PEC 3 ANALISIS DE DATOS OMICOS

Autor: Diego Vázquez Zambrano

Contenido

1. Resumen

Este estudio se enfoca en la evaluación de una variante genómica específica mediante técnicas de secuenciación de nueva generación y análisis bioinformático. Utilizando el genoma de referencia humano hg38, se llevó a cabo el alineamiento de secuencias de lectura de la variante en estudio. El objetivo principal fue identificar y caracterizar las diferencias entre la variante y el genoma de referencia, con un enfoque particular en el cromosoma 1, donde se observó un número significativo de alineaciones.

Se emplearon herramientas de análisis, como Samtools y programas de visualización como IGB, para evaluar la calidad del alineamiento y examinar las regiones donde la variante presenta diferencias respecto al genoma de referencia. Las variantes detectadas fueron posteriormente clasificadas en diferentes tipos, incluyendo polimorfismos de nucleótido único (SNP), inserciones (INS), y deleciones (DEL), y se evaluó su impacto potencial en las funciones genéticas utilizando SnpEff.

Los resultados proporcionan una visión detallada de la variante genómica en estudio, revelando información sobre las diferencias en la secuencia de nucleótidos y su potencial efecto en la función de genes y proteínas.

2. Objetivo del estudio

El objetivo principal de este estudio es utilizar técnicas de secuenciación de nueva generación y herramientas bioinformáticas para identificar y caracterizar variantes genéticas presentes en una muestra de ADN humano. Este análisis incluye el alineamiento de las secuencias de lecturas con un genoma de referencia, la identificación de diferencias entre las secuencias alineadas y el genoma de referencia, y la evaluación del impacto potencial de las variantes identificadas en la función genética. Los resultados podrían tener implicaciones para la comprensión de la biología humana y, potencialmente, para el diagnóstico y tratamiento de enfermedades genéticas.

3. Alineamiento de secuencias de lecturas con el genoma de referencia (HG38)

En este caso se ha decidido utilizar como referencia el genoma hg38. Hemos obtenido el siguiente resultado en el Samtools despues de realizar el mapeo y el Sortsam correspondientes.

Column 1	Column 2	Column 3
chr1	248956422	214754
chr1_KI270706v1_random	175055	36
chr1_KI270707v1_random	32032	0

Figura 1. Tabla Samtools

Como se puede ver en la Figura 1, tenemos 3 columnas que se refieren a lo siguiente:

La columna 1 contiene el nombre de los contigs o segmentos cromosómicos.

La columna 2 representa la longitud total del contig, es decir, el número total de bases.

La columna 3 es el número total de reads alineadas o mapeadas a cada contig. Es decir, cuantas secuencias de la muestra (nuestra variante) se alinearon con esa región en particular.

4. Análisis de las diferencias entre las secuencias alineadas y el genoma de referencia

Después de revisar todos los cromosomas, hemos observado que el cromosoma 1 es el que más coincidencias ha tenido con un total de 214754 reads. Por lo tanto, en esta práctica nos vamos a centrar en esta región para llevar a cabo el análisis.

