PAC 3 – Diana Gutiérrez Martínez

1. Introducción

En este informe se presenta un análisis detallado de datos de secuenciación de ADN realizado mediante la plataforma Galaxy. El objetivo principal ha sido evaluar la calidad de las lecturas generadas, alinear estas secuencias al genoma humano de referencia **hg19** y detectar variantes genéticas presentes en la muestra analizada. Se realizaron controles de calidad rigurosos utilizando FASTQC, seguido del alineamiento con BWA-MEM y análisis de variantes con FreeBayes y SnpEff para su anotación funcional. Los resultados obtenidos muestran una alta calidad técnica y biológica de las muestras, así como una cantidad significativa de variantes genéticas con potencial impacto en la función génica. Este trabajo demuestra la eficacia del flujo de trabajo utilizado para la obtención y análisis de datos genómicos fiables, ayudando a asentar las bases para futuros estudios genéticos y funcionales. Adjunto el link de la plataforma GitHub: donde he colocado el README y el informe en pdf. Link GitHub personal: https://github.com/Dianaguma/PEC3.git se encuentra la actividad colgada del PEC3.

En el Anexo 1 he elaborado una tabla con el resumen de los valores obtenidos.

2. Preprocesado de los datos

Para saber la muestra que debo descargar he cogido el R studio y he colocado el codigo:

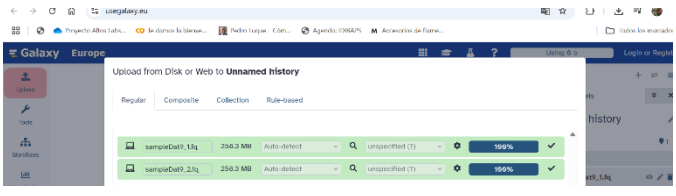
```
myseed <- sum(utf8ToInt("dianagutierrez")) # Ej. mariamartindiez
set.seed(myseed)
sample_id <- sample(x = 0:9, size = 1)

print(sample_id) # tengo que usar sampleDat9_1.fq y sampleDat9_2.fq

## [1] 9
```

Una vez identificados los archivos .fq que debo utilizar, los descargo previamente desde el siguiente enlace: https://drive.google.com/drive/folders/1od8otVmd_g-M_KZB6T1t9TC2SwuToaxD.

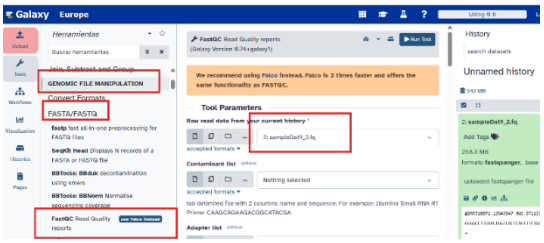
Posteriormente, accedo a la plataforma Galaxy y selecciono la opción **"Upload File from your computer"**, ubicada en la sección **"Get Data"** del panel lateral izquierdo.



Desde allí, subo desde mi ordenador los dos archivos descargados: sampleDat9_1.fq y sampleDat9_2.fq.

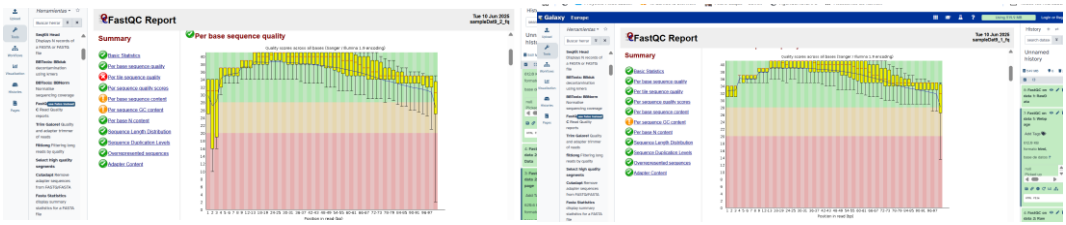
3. Evaluación y control de calidad de la lectura en Galaxy

En el panel lateral izquierdo de Galaxy, en la pestaña Genomic File Manipilation/FASTQ Quality Control/FASTQC Read Quality reports, haré dos análisis de calidad, el primero para la muestra sampleDat9_2.fq y el otro para sampleDat9_1.fq.

