Análisis de variantes minoritarias con Galaxy.

Tabla de contenido

Abstract.

Este estudio realiza un análisis genómico de variantes minoritarias (SNVs e indels) en el genoma humano utilizando datos del Proyecto de los 1000 Genomas y el genoma de referencia GRCh37. A través de la plataforma Galaxy, se implementó un pipeline bioinformático que incluye la carga y preprocesamiento de datos, control de calidad, alineación al genoma de referencia, identificación de variantes genéticas, y su anotación funcional.

Los resultados destacan una alta calidad del mapeo, con un 99.8% de lecturas alineadas correctamente y una razón Ts/Tv de 2.2737, característica de datos genómicos robustos. Se identificaron 27,662 SNPs y 1,601 indels, con predominancia de transiciones sobre

transversiones. Las variantes Missense (50.805%) y Silent (47.928%) fueron las más frecuentes, mientras que las variantes Nonsense y Frameshift, aunque menos comunes, tienen implicaciones funcionales significativas. La mayoría de las variantes se localizaron en regiones intergénicas e intrónicas, reflejando patrones evolutivos conocidos.

El estudio confirma la eficacia del pipeline empleado, aunque reconoce limitaciones como la baja cobertura en ciertas regiones genómicas y el uso de una versión del genoma de referencia obsoleta. Estos factores podrían mejorarse mediante técnicas de secuenciación de mayor profundidad y actualización del genoma de referencia. A pesar de estas limitaciones, los hallazgos proporcionan información valiosa sobre las variantes genéticas humanas y sus potenciales efectos funcionales.

Objetivos del estudio.

El presente estudio pretende realizar un análisis genómico de variantes minoritarias-pequeñas, como SNVs (Single Nucleotide Variants) e indels (inserciones y deleciones) en el genoma humano, utilizando como referencia una pequeña muestra de datos genómicos del Proyecto de los 1000 genomas y comparándolos con GRCh37.

Materiales y métodos.

Los datos seleccionados en el estudio corresponden a sampleData8_1.fq y sampleData8_2.fq, correspondientes a la muestra HG00128 del Proyecto de los 1000 Genomas. Estos datos consisten en lecturas apareadas (*Paired End reads*).

Así pues, todo el análisis se realizó a través de la plataforma bioinformática Galaxy, en específico en la instancia de Europa.

Así pues, el análisis ha sido realizado a través del siguiente pipeline:

1. Carga y preprocesamiento de los datos.

Una vez se accede a Galaxy, es necesario generar un nuevo Historial:

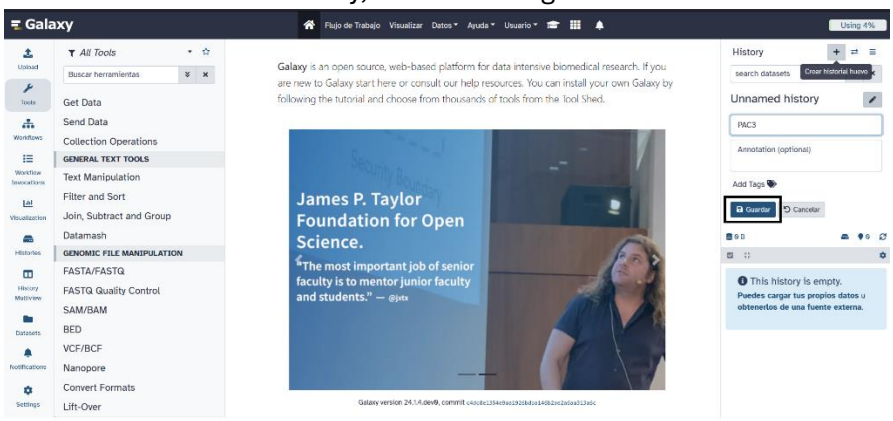

Ilustración 1: Generación de nuevo historial en Galaxy.

Tras ello, se podrán cargar los datos. En este caso, al estar descargados en el archivo local, se debe seleccionar "*Upload File from your computer*", dentro de *Get Data,* en el panel izquierdo de Galaxy. Tras cargarlos, se renombrarán para facilitar su identificación:

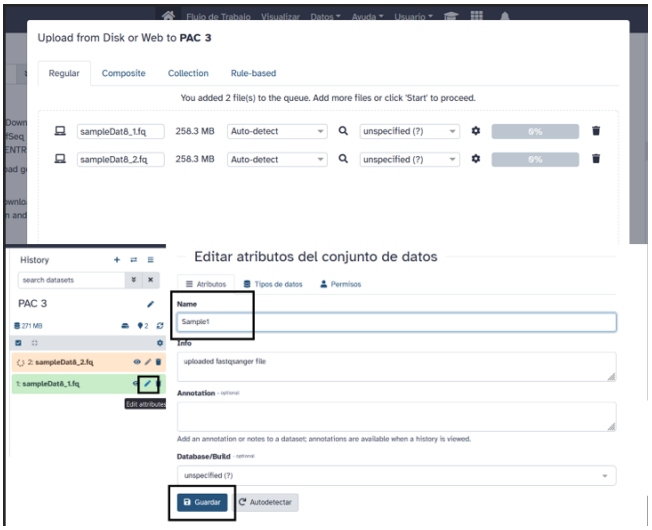

Ilustración 2: Carga de archivos y renombre.

2. Evaluación y control de calidad de las lecturas.

Desde el panel lateral izquierdo de Galaxy, en *Genomic File Manipulation / FASTQ Quality Control / FastQC Read Quality reports*, se realizarán dos análisis de calidad: uno por cada muestra.

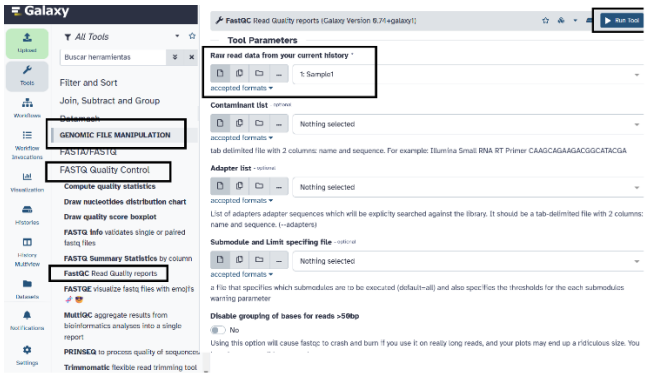

Ilustración 3: Configuración de FastQC Read Quality reports.