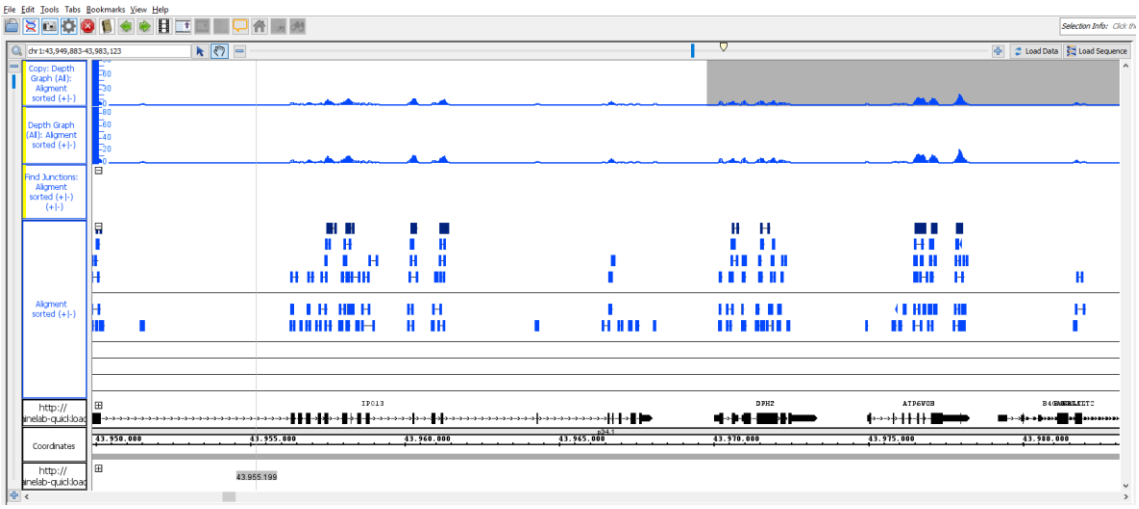


Figura 2. Recorte del software IGB en una región específica del cromosoma 1

Para llevar a cabo un análisis visual de los datos se ha utilizado el programa IGB ya que IGV daba problemas al introducir los datos. Este programa no nos permite hacer una visualización tan limpia como IGV, pero hemos podido observar utilizando la función Depth Graph que existen varias coincidencias, así como aparecía en la tabla SortSam. Con una imagen global

podemos observar mejor estas coincidencias:

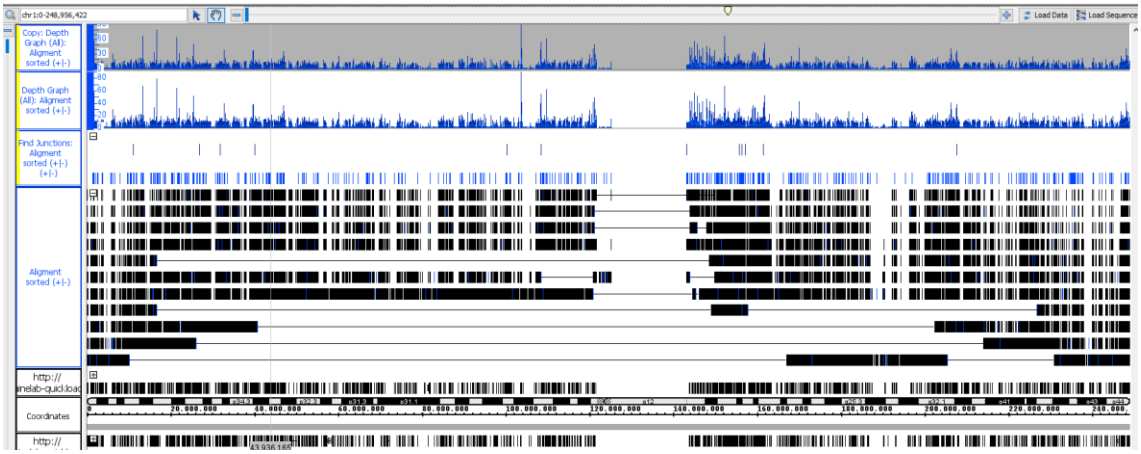


Figura 3. Recorte del software IGB donde se visualiza todo el alineamiento con el genoma de referencia en el cromosoma 1

Luego realizamos el Pileup para valorar la calidad de nuestro alineamiento.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
chr1	14564	G	G	30	0	0	1	^!,	@
chr1	14565	C	C	30	0	0	1	.	@
chr1	14566	T	T	30	0	0	1	.	@
chr1	14567	G	G	30	0	0	1	.	F
chr1	14568	G	G	30	0	0	1	.	F
chr1	14569	T	T	30	0	0	1	.	E
chr1	14570	C	C	30	0	0	1	.	F
chr1	14571	T	T	30	0	0	1	.	F
chr1	14572	C	C	30	0	0	1	.	H
chr1	14573	C	C	30	0	0	1	.	G
chr1	14574	A	A	30	0	0	1	.	D
chr1	14575	C	C	30	0	0	1	.	H
chr1	14576	A	A	30	0	0	1	.	H
chr1	14577	C	C	30	0	0	1	.	I
chr1	14578	A	A	30	0	0	1	.	I
chr1	14579	G	G	30	0	0	1	.	H
chr1	14580	T	T	30	0	0	1	.	F
chr1	14581	G	G	30	0	0	1	.	H
chr1	14582	C	C	30	0	0	1	.	I
chr1	14583	T	T	30	0	0	1	.	J
chr1	14584	G	G	30	0	0	1	.	J
chr1	14585	G	G	30	0	0	1	.	I
chr1	14586	T	T	30	0	0	1	.	I
chr1	14587	T	T	30	0	0	1	.	I
chr1	14588	C	C	30	0	0	1	.	I
chr1	14589	C	C	30	0	0	1	.	I

Figura 4. Pileup

Observando la Figura 4, vemos que la calidad del mapeo y la calidad de la base son altas en la mayoría de posiciones (calidad del consenso = 30).

