294790 11294790 T C comments: rs104894737, Mutation, X-linked dominant, 300391, pathogenic
exonic AMELX X 11294799 11294799 G C comments: rs104894738, Mutation, X-linked dominant, 300391, pathogenic
exonic AMELX X 11298569 11298569 C A comments: rs104894736, Mutation, X-linked dominant, 300391, pathogenic
```

Primeras líneas del archivo .variant\_function, donde la primer columna es el valor de las regiones anotadas, en este caso *exonic*

**Figura 3c**

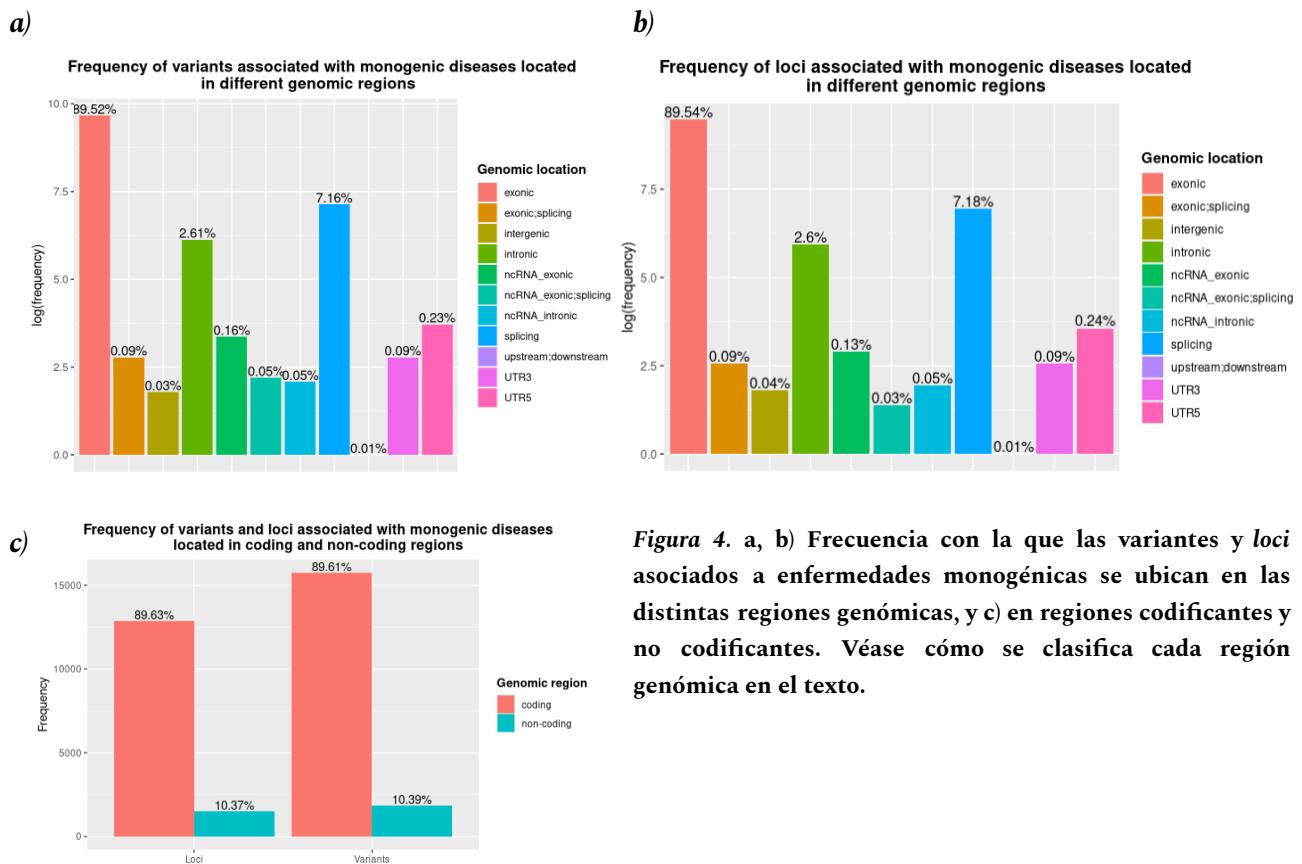
```
line1 nonsynonymous SNV CLCN4:NM_001830:exon11:c.G1606A:p.V536M,CLCN4:NM_001256944:exon9:c.G1324A:p.V442M, X 10213710
10213710 G A comments: rs1569231897, Mutation, X-linked dominant, 302910, pathogenic
line2 nonsynonymous SNV CLCN4:NM_001830:exon12:c.G2191C:p.G731R,CLCN4:NM_001256944:exon10:c.G1909C:p.G637R, X 10220876
10220876 G C comments: rs1569233549, Mutation, X-linked dominant, 302910, pathogenic
line3 nonsynonymous SNV CLCN4:NM_001830:exon4:c.G232A:p.G78S, X 10187602 10187602 G A comments: rs1569226551, Mutation, X-
linked dominant, 302910, pathogenic
```

Primeras líneas del archivo .exonic\_variant\_function en donde se muestran el tipo resultado de las mutaciones en regiones codificantes. En este caso, variación en un sólo nucleótido no sinónima (nonsynonymous SNV)

Posteriormente se hicieron algunos gráficos para determinar la frecuencia con la que las variantes y *loci* asociados a enfermedades monogénicas se ubican en regiones codificantes y no codificantes, así como en ciertas regiones genómicas específicas (*Figura 