3. Alineación de las secuencias al genoma de referencia.

Para realizar la alineación respecto al genoma de referencia, se escoge hg19, también conocido como GRCh37. Para ello, se genera un archivo BAM con los metadatos de dichas lecturas desde el panel lateral izquierdo de Galaxy en *Genomics Analysis / Mapping / Map with BWA-MEM*. En la configuración de esta herramienta es indicar que estamos trabajando con datos *Paired-end reads*, para poder seleccionar ambos fragmentos.

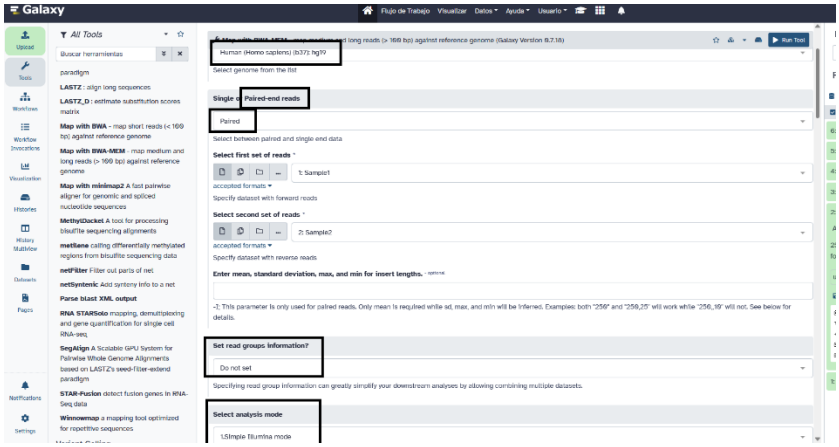

Ilustración 4: Configuración de Map with BWA-MEM.

El siguiente paso es generar un listado de lecturas mapeadas al genoma de referencia. Para ello, el primer paso es generar un archivo con las alineaciones ordenadas por coordenadas, a través de la herramienta *SortSam*:

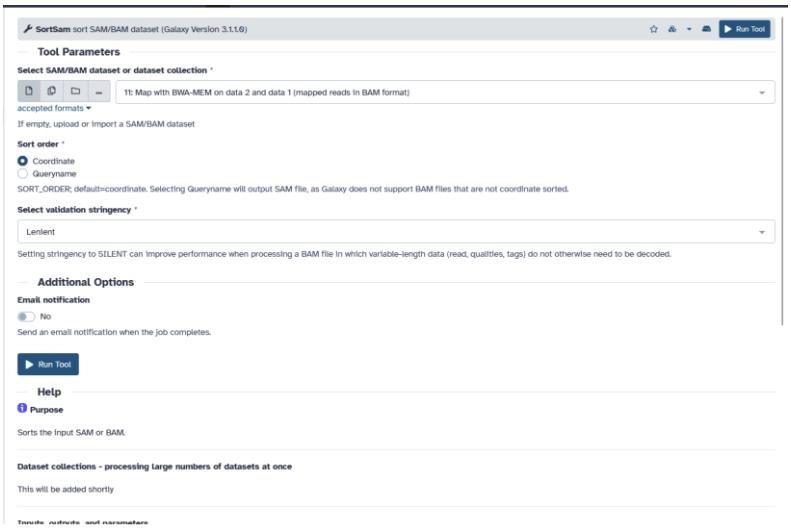

Ilustración 5: Configuración de la herramienta SortSam.

Tras esto, se utiliza el archivo generado BAM generado para ejecutar la herramienta *Samtools idxstats* para obtener el cromosoma, el número de lecturas mapeadas y el número de lecturas no mapeadas por cada referencia en el archivo de alineación:

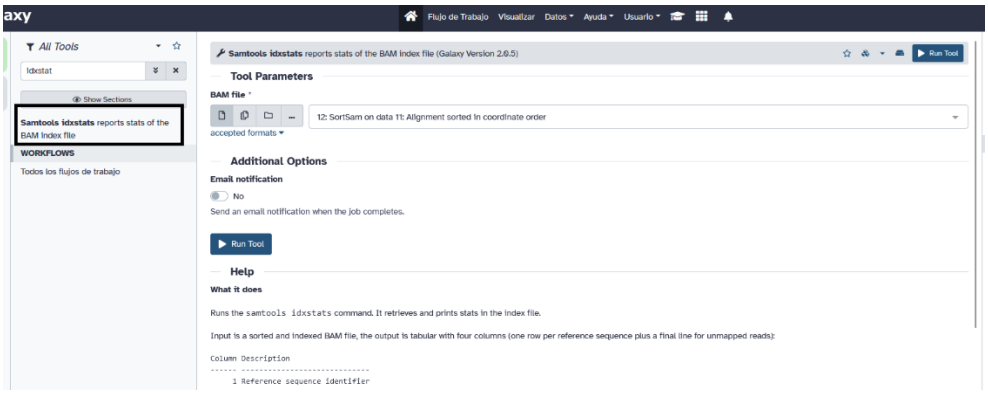

Ilustración 6: Configuración de Samtools idxstats.

Tras ello, se utilizará la herramienta *Samtools flagstat* sobre el archivo BAM para generar un resumen estadístico breve, incluyendo la calidad del alineamiento, la totalidad de lecturas mapeadas, así como los duplicados:

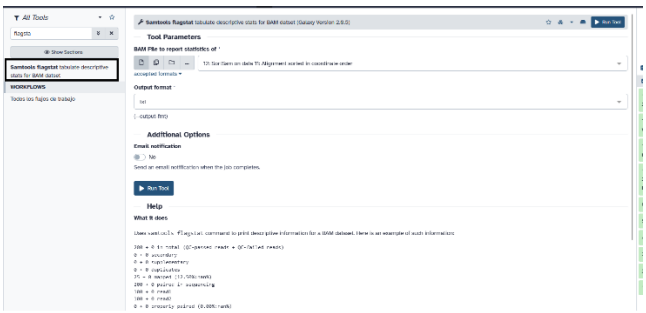

Ilustración 7: Configuración de la herramienta Samtools flagstat.