Las columnas de la tabla Pileup significan lo siguiente:

Columna 1: nombre de la secuencia (“chr1”)

Columna 2: posición

Columna 3: bases de referencia en esa posición

Columna 4: bases de consenso

Columna 5: calidad del consenso

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
chr1	17315	T	T	47	0	0	10	......A,,,	CCBICI6BDD
chr1	17385	G	G	36	0	0	10	.Aa,,,,,a,	B63JJ6DB6B
chr1	35198	T	T	47	0	0	10	,,,A,,.,,,	FHJ8JJ@EF@
chr1	69318	T	T	62	0	0	18	...c.,,.,,,,.,,.,	DDE:FIJHGIIIJHFIDD
chr1	69799	A	A	81	0	1	57	,,,,,,.,,.,.,,.,,,,,,.,,,,,,G.,,,,.,.,,.,,,,,,,,,,,,,,,,..	CFFFFHDGIEEJ4FJIIJJ+HJGJJH#
chr1	69932	C	C	52	0	0	22	.t,..,.,.,T,.,.,t.,..^!,	EHIII??IGJGGJIE7HCE?F;
chr1	826705	G	R	8	8	52	18	..,Aa...,.,.,,.,,.	##JJHJGJJJIJFDJ#B@
chr1	826729	C	C	47	0	52	16	t...,.,.,.,,.,,..,	:FGJJJIEIJJ#DEJ@
chr1	980060	A	A	45	0	54	12	..,.,g,,...,	##J#G95BIJHD
chr1	1452384	G	R	59	88	58	11	aAAA.aaa,A.	4888A688?<@
chr1	1459247	G	G	9	0	37	10	.$,a,,aa,,,	>F5JJ9:DDD
chr1	1486536	T	T	34	0	33	19	.,C.c.,..........^!.^!.	CJ7J=IDJJJJIJIHFCBB
chr1	1486604	A	A	71	0	15	30	......C..CC,,C,,,,C,,,,,,,..,,,	5#&&&&&&0#5IJ#JEFC&CCC(CC
chr1	1486611	G	G	16	0	18	30	.A.AA..A.A,,Taaa,.,,a,aa..aaa,	#0<3#09-##GJ#344I5C?:C>:JJ9
chr1	1516000	T	Y	61	61	58	11	C.,CccC,c,c	5DJ84;/D5D5
chr1	1524202	G	R	9	9	39	11	A..AA.A.Aaa	:??><B:J690
chr1	1524230	G	G	26	0	47	14	AA.A..aa,,,,.,	;<D>DB=?@DDDFD
chr1	1524244	C	C	45	0	55	12	..tt,,,,.,,.	CD8:CCDDJDDF
chr1	1524340	G	G	50	0	29	15	...,,,,,.,,,A,^!a	##<J#GJF7DDD5D(
chr1	1524385	C	C	33	0	27	12	,.,,,.,,t.,^!.	H#JJJFDD<HD?
chr1	1574492	C	Y	12	12	59	10	Tt...,.,,,	?EDC#BIDD9
chr1	1641157	A	A	83	0	22	30	,...,G.,.,,,.g,,,.,,,,.,.,,,,^Y,	CDBDH5DJDIBA9;IBIDIIGGBG:)F(
chr1	1641946	G	G	58	0	0	21	,$.,.....t..,,.,.,,,,,.	@BG8<#BBIHJHGJICDCDDF
chr1	1657220	G	R	123	123	55	21	.a..,A.A..,a..,a,Aa,,	EHIJJJJIJIGBIJHHDJEBD
chr1	1657266	T	T	66	0	50	20	,..,,c.,,,,,.,,,,...	FFFJJ6JIJJJJGJIFEHFF

Figura 5. Filter pileup

Según los resultados obtenidos en la Figura 5 vemos que, la cobertura, que es el número de lecturas que se alinean en cada posición, es relativamente alta en todas las posiciones. Esto es importante porque una mayor cobertura aumenta la confianza en las observaciones realizadas en cada posición. Además, la calidad de las bases (columna 5), que es una medida de la confianza en la identificación de cada base, también es alta en la mayoría de las posiciones. Esto sugiere que las bases se han identificado con precisión en las lecturas. Finalmente, la concordancia entre el nucleótido de referencia y el nucleótido más comúnmente observado en las lecturas alineadas en cada posición también sugiere que el alineamiento es preciso.