4*). También visualizamos la frecuencia del tipo de herencia de las enfermedades asociadas a las variantes y *loci* (*Figura 5d*), la frecuencia de los distintos tipo de mutaciones que generan a las variantes y *loci* exónicas y genómicas asociadas a enfermedades monogénicas (*Figura 5a-c*), el número de genes con cierto número de variantes (figura 6a) y la distribución de las variantes y *loci* asociadas a enfermedades monogénicas entre los distintos cromosomas (*Figura 6b-c*). Los códigos con los

que se generaron los gráficos se encuentran en los archivos *plots\_variants.rmd* y *plots\_loci.rmd*, respectivamente.

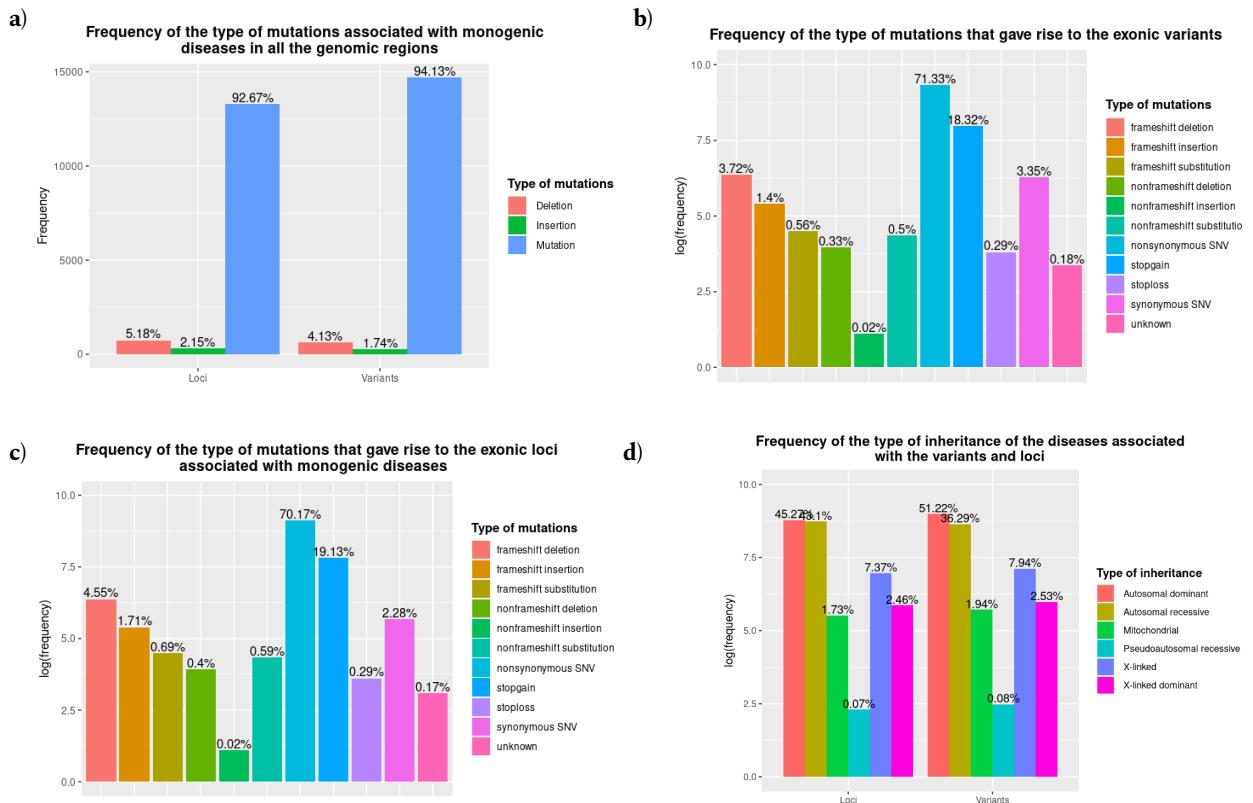
Claramente las regiones exónicas son las que contienen más *loci* y variantes asociadas a enfermedades monogénicas, seguidas por los sitios de splicing y las regiones intrónicas (*Figura 4a-b*). Las demás regiones contienen menos del 1% de las variantes y *loci* asociados a enfermedades monogénicas. Para determinar cómo se distribuyen las variantes y los *loci* asociados a enfermedades monogénicas entre regiones codificantes y no codificantes se colapsaron los datos de las regiones genómicas de tal manera que se clasificó como codificante a las regiones *exonic* y *exonic;splicing*, y el resto se clasificó como no codificante. Encontramos que tanto los *loci* como las variantes se encuentran predominantemente en regiones codificantes (*Figura 4c*). Además, desde aquí podemos notar que al hacer el análisis con las variantes o *loci* obtenemos resultados muy parecidos.