4. Identificación de diferencias genéticas entre las lecturas alineadas y el genoma de referencia.

El siguiente paso es generar un pileup con la herramienta *Generate pileup from BAM dataset*. Esto es útil para visualizar las variantes de la secuencia, calcular la cobertura de las lecturas e identificar variantes, como las que son de nuestro interés (SNPs o indels):

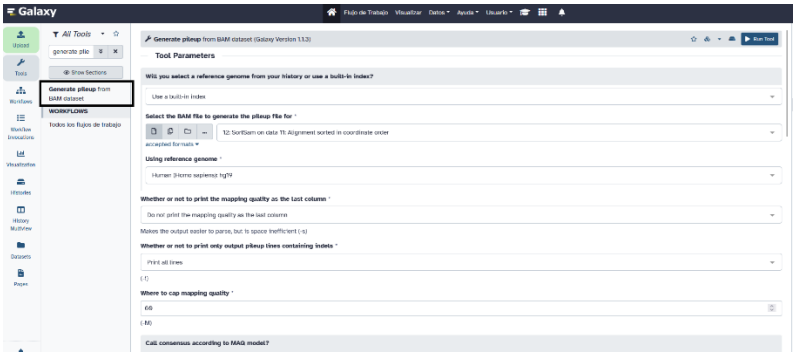

Ilustración 8: Configuración de la herramienta de Pileup.

Para poder seguir con el análisis, es importante cambiar el tipo del archivo a pileup una vez generado, editando este:

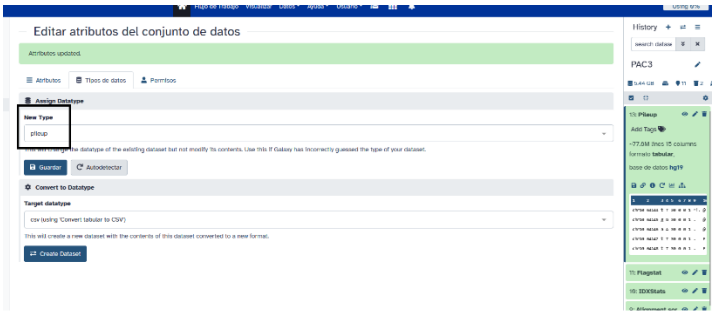

Ilustración 9: Edición del archivo Pileup.

Tras obtener el archivo, filtraremos este con la herramienta *Filter pileup on coverage and SNPs*, descartando posiciones con una cobertura menor a 10 y también aquellas bases con una calidad menor a 20. Este filtrado se realizará dos veces, en una se reportarán solo las variantes y en otra no:

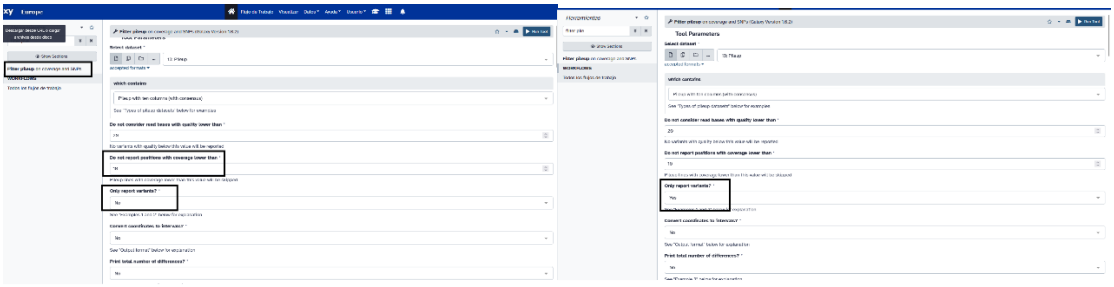

Ilustración 10: Filtrado del Pileup.

Para tener un listado de las variantes genéticas, se utilizará *FreeBayes bayesian genetic variant detector*, el cual genera una puntuación de calidad de variantes que se utilizarán posteriormente para el filtrado. Esta herramienta debe aplicarse sobre el archivo BAM de la alineación ordenada por coordenadas:

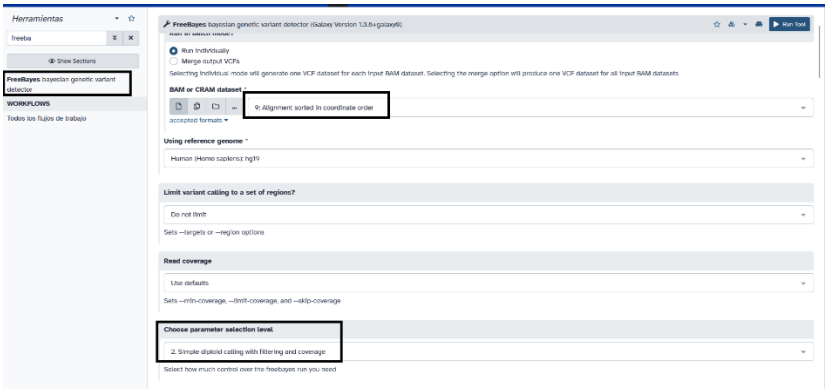

Ilustración 11: Aplicación de la herramienta FreeBayes.

Con este archivo es posible realizar una visualización de estas variables genéticas a través de UCSC:

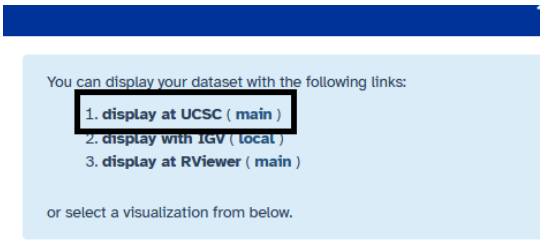

Ilustración 12: Herramientas para visualizar las variantes genéticas obtenidas por FreeBayes.