5. Variant calling y análisis de variantes

Observamos el archivo VCF desde el programa IGV

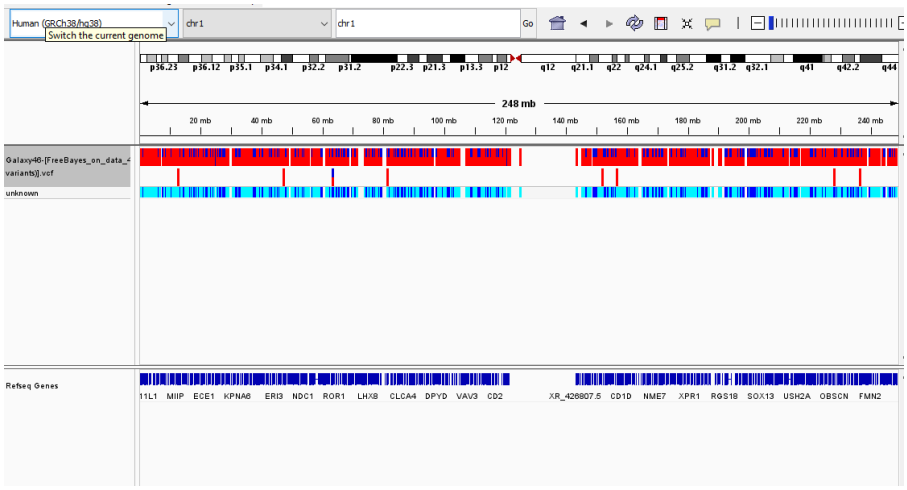


Figura 6. Recorte del software IGV donde se visualiza el archivo VCF obtenido tras utilizar la herramienta Free Bayes.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
chr1	189514	.	C	A	59.8681	.
chr1	826705	.	G	A	0.0322238	.
chr1	827209	.	GGCC	CGCG	144.117	.
chr1	827221	.	T	C	148.762	.
chr1	827252	.	T	A	118.932	.
chr1	914618	.	A	G	57.1766	.
chr1	930939	.	G	A	74.9004	.
chr1	941119	.	A	G	76.4741	.
chr1	946247	.	G	A	114.366	.
chr1	948245	.	A	G	117.304	.
chr1	952180	.	A	C	120.982	.
chr1	953259	.	T	C	185.889	.

Figura 7. Tabla obtenida tras realizar un filtrado en el archivo VCF obtenido por Free Bayes

Si observamos el valor QUAL en la Figura 7, vemos que en general tiene un valor alto (>30). Esto indica buena confianza en la variante. El valor QUAL representa la calidad de la llamada de la variante.

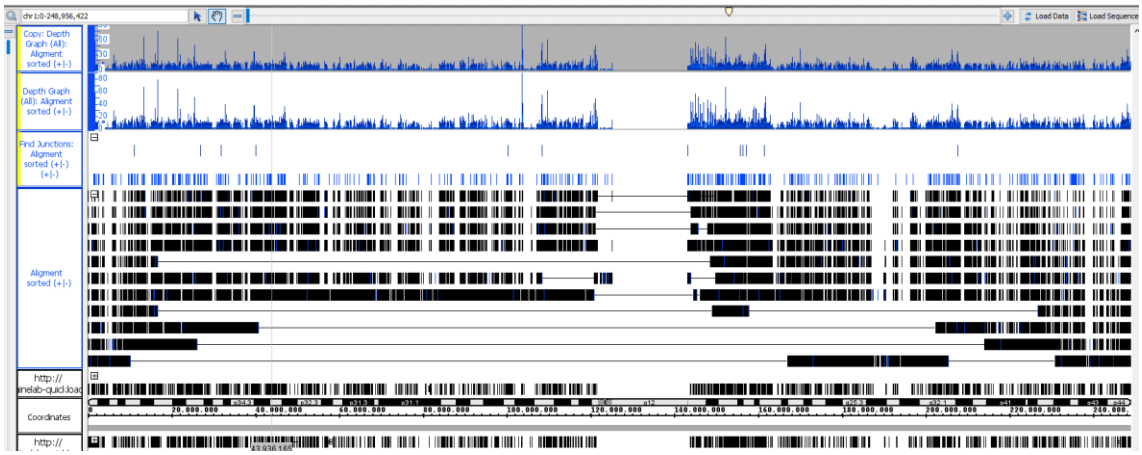


Figura 3. Recorte del software IGB donde se visualiza todo el alineamiento con el genoma de referencia

Volviendo a la Figura 3 podemos observar la relación entre ambas imágenes. El variant calling nos ha detectado una serie de variantes. Ambas imágenes son bastante similares.