**Figura 4. a, b)** Frecuencia con la que las variantes y *loci* asociados a enfermedades monogénicas se ubican en las distintas regiones genómicas, y c) en regiones codificantes y no codificantes. Véase cómo se clasifica cada región genómica en el texto.

La mayoría de las variantes y *loci* asociados a enfermedades monogénicas se derivan de mutaciones de un sólo nucleótido, mientras sólo alrededor del 6% se derivan de inserciones o delecciones (*Figura 5a*). En las regiones exónicas específicamente, alrededor del 70% de las mutaciones que dan lugar a los loci y variantes son de un sólo nucleótido, mientras que el ~18%

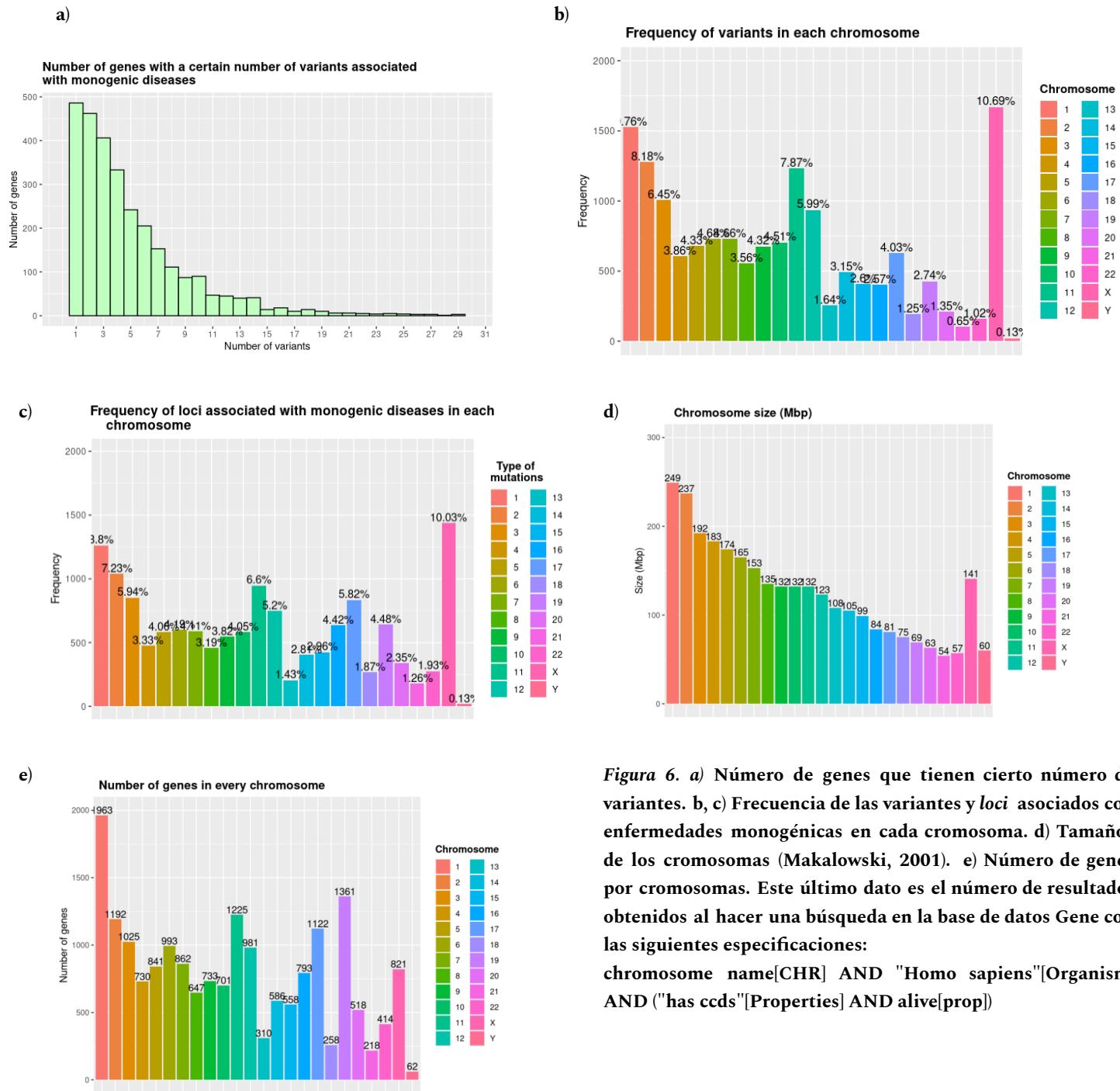
generan un nuevo codón de paro y sólo el 0.02% de las variantes y loci son generados por inserciones que cambian el marco de lectura (*Figura 5b-c*). Además, alrededor del 48% de las enfermedades asociadas a las variantes y loci se heredan de forma autosómica dominante, ~39% de manera autosómica recesiva y ~10% se heredan de manera ligada al sexo, específicamente al cromosoma X (*Figura 5d*). De hecho, el 10% de las variantes y loci se encuentran en el cromosoma X, mientras tan solo el 0.13% de las variantes se encuentran en el cromosoma Y (*Figura 6b-c*).



**Figura 5.** a) Frecuencia de los tipos de mutaciones que dan lugar a variantes y loci asociados con enfermedades monogénicas. b, c) Frecuencia de los tipos de mutaciones que dan lugar a variantes y loci exónicos asociados con enfermedades monogénicas. d) Frecuencia de los tipos de herencia de las variantes y loci asociados con enfermedades monogénicas.

Es interesante notar que la distribución del número de variantes y loci asociadas a enfermedades mendelianas en los cromosomas (*Figura 6b*) no se parece en nada a la del tamaño de los cromosomas (*Figura 6d*), lo cual indica que aunque un cromosoma tenga mayor tamaño y, por tanto, mayor probabilidad de adquirir mutaciones, el tamaño del cromosoma no es un