Tras obtener el archivo VCF correspondiente, es posible filtrarlo con la herramienta *Filter a VCF file* para descartar aquellas calidades menores a 50:

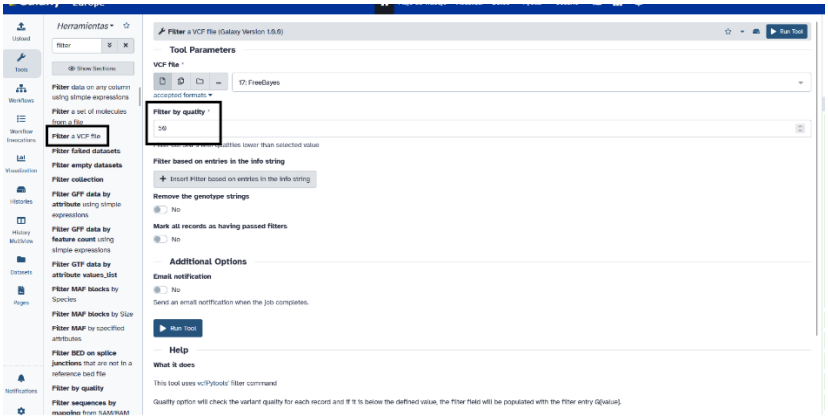

Ilustración 13: Configuración para filtrar FreeBayes.

5. Visualización de los resultados intermedios mediante un visor de genomas.

Para visualizar este alineamiento, es posible hacer clic en el icono de visualización del archivo BAM del alineamiento ordenado por coordenadas:

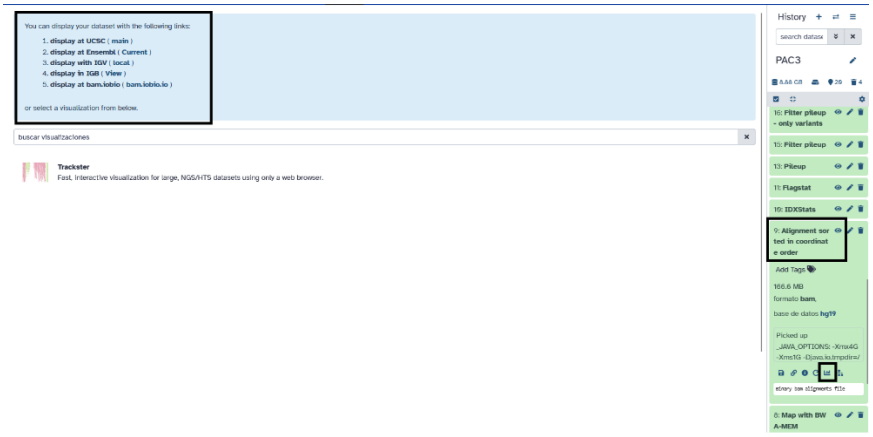

Ilustración 14: Herramientas de visualización del alineamiento.

Esta acción proporciona distintas fuentes de visualización del alineamiento en referencia al genoma de referencia.

En este caso, se visualizarán las tres siguientes herramientas:

- **UCSC Genome Browser:** para poder comparar la alineación directamente desde el navegador.
- **Integrative Genomics Viewer:** para realizar la revisión desde una aplicación de escritorio.
- **BAM.iobio:** para observar estadísticas como profundidad de cobertura, tasas de mapeo, etc. En este caso, será necesario hacer clic en la flecha dentro de las lecturas para ampliar el subconjunto analizado:

Ilustración 15: Ampliación del número de lecturas en BAM.iobio.

6. Filtrado y anotación de variantes genéticas.

Finalmente, se aplica la herramienta *SnpEff eff* sobre el archivo FreeBayes con tal de anotar las variantes genéticas con información procedente de datos de referencia conocidos, como por ejemplo si la variante corresponde a una previamente observada en otras muestras, si está dentro o cerca de un gen conocido, si se encuentra en un tipo concreto de región genómica o si se prevé que cause un efecto patogénico en el gen cuando mita:

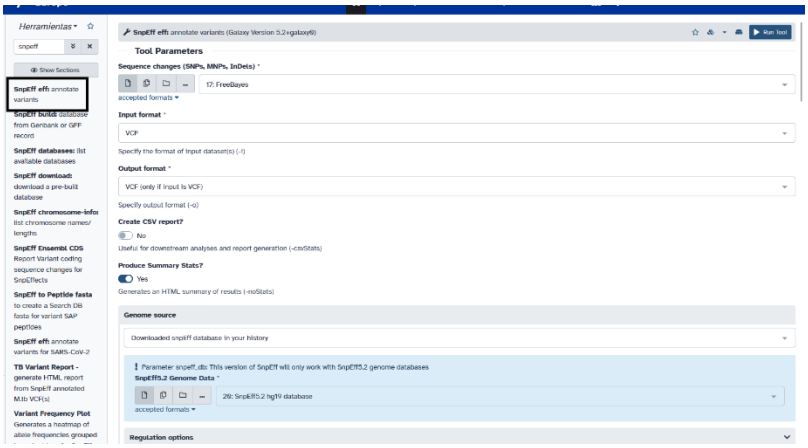

Ilustración 16: Configuración de la herramienta SnpEff.

Esta herramienta genera un informe en formato HTML con un resumen de las anotaciones aplicadas a las variantes observadas, como el tipo de variante, el impacto de estas, regiones funcionales, etc (Universidad de Melbourne, s.f.).

Resultados.