Ahora vamos a coger una región específica para llevar a cabo el siguiente paso. Podemos observar las variantes detectadas en esta región específica del cromosoma 1

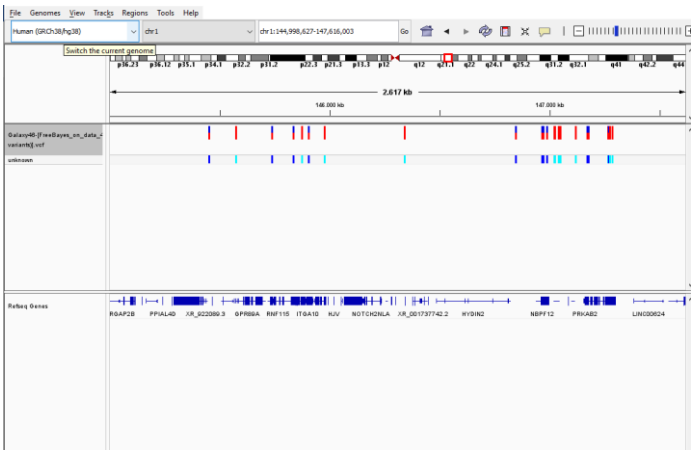


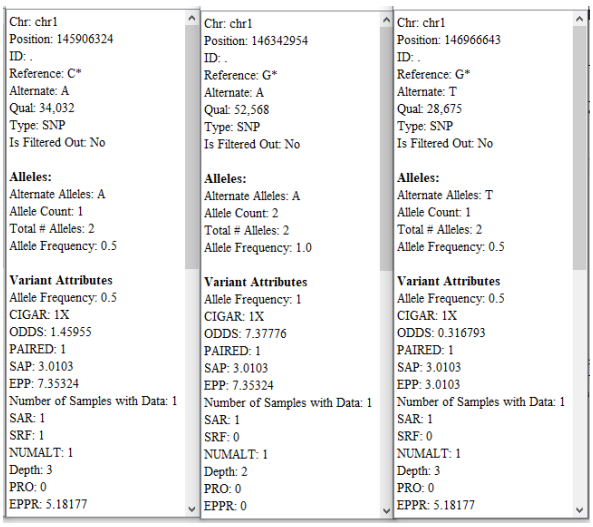
Figura 8. Recorte del software IGV donde se visualiza una región específica del alineamiento entre la variante y el genoma de referencia

Esta región es: chr1:144,998,627-147,616,003

Aquí podemos observar las variantes que se encuentran en esta región del cromosoma 1:

Chrom	Pos	ID	Ref	Alt	Qual	Fi
chr1	145454268	.	G	C	18.6296	.
chr1	145574212	.	C	G	114.928	.
chr1	145738392	.	C	T	63.217	.
chr1	145836695	.	A	G	54.0763	.
chr1	145837693	.	G	T	14.4873	.
chr1	145875024	.	T	C	0.11623	.
chr1	145876257	.	T	G	138.987	.
chr1	145876488	.	G	A	60.1041	.
chr1	145906324	.	C	A	34.0317	.

Figura 9. Tabla obtenida por la herramienta Free Bayes de la región específica



Chr: chr1
Position: 145906324
ID: .
Reference: C*
Alternate: A
Qual: 34,032
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: A
Allele Count: 1
Total # Alleles: 2
Allele Frequency: 0.5

Variant Attributes
Allele Frequency: 0.5
CIGAR: 1X
ODDS: 1.45955
PAIRED: 1
SAP: 3.0103
EPP: 7.35324
Number of Samples with Data: 1
SAR: 1
SRF: 1
NUMALT: 1
Depth: 3
PRO: 0
EPPR: 5.18177

Chr: chr1
Position: 146342954
ID: .
Reference: G*
Alternate: A
Qual: 52,568
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: A
Allele Count: 2
Total # Alleles: 2
Allele Frequency: 1.0