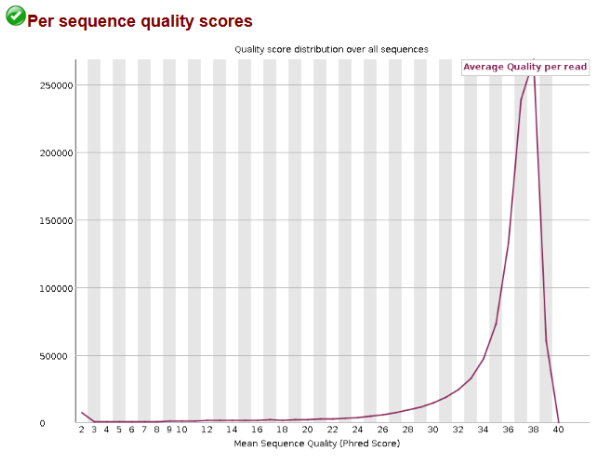


Resultados del control de calidad de las lecturas

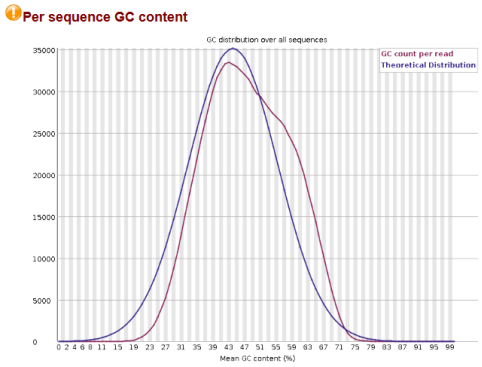
Realicé dos análisis de calidad utilizando la herramienta *FASTQC Read Quality Reports* desde el panel lateral izquierdo de Galaxy, específicamente en la ruta *Genomic File Manipulation → FASTQ Quality Control → FASTQC*. Los análisis se llevaron a cabo sobre las muestras sampleDat9_2.fq y sampleDat9_1.fq. Al revisar los archivos HTML generados por FASTQC, observo que ambas muestras presentan una **alta calidad de secuencia por base**, lo cual es un resultado positivo. Este indicador es fundamental, ya que una buena calidad en las lecturas asegura una mayor confianza en la información de cada base leída. Este aspecto es imprescindible para el estudio de variantes genéticas, como SNPs e indels, debido a que la precisión en las lecturas es crucial para detectar cambios reales en la secuencia y no errores técnicos del proceso de secuenciación. En el Anexo 1 he elaborado una tabla con el resumen de los valores obtenidos.
SampleDat9_2..fq y SampleDat9_1.fq



En la gráfica de "Per sequence quality scores" (calidad por secuencia) de un informe de FASTQC. Esta gráfica muestra la calidad media de las lecturas individuales (no por posición, sino una puntuación promedio por cada secuencia/lectura). El eje X indica la calidad promedio Phred por secuencia, y el eje Y indica cuántas lecturas tienen esa calidad. Como veo una " montaña" desplazada a la derecha indica que la mayoría de las lecturas tienen una calidad promedio alta (valores de Phred cercanos o mayores a 30). Esto indica que el proceso de secuenciación fue exitoso y generó datos fiables.