buen predictor del número de variantes o loci que puede tener asociadas a una enfermedad mendeliana. Esto es lógico al pensar que nuestros resultados nos muestran que estas variantes se encuentran mayormente en regiones codificantes (*Figura 4c*), entonces el número de genes que tiene cada cromosoma debería de ser un mejor predictor. Aunque la distribución se parece un

poco más (*Figura 6e*), podemos notar que el número de genes que tiene un cromosoma no explica totalmente el número de variantes o *loci* asociados a enfermedades mendelianas que contiene.



**Figura 6.** a) Número de genes que tienen cierto número de variantes. b, c) Frecuencia de las variantes y *loci* asociados con enfermedades monogénicas en cada cromosoma. d) Tamaños de los cromosomas (Makalowski, 2001). e) Número de genes por cromosomas. Este último dato es el número de resultados obtenidos al hacer una búsqueda en la base de datos Gene con las siguientes especificaciones:

```
chromosome name[CHR] AND "Homo sapiens"[Organism]
AND ("has ccds"[Properties] AND alive[prop])
```

Por último, los genes que tomamos en cuenta en este análisis en promedio tienen ~5 variantes asociadas a una enfermedad monogénica (*Figura 6a*). Sin embargo, es claro que la mayoría tiene

una o dos variantes. Encontramos que el factor de coagulación VIII, tiene 173 variantes y 137 loci asociados a la hemofilia A. La hemofilia es una enfermedad ligada al cromosoma X, y el número de variantes asociadas con este gen son el 12% de todas las variantes que encontramos en este cromosoma. Después del cromosoma X, el cromosoma 1 es el que tiene mayor número de variantes y loci asociados a una enfermedad monogénica (*Figura 6b-c*) y, curiosamente, este es el cromosoma con el mayor número de genes (*Figura 6e*). Otro gen interesante es el que codifica para la hemoglobina beta locus, el cual está asociado con seis enfermedades mendelianas diferentes y encontramos 75 loci asociados a estas.