1. Evaluación y control de calidad de las lecturas.

Tras revisar los archivos Webpage generados, se puede proceder con el análisis, teniendo ambas muestras una buena calidad de la secuencia por base, característica imprescindible para el estudio de las variantes, ya que refleja la precisión de las lecturas de DNA en cada posición a lo largo de la secuencia:

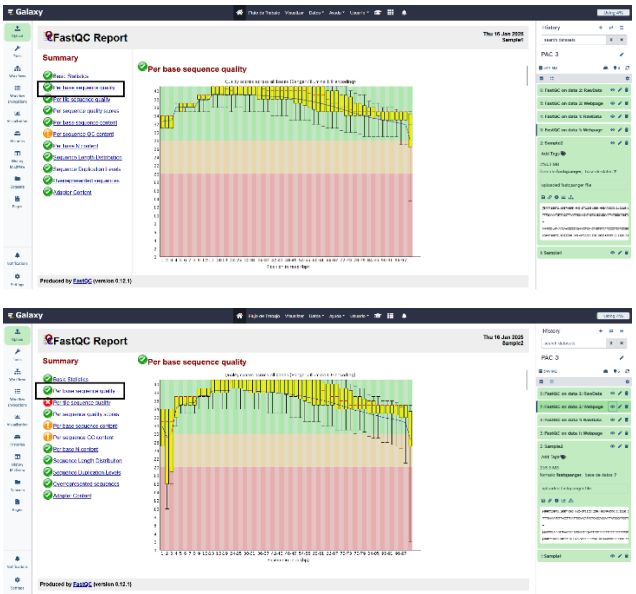

Ilustración 17: Análisis de calidad de las muestras.

2. Alineación de las secuencias al genoma de referencia.

Se obtiene un archivo BAM con las alineaciones ordenadas por las coordenadas. Este archivo es muy amplio, habiéndose ordenado primero las alineaciones correspondientes al cromosoma 10, pero estando todos los cromosomas presentes en este:

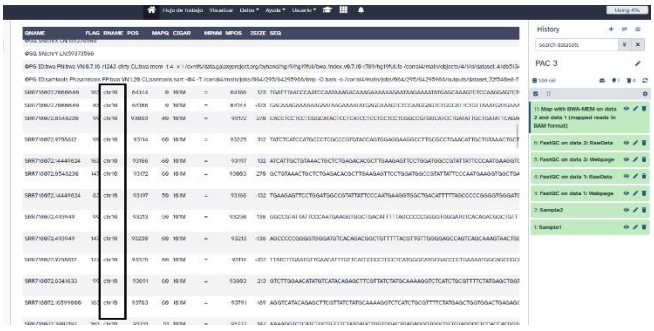

Ilustración 18: Primeras filas de las alineaciones ordenadas por coordenadas.

Se observan coincidencias en todos los cromosomas, siendo las más altas en el cromosoma1, cromosoma2 y cromosoma3, lo cual puede deberse al tamaño superior de estos. Así mismo, las secuencias de tipo “random”, que no han sido asignadas a una región específica del cromosoma, tienen pocas coincidencias:

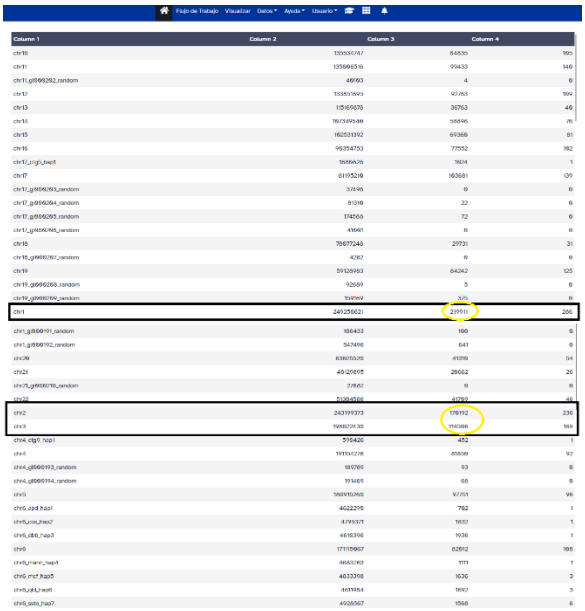

Ilustración 19: Coincidencias en la alineación por cromosomas.

En cuanto los datos de Flagstat, se obtiene:

```
2000463 + 0 in total (QC-passed reads + QC-failed reads)
2000000 + 0 primary
0 + 0 secondary
463 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1995924 + 0 mapped (99.77% : N/A)
1995461 + 0 primary mapped (99.77% : N/A)
2000000 + 0 paired in sequencing
1000000 + 0 read1
1000000 + 0 read2
1978722 + 0 properly paired (98.94% : N/A)
1992930 + 0 with itself and mate mapped
2531 + 0 singletons (0.13% : N/A)
2388 + 0 with mate mapped to a different chr
1573 + 0 with mate mapped to a different chr (mapQ>=5)
```

Ilustración 20: Resultados de las métricas de Flagstat.

Con esto es posible asegurar la alta calidad del mapeo, teniendo un 99.77% de las lecturas mapeadas. También se observa que un 98.94% de las lecturas emparejadas están correctamente alineadas. Así mismo, hay pocas lecturas mapeadas a diferentes cromosomas.

3. Identificación de diferencias genéticas entre las lecturas alineadas y el genoma de referencia.

Con la generación del Pileup es posible obtener una representación en columnas de las lecturas alineadas con el genoma hg19. Este archivo permite evaluar la cobertura, calidad y diferencias genéticas en cada posición. Así mismo, se obtiene un archivo con 10 columnas, representando:

- **Columna 1:** Cromosoma donde se encuentra la posición.
- **Columna 2:** Posición en el genoma de referencia.
- **Columna 3:** Base de referencia en esa posición.
- **Columna 4:** Base consenso observada en las lecturas.
- **Columna 5:** Número de lecturas alineadas que cubren esa posición (cobertura).
- **Columna 6-8:** Datos sobre variantes específicas (ejemplo: calidad de inserciones/deleciones).
- **Columna 9:** Bases observadas en las lecturas alineadas.
- **Columna 10:** Calidad de las bases observadas (Phred score).