Variant Attributes
Allele Frequency: 1
CIGAR: 1X
ODDS: 7.37776
PAIRED: 1
SAP: 3.0103
EPP: 7.35324
Number of Samples with Data: 1
SAR: 1
SRF: 0
NUMALT: 1
Depth: 2
PRO: 0
EPPR: 0

Chr: chr1
Position: 146966643
ID: .
Reference: G*
Alternate: T
Qual: 28,675
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: T
Allele Count: 1
Total # Alleles: 2
Allele Frequency: 0.5

Variant Attributes
Allele Frequency: 0.5
CIGAR: 1X
ODDS: 0.316793
PAIRED: 1
SAP: 3.0103
EPP: 3.0103
Number of Samples with Data: 1
SAR: 1
SRF: 0
NUMALT: 1
Depth: 3
PRO: 0
EPPR: 5.18177

Figura 10. Recorte de algunas variantes filtradas en la región específica

Observamos algunas de las variantes filtradas.

6. Análisis funcional de variantes con SnpEff

Realizamos a continuación un examen SnpEff:

Genome	hg38
Date	2023-06-20 13:21
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /corral4/main/jobs/051/108/51108290/outputs/galaxy_dataset_905e884d-7d81-4934-aa29-69a033dd95af.dat hg38 /corral4/main/objects/e/7/3/dataset_e73cedd9-28b6-4c92-8988-4cf12888da07.dat
Warnings	0
Errors	0
Number of lines (input file)	29
Number of variants (before filter)	29
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	29
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of effects	60
Genome total length	3,209,286,106
Genome effective length	248,956,422
Variant rate	1 variant every 8,584,704 bases

Figura 11. Resumen del examen SnpEff

Number variants by type

Type	Total
SNP	25
MNP	1
INS	2
DEL	1
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	29

Figura 12. Número de variantes por tipo

En esta tabla se observa un resumen del número de variantes genéticas por tipo presentes en los datos. Los tipos de variantes son los siguientes:

SNP: Single Nucleotide Polymorphism, que es un cambio en un solo nucleótido en la secuencia del ADN.

MNP: Multiple Nucleotide Polymorphism, que es un cambio en más de un nucleótido en la secuencia de ADN.

INS: Insertion, que es la adición de uno o más nucleótidos en la secuencia del ADN.

DEL: Deletion, que es la eliminación de uno o más nucleótidos de la secuencia del ADN.

MIXED: Se refiere a variantes que tienen más de un tipo de cambio en una sola ubicación, por ejemplo, una inserción y una eliminación.

INV: Inversion, que es cuando un segmento de ADN se invierte de extremo a extremo.

DUP: Duplication, que es la repetición de uno o más nucleótidos en la secuencia del ADN.

BND: Breakend, que se refiere a eventos de reorganización estructural más complejos que implican roturas y uniones de diferentes regiones de ADN.

INTERVAL: Intervalo, no es un tipo de variante genética, sino que se refiere a un intervalo específico de ADN.

En los datos, tenemos 25 SNPs, 1 MNP, 2 inserciones y 1 deleción, y no tenemos ningún tipo de variantes mixtas, inversiones, duplicaciones, breakends o intervalos.

Base changes (SNPs)

	A	C	G	T
A	0	0	6	0
C	1	0	3	3
G	3	1	0	2
T	0	5	1	0

Figura 13. Cambios de bases

Los SNPs son variaciones en un solo nucleótido que ocurren en una posición específica en el genoma.

La tabla se lee de la siguiente manera:

En la fila "A", la base de referencia es "A". Hubo 0 cambios a "A", 0 cambios a "C", 6 cambios a "G", y 0 cambios a "T".

En la fila "C", la base de referencia es "C". Hubo 1 cambio a "A", 0 cambios a "C", 3 cambios a "G", y 3 cambios a "T".

En la fila "G", la base de referencia es "G". Hubo 3 cambios a "A", 1 cambio a "C", 0 cambios a "G", y 2 cambios a "T".

En la fila "T", la base de referencia es "T". Hubo 0 cambios a "A", 5 cambios a "C", 1 cambio a "G", y 0 cambios a "T".

Transitions	27
Transversions	10
Ts/Tv ratio	2.7

All variants:

```
Sample ,Total
Transitions ,27,27
Transversions ,10,10
Ts/Tv ,2.700,2.700
```

Figura 14. Contaje de transiciones y transversiones.