## Discusión y Conclusiones

En la realización de este análisis utilizamos varias herramientas bioinformáticas que fueron de gran utilidad, sin lugar a dudas, el desarrollo de nuevas técnicas ha permitido que los avances en el área de la investigación sean más frecuentes que nunca. Una de las herramientas más destacables son las bases de datos, ya que sin ellas, este y muchos otros trabajos de investigación serían casi imposibles de realizar.

En cuanto a los resultados obtenidos, el más interesante es la relación de mutaciones generadoras de enfermedades monogénicas en regiones exónicas y regiones intrónicas, pues hay una disparidad impresionante, siendo aproximadamente 80% más común encontrar una en regiones exónicas. Esto tiene sentido ya que ahí es donde podemos encontrar las secuencias que expresan proteínas, a diferencia de las regiones intrónicas, que comúnmente son asociadas a elementos regulatorios. Sin embargo, es posible que el número de mutaciones en regiones intrónicas sea mayor, o incluso menor, debido a que trabajamos en un área que está constantemente evolucionando y aprendiendo, y los datos disponibles en unos años podrían contar una historia totalmente diferente a la que vemos hoy.

Los datos encontrados sobre los tipos de mutaciones y sus respectivas frecuencias tienen bastante sentido matemáticamente hablando. Las mutaciones puntuales son más comunes porque el efecto de las mismas es mucho menor al de una inserción o delección, por ejemplo, las mutaciones puntuales pueden resultar en sinónimos, donde el aminoácido producido es el mismo, también pueden generar pérdidas de función. Por otro lado, las inserciones y delecciones tienen un efecto mucho mayor en las funciones del gen, llegando a no sólo una pérdida de función, sino a perder la viabilidad del organismo, haciendo que sea menos común encontrar una.

Todo el código utilizado puede ser consultado en:

<https://github.com/MichMourra/HumanGenomicsProject>

## Referencias:

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., & Majounie, E. et al. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain*, 138(12), 3673-3684. doi: 10.1093/brain/awv268

Horikawa, Y. (2018). Maturity-onset diabetes of the young as a model for elucidating the multifactorial origin of type 2 diabetes mellitus. *Journal Of Diabetes Investigation*, 9(4), 704-712. doi: 10.1111/jdi.12812

Makalowski, W. (2001). The human genome structure and organization. *Acta Biochimica Polonica*, 48(3), 587-598. Retrieved from [https://www.researchgate.net/publication/11526566\\_The\\_human\\_genome\\_structure\\_and\\_organization/download](https://www.researchgate.net/publication/11526566_The_human_genome_structure_and_organization/download)

OMIM Gene Map Statistics. (2021). Retrieved 25 April 2021, from <https://www.omim.org/statistics/geneMap>

Peltonen, L., Perola, M., Naukkarinen, J., & Palotie, A. (2006). Lessons from studying monogenic disease for common disease. *Human Molecular Genetics*, 15(suppl\_1), R67-R74. <https://doi.org/10.1093/hmg/ddl060>

"Search results - chromosome name[CHR] AND "Homo sapiens"[Organism] AND ("has ccds"[Properties] AND alive[prop]) - Gene". NCBI. CCDS Release 20 for *Homo sapiens*. Retrieved 2021-06-07.

Thaventhiran, J., Lango Allen, H., Burren, O., Rae, W., Greene, D., & Staples, E. et al. (2020). Whole-genome sequencing of a sporadic primary immunodeficiency cohort. *Nature*, 583(7814), 90-95. doi: 10.1038/s41586-020-2265-1

Valente, E., & Bhatia, K. (2018). Solving Mendelian Mysteries: The Non-coding Genome May Hold the Key. *Cell*, 172(5), 889-891. <https://doi.org/10.1016/j.cell.2018.02.022>

Wang, K. (2015). ANNOVAR website. Retrieved 25 April 2021, from  
[http://www.openbioinformatics.org/annovar/annovar\\_input.html](http://www.openbioinformatics.org/annovar/annovar_input.html)

Yang H, Wang K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR Nature Protocols, 10:1556-1566