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Colu
chr19	64144	t	T	36	0	0	1	^L	@	
chr19	64145	g	G	36	0	0	1	.	@	
chr19	64146	a	A	36	0	0	1	.	@	
chr19	64147	t	T	36	0	0	1	.	F	
chr19	64148	t	T	36	0	0	1	.	F	
chr19	64149	t	T	36	0	0	1	.	F	
chr19	64150	a	A	36	0	0	1	.	F	
chr19	64151	a	A	36	0	0	1	.	F	
chr19	64152	c	C	36	0	0	1	.	H	
chr19	64153	c	C	36	0	0	1	.	H	
chr19	64154	c	C	36	0	0	1	.	H	
chr19	64155	a	A	36	0	0	1	.	H	
chr19	64156	a	A	36	0	0	1	.	H	
chr19	64157	t	T	36	0	0	1	.	J	
chr19	64158	c	C	36	0	0	1	.	J	
chr19	64159	c	C	36	0	0	1	.	J	
chr19	64160	a	A	36	0	0	1	.	J	
chr19	64161	a	A	36	0	0	1	.	J	
chr19	64162	t	T	36	0	0	1	.	J	
chr19	64163	a	A	36	0	0	1	.	J	
chr19	64164	a	A	36	0	0	1	.	J	
chr19	64165	a	A	36	0	0	1	.	J	
chr19	64166	g	G	33	0	0	2	.^L	J;	
chr19	64167	a	A	33	0	0	2		J;	

Ilustración 21: Primeras filas del resultado del Pileup.

Tras filtrar estos resultados por coberturas mayores a 10 y con calidades mayores a 20, se puede observar una diferencia importante entre el archivo completo, teniendo 2.1 millones de posiciones del genoma cumpliendo estas características, y el archivo de solo las variantes, siendo estas de 7.121:

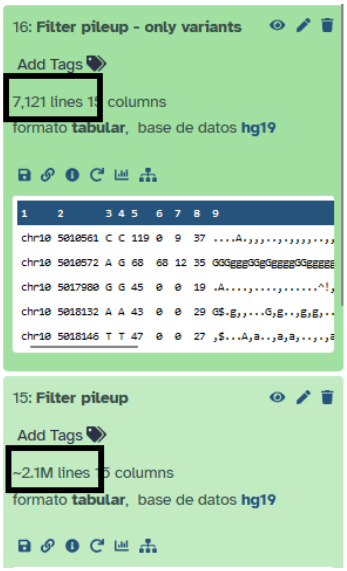

Ilustración 22: Resultados del filtrado de resultados.

En cuanto la detección de variantes genéticas (SNPs, Indels y variantes estructurales menores) sobre el archivo SortSam, obtenemos un total de 29.876 variantes genéticas al comparar las lecturas alineadas con el genoma de referencia (hg19). Este archivo nos indica:

- **Chrom:** Cromosoma donde se encuentra la variable.
- **Pos:** Posición en el genoma de la variante.
- **Ref y Alt:** Bases de referencia y las variantes.
- **Qual:** Calidad de la variante.
- **Info:** Datos adicionales.

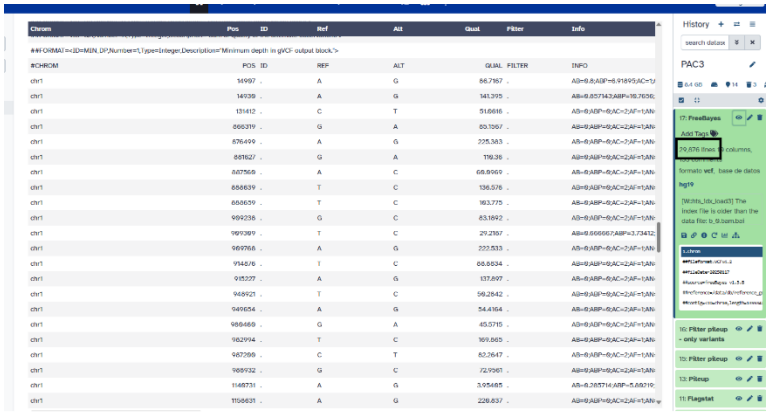

Ilustración 23: Primeras filas de la detección de variantes genéticas.

Estos resultados confirman la presencia de múltiples variantes (SNPs e indels).

4. Visualización de los resultados intermedios mediante un visor de genomas.

Al visualizar las alineaciones a través de IGB para todos los cromosomas, se observa una alta densidad de cobertura en las regiones, sugiriendo una buena profundidad de secuenciación:

Ilustración 24: Ejemplo de las alineaciones en el chr1.

Esto puede observarse del mismo modo con el navegador UCSC, por ejemplo, en el chr7:

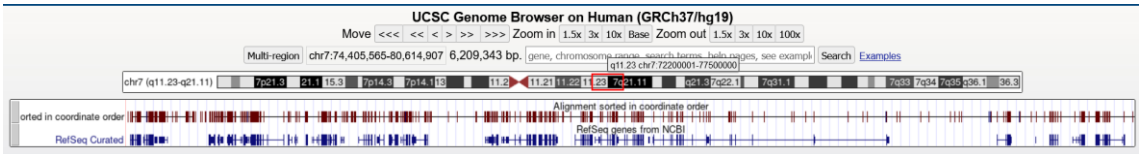

Ilustración 25: Ejemplo de las alineaciones en el chr7.

Finalmente, a través de bam.iobio podemos confirmar que hay regiones de cobertura en todos los cromosomas:

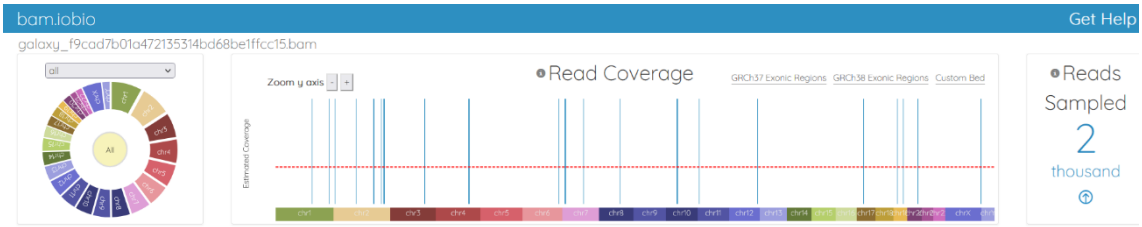

Ilustración 26: Cobertura en los cromosomas para un subconjunto de 2.000 lecturas.

Así mismo, se obtienen los siguientes datos:

- 99.8% de las lecturas están alineadas con el genoma de referencia.
- El 50% de las lecturas están alineadas con la hebra directa, lo que sugiere una distribución uniforme entre estas.
- El 99.3% de las lecturas pareadas están alineadas correctamente como pares, estando en las posiciones esperadas y en la orientación correcta en el genoma hg19.
- Solo hay 1 lecutra no pareada.
- El 99.9% de los pares tienen ambos fragmentos pareados.
- Hay un 0% de lecturas duplicadas.