Las transiciones y las transversiones son dos categorías diferentes de SNPs (polimorfismos de un solo nucleótido) que pueden ocurrir en un genoma:

Una transición es un cambio de una base de purina a otra base de purina (A ↔ G) o de una pirimidina a otra pirimidina (C ↔ T). Por ejemplo, si una adenina (A) se cambia por una guanina (G), o si una citosina (C) se cambia por una timina (T), eso se considera una transición.

Una transversión es un cambio de una base de purina a una base de pirimidina o viceversa (A ↔ C, A ↔ T, G ↔ C, G ↔ T). Por ejemplo, si una adenina (A) se cambia por una citosina (C) o una timina (T), o si una guanina (G) se cambia por una citosina (C) o una timina (T), eso se considera una transversión.

En este caso:

Hay 27 transiciones (Ts) y 10 transversiones (Tv).

La relación Ts/Tv es de 2.7, lo que significa que hay 2.7 veces más transiciones que transversiones en este conjunto de datos.

La relación Ts/Tv es una métrica importante en la genómica. En genomas humanos normales, la relación Ts/Tv es generalmente de alrededor de 2 a 2.1 debido a que las transiciones son más propensas a ocurrir que las transversiones. Cuando la relación Ts/Tv es significativamente más baja o más alta que la esperada, puede indicar posibles errores de secuenciación o análisis, o puede ser una señal de un tipo particular de presión de selección o mutación.

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	1	1.667%
LOW	2	3.333%
MODERATE	8	13.333%
MODIFIER	49	81.667%

Figura 15. Número de efectos por impacto

Esta tabla nos proporciona una clasificación de las variantes genéticas según su impacto potencial en las funciones genéticas. Las categorías significan lo siguiente:

High: variantes con impacto alto que tienen un efecto disruptivo en la proteína, por ejemplo, cambios de sentido erróneo que generan un codón de parada prematuro. Estas variantes pueden causar perdida de función de la proteína.

Moderate: las variantes con impacto moderado son aquellas que podrían cambiar la efectividad de la proteína, pero no necesariamente la desactivan por completo. Ejemplos de esto podrían ser variantes missense (un cambio de un solo nucleótido que resulta en la codificación de un aminoácido diferente).

Low: las variantes de impacto bajo son aquellas que se espera que tengan poco o ningún efecto en la función de la proteína.

Modifier: las variantes con impacto modificador son aquellas que se encuentran en regiones no codificantes del genoma o que afectan a regiones no codificantes de un transcrito. No se espera que estas variantes cambien la secuencia de aminoácidos de la proteína, pero pueden afectar la regulación del gen o la estabilidad del ARN mensajero.

#GeneName	GeneId	TranscriptId	BioType	variants_impact_HIGH	variants_impact_LOW	variants_impact_MODERATE	variants_impact_MODIFIER	variants_effect_downstream_gene_variant	variants_effect_frameshift_variant	variants_effect_intron_variant	variants_effect_missense_variant	variants_effect_non_coding_transcript_exon_variant	variants_effect_splice_region_variant
ANKRD35	ANKRD35	NM_001280799.1	protein_coding	0	0	3	0	0	3	0	0	0	
ANKRD35	ANKRD35	NM_144698.4	protein_coding	0	0	3	0	0	3	0	0	0	
CD160	CD160	NM_007053.3	protein_coding	0	0	1	0	0	1	0	0	0	
CD160	CD160	NR_103845.1		0	0	1	0	0	1	0	0	0	
CHD1L	CHD1L	NM_001256336.1	protein_coding	0	0	1	3	0	0	3	1	0	0
CHD1L	CHD1L	NM_001256337.1	protein_coding	0	0	1	3	0	0	3	1	0	0
CHD1L	CHD1L	NM_001256338.1	protein_coding	0	0	1	3	0	0	3	1	0	0
CHD1L	CHD1L	NM_004284.4	protein_coding	0	0	1	3	0	0	3	1	0	0
CHD1L	CHD1L	NM_024568.2	protein_coding	0	0	1	3	0	0	3	1	0	0
CHD1L	CHD1L	NR_046070.1		0	0	0	4	0	0	3	0	1	0
FMO5	FMO5	NM_001144829.2	protein_coding	0	0	0	1	1	0	0	0	0	0
ITGA10	ITGA10	NM_001303040.1	protein_coding	0	0	0	1	0	0	1	0	0	0
ITGA10	ITGA10	NM_001303041.1	protein_coding	0	0	0	1	0	0	1	0	0	0
ITGA10	ITGA10	NM_003637.4	protein_coding	0	0	0	1	0	0	1	0	0	0
LOC728989		LOC728989		NR_024442.2		0	0	0	3	0	0	0	3	0
NBPF12	NBPF12	NM_001278141.1	protein_coding	1	1	3	0	0	1	1	3	0	1
NBPF25P	NBPF25P	NR_104217.1		0	1	0	0	0	1	0	0	1	
PDIA3P1	PDIA3P1	NR_002305.1		0	0	0	1	0	0	0	0	1	0
POLR3C	POLR3C	NM_001303456.1	protein_coding	0	0	0	2	0	0	2	0	0	0
POLR3C	POLR3C	NM_006468.7	protein_coding	0	0	0	2	0	0	2	0	0	0
POLR3GL	POLR3GL	NM_032305.1	protein_coding	0	0	0	1	0	0	1	0	0	0
PRKAB2	PRKAB2	NM_005399.4	protein_coding	0	0	0	1	0	0	1	0	0	0
PRKAB2	PRKAB2	NR_103870.1		0	0	0	1	0	0	1	0	0	0
PRKAB2	PRKAB2	NR_103871.1		0	0	0	1	0	0	1	0	0	0