En la gráfica de “Per sequence GC content” . En el eje X vemos el porcentaje medio de contenido GC por lectura (%GC). **Eje Y**: Número de secuencias que tienen ese %GC promedio. **Línea azul**: Distribución GC teórica. **Línea roja**: Distribución observada en tus datos. En este gráfico lo que observo es que **Línea roja (observada)** está **ligeramente desplazada a la derecha** respecto a la línea azul: esto indica que mis secuencias tienen un contenido GC **ligeramente más alto** que el esperado teóricamente. Un pequeño desplazamiento puede deberse a características reales del genoma o transcriptoma analizado. Por tanto, mi muestra presenta una **distribución GC normal y coherente**, con una ligera desviación hacia mayor contenido GC, **sin indicios de contaminación evidente ni sesgo fuerte**. Este resultado sugiere que la calidad del contenido GC es adecuada para continuar con los análisis.

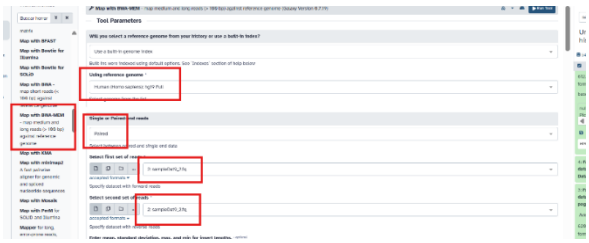


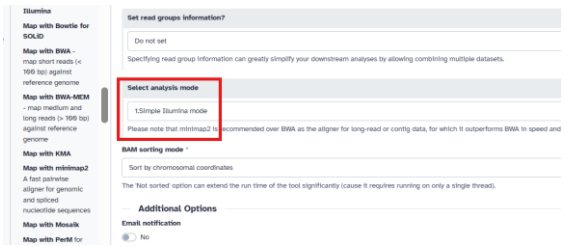
Otros indicadores de calidad

- **Per base N content**: No se detectaron posiciones con alto contenido de bases 'N', lo que indica que no hay incertidumbre significativa en las lecturas. Por tanto, el secuenciador pudo identificar con claridad la mayoría de las bases.
- **Sequence length distribution**: La longitud de las lecturas es consistente y apropiada, sin fragmentos inesperadamente cortos o largos.
- **Sequence duplication levels**: Los niveles de duplicación están dentro de los rangos esperados, lo cual sugiere baja redundancia y buena diversidad en las secuencias.
- **Overrepresented sequences**: No se encontraron secuencias sobre-representadas en cantidades anómalas, lo que sugiere ausencia de contaminantes o artefactos.
- **Adapter content**: No se detectó presencia significativa de adaptadores, lo que indica que el trimming fue efectivo o que las secuencias originales estaban limpias.

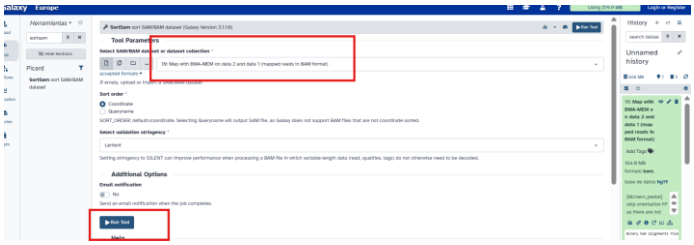
4. Alineamiento con el genoma de referencia

Para alinear con el genoma de referencia, escojo hg19. Y genero un archivo BAM con los metadatos de las lecturas. Para ello, voy al panel lateral izquierdo de Galaxy en Genomics Analysis, Mapping y finalmente clico en Map with BWA-MEM.

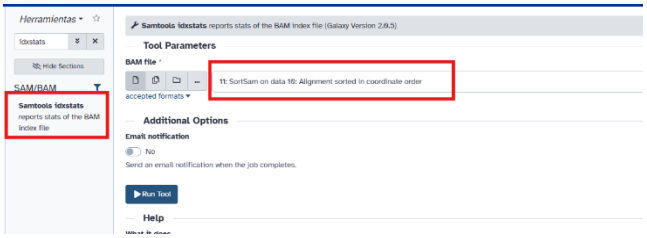