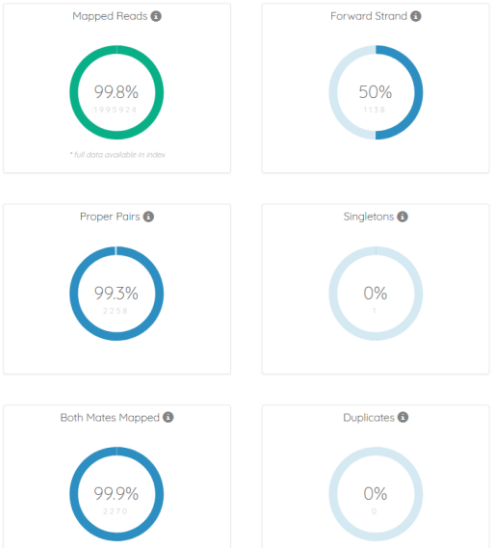

Ilustración 27: Métricas del alineamiento.

Por otro lado, en el gráfico de Read Coverage Distribution, se observan coberturas bajas, estando la mayor parte de esta en 0X.

Finalmente, si comparamos la misma región del cromosoma 7 entre el genoma de referencia, el alineamiento y las variantes a través del navegador genómico UCSC:

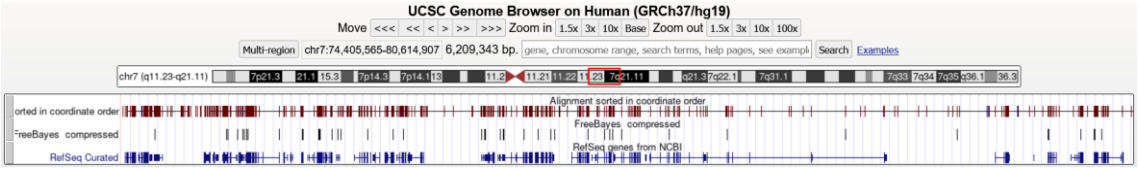

Ilustración 28: Visualización de la comparativa entre hg19, SortSam y FreeBayes.

5. Filtrado y anotación de variantes genéticas.

El número total de variantes detectadas tras el filtrado es de 29.943, siendo de los siguientes tipos:

Number variants by type

Type	Total
SNP	27,622
MNP	628
INS	660
DEL	941
MIXED	92
INV	0
DUP	0
BND	0
INTERVAL	0
Total	**29,943**

Ilustración 29: Número de variantes por tipo.

En cuanto a las anotaciones funcionales se obtiene:

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	13,632	50.805%
NONSENSE	340	1.267%
SILENT	12,860	47.928%

Ilustración 30: Número de variantes por efectos funcionales.

Por otro lado, en referencia a las regiones afectadas, las estadísticas son:

Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	2,641	2.933%
5_prime_UTR_premature_start_codon_gain_variant	161	0.179%
5_prime_UTR_variant	1,266	1.406%
conservative_inframe_deletion	48	0.053%
conservative_inframe_insertion	30	0.033%
disruptive_inframe_deletion	28	0.031%
disruptive_inframe_insertion	12	0.013%
frameshift_variant	497	0.552%
initiator_codon_variant	16	0.018%
intergenic_region	4,300	4.776%
intron_variant	46,460	51.602%
missense_variant	13,885	15.422%
non_coding_transcript_exon_variant	3,754	4.169%
splice_acceptor_variant	121	0.134%
splice_donor_variant	181	0.201%
splice_region_variant	3,315	3.682%
start_lost	18	0.02%
start_retained_variant	4	0.004%
stop_gained	342	0.38%
stop_lost	35	0.039%
stop_retained_variant	12	0.013%
synonymous_variant	12,909	14.338%

Region

Type (alphabetical order)	Count	Percent
EXON	31,158	36.062%
INTERGENIC	4,300	4.977%
INTRON	43,801	50.694%
SPLICE_SITE_ACCEPTOR	121	0.14%
SPLICE_SITE_DONOR	174	0.201%
SPLICE_SITE_REGION	2,780	3.218%
UTR_3_PRIME	2,641	3.057%
UTR_5_PRIME	1,427	1.652%

Ilustración 31: Número de variantes por región afectada.

Otros datos destacables son la frecuencia alélica, así como la previsibilidad del impacto de estas:

Allele frequency

Min	0
Max	100
Mean	81.873
Median	100
Standard deviation	24.314
Values	0,50,100
Count	80,10671,19125

Ilustración 32: Frecuencia alélica de las variantes.

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	1,182	1.368%
LOW	15,591	18.045%
MODERATE	13,963	16.161%
MODIFIER	55,666	64.427%

Ilustración 33: Impacto predecible de las variantes.

En cuanto los cambios entre bases, se dispone de la siguiente tabla:

Base changes (SNPs)

	A	C	G	T
A	0	836	4,699	878
C	1,106	0	1,205	4,654
G	4,834	1,197	0	1,489
T	1,005	4,841	878	0

Ilustración 34: Cambios específicos entre bases.

Finalmente, esto queda relacionado con la relación Transiciones/Transversiones:

Ts/Tv (transitions / transversions)

Transitions	31,257
Transversions	13,747
Ts/Tv ratio	2.2737

Ilustración 35: Razón Ts/Tv.

Discusión y limitaciones.

El presente estudio confirma la eficacia del pipeline utilizado para el análisis genómico, obteniendo resultados robustos y congruentes con la bibliografía. En primer lugar, el 99.8% de las lecturas se alinearon correctamente al genoma de referencia, con un 99.3% de lecturas pareadas alineadas correctamente como pares y un 99.9% de los pares teniendo ambos fragmentos pareados. Estas métricas indican una alta calidad del alineamiento y un procesamiento adecuado de los datos.

La distribución de cobertura de lectura (Read Coverage Distribution) muestra coberturas bajas, lo cual es consistente con la naturaleza de los datos del Proyecto de los 1000 Genomas. Este proyecto utiliza lecturas fragmentadas, lo que genera una distribución desigual a lo largo del genoma.