Figura 16. Informe generado por SnpEff sobre las variantes genéticas y sus efectos potenciales en genes y transcritos específicos

Podemos observar el número de variantes en cada gen transcrito que tienen un impacto alto, bajo, moderado o modificador, según la predicción de SnpEff.

Por ejemplo, centrémonos en el gen NBPF12. Este gen pertenece a la familia de genes de ruptura del neuroblastoma (NBPF), la cual está compuesta por múltiples genes duplicados que se encuentran principalmente en duplicaciones segmentales en el cromosoma 1 humano. Este es un gen que codifica una proteína y está relacionado con enfermedades como neuroblastomas o Síndrome de Hydrolethalus 2.

Según la Figura 16, si observamos esta proteína, podemos ver que este gen que codifica una proteína tiene:

- 1 variante con impacto alto
- 1 variante con impacto bajo
- 3 variantes con impacto moderado
- 1 variante con efecto intron_variant
- 1 variante con efecto missense_variant
- 3 variantes con efecto non_coding_transcript_exon_variant

7. Discusión

En este estudio, se analizó una variante genómica mediante alineamiento de secuencias de lectura contra el genoma de referencia humano hg38. Utilizando herramientas como Samtools y programas de visualización como IGB, se pudo identificar y caracterizar las diferencias entre la variante y el genoma de referencia, enfocándose particularmente en el cromosoma 1.

Se categorizaron varias variantes detectadas, incluyendo polimorfismos de nucleótido único (SNP), inserciones (INS) y deleciones (DEL), y se evaluó su impacto potencial en la función de genes y proteínas mediante SnpEff. Además, se realizó un análisis de la calidad del alineamiento, lo que permitió evaluar la confiabilidad de los datos obtenidos.

Sin embargo, el estudio presenta algunas limitaciones. En primer lugar, el análisis se centró principalmente en el cromosoma 1, lo que significa que podrían existir otras diferencias significativas en otros cromosomas que no se han examinado en detalle. Además, la interpretación de las variantes y su impacto en la función genética requiere un conocimiento detallado de la biología del organismo, y la correlación entre las variantes genéticas y su función fenotípica puede ser compleja.

En cuanto a las conclusiones, se pudo identificar un conjunto de variantes en la secuencia en estudio con respecto al genoma de referencia. Esto sienta las bases para futuras investigaciones que podrían enfocarse en entender el papel funcional de estas variantes, y cómo pueden estar involucradas en características fenotípicas o susceptibilidad a enfermedades.

Es importante destacar que la secuenciación y el análisis de variantes genómicas son campos en constante evolución, y futuras investigaciones podrían beneficiarse de técnicas y herramientas más avanzadas para un análisis más profundo y preciso. Además, la colaboración con expertos en genética y biología molecular podría enriquecer la interpretación de los datos y llevar a descubrimientos más significativos.

En definitiva, este estudio representa un paso inicial en el análisis de una variante genómica específica y destaca la importancia de la metodología bioinformática en la comprensión de la diversidad genética.