Referente a la la identificación de variantes, se observaron 27.622 SNPs y 1.601 Indels. La proporción predominante de SNPs es esperada, ya que constituyen aproximadamente el 90% de todas las variaciones humanas (Lavebratt & Sengul, 2006). Así mismo, en este estudio, la frecuencia fue de 1 variante cada 103.841 bases.

En el análisis funcional, las variantes Missense representan el 50.805% del total, seguidas de las Silent (47.928%). Las variantes Silent no alteran la secuencia de aminoácidos debido

a la redundancia del código genético, mientras que las variantes Missense generan una alteración en un aminoácido, pudiendo ser responsables de enfermedades como la anemia falciforme. Las variantes Nonsense, aunque poco frecuentes (1.267%), son significativas porque producen codones de parada prematuros, truncando las proteínas y generando potenciales efectos patogénicos (*Types of Mutations*, n.d.).

En cuanto a las regiones genómicas afectadas, la mayor proporción de variantes se localizó en regiones intergénicas (51.602%), las cuáles no codifican proteínas, seguidas de intrones (50.694%) y exones (36.062%), lo cual también es consistente ya que las variantes intrónicas son frecuentes en genes evolutivamente antiguos que están muy conservados a nivel de secuencia proteica. Este resultado contrasta con las pérdidas que solapan exones, que se observan con menos frecuencia de lo esperado por azar (Rigau et al., 2019). Por otro lado, las variantes Frameshift son especialmente relevantes pese a su baja frecuencia (0.552%) debido a su impacto funcional severo al provocar la creación de un codón STOP, generando una inserción o deleción de pares de bases no múltiples de tres, alterando la lectura de tripletes, quedando el producto proteico truncado (*Definition of Frameshift Variant - NCI Dictionary of Genetics Terms - NCI*, n.d.).

La distribución alélica promedio es del 81.87%, indicando que la mayoría de las variantes identificadas son frecuentes. Respecto al impacto funcional, la mayoría de las variantes tiene un impacto modificador (64.4%), seguido de impacto bajo (18.045%), moderado (16.161%) y alto (1.368%).

Así pues, es importante que, aunque con poca frecuencia, se realice un análisis más profundo de las variantes con impacto funcional alto, generalmente correspondientes a Nonsense y Frameshift (Rašić et al., 2014), las cuales suman un 1.819%, puesto que su relevancia biológica es importante.

En el análisis de cambios específicos de base, los más frecuentes fueron C -> T (4.841%) y A -> G (4.834%), seguidos de G -> A (4.699%) y T -> C (4.654%). Este patrón refleja una mayor proporción de transiciones respecto a transversiones, con una razón Ts/Tv de 2.2737, característica de datos genómicos de alta calidad (Wang et al., 2014).

Por otro lado, este estudio presenta una serie de limitaciones, destacando el uso del genoma hg19 como referencia, lo cual puede excluir variantes presentes en regiones recientemente anotadas en versiones más actualizadas del genoma humano.

Así mismo, sería de gran interés validación experimental para las variantes detectadas, para disminuir el riesgo de excluir variantes relevantes o de falsos positivos.

Finalmente, la baja cobertura en determinadas regiones del genoma puede haber limitado la detección de variantes, lo que podría solventarse con técnicas de secuenciación de mayor profundidad.

Conclusiones.

1. El 99.8% de las lecturas se alinearon correctamente al genoma de referencia, confirmando la robustez del pipeline empleado.
2. Se identificaron 27.662 SNPs y 1.601 Indels, siendo la proporción de SNPs de un 94.52%, en línea con otros estudios.
3. Las variantes Missense y Silent fueron las más frecuentes, suman
4. do 98.733%.
5. Aunque poco frecuentes, las variantes Frameshift y Nonsense tienen un impacto funcional significativo, por lo que es interesante seguir investigándolas.
6. La gran mayoría de variantes se localizan en regiones intergénicas e intrónicas, lo cual concuerda con patrones evolutivos conocidos.
7. Los cambios más comunes fueron las transiciones, con una relación Ts/Tv de 2.2737, consistente con datos de alta calidad.
8. Las limitaciones principales incluyen utilizar el genoma hg19 como referencia, la baja cobertura en algunas regiones y la ausencia de validación experimental, lo que puede solventarse con técnicas de secuenciación más avanzadas y actualización del genoma de referencia.

Bibliografía.

Average TsTv Ratio Calculation · Issue #1526 · samtools/bcftools. (n.d.). Retrieved January 19, 2025, from https://github.com/samtools/bcftools/issues/1526

Definition of frameshift variant - NCI Dictionary of Genetics Terms - NCI. (n.d.). Retrieved January 19, 2025, from https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/frameshift-variant

Lavebratt, C., & Sengul, S. (2006). Single nucleotide polymorphism (SNP) allele frequency estimation in DNA pools using Pyrosequencing. *Nature Protocols*, *1*(6), 2573–2582. https://doi.org/10.1038/NPROT.2006.442

Number of SNP effects by impact. SNP effects were categorized by impact... | Download Scientific Diagram. (n.d.). Retrieved January 19, 2025, from https://www.researchgate.net/figure/Number-of-SNP-effects-by-impact-SNP-effects-were-categorized-by-impact-as-high_fig3_261607416

Output summary files - SnpEff & SnpSift. (n.d.). Retrieved January 19, 2025, from https://pcingola.github.io/SnpEff/snpeff/outputsummary/

Rašić, G., Filipović, I., Weeks, A. R., & Hoffmann, A. A. (2014). Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, Aedes aegypti. *BMC Genomics*, *15*(1). https://doi.org/10.1186/1471-2164-15-275

Rigau, M., Juan, D., Valencia, A., & Rico, D. (2019). Intronic CNVs and gene expression variation in human populations. *PLoS Genetics*, *15*(1), e1007902. https://doi.org/10.1371/JOURNAL.PGEN.1007902

Types of Mutations. (n.d.). Retrieved January 19, 2025, from https://ib.bioninja.com.au/types-of-mutations/

Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2014). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*, *31*(3), 318. https://doi.org/10.1093/BIOINFORMATICS/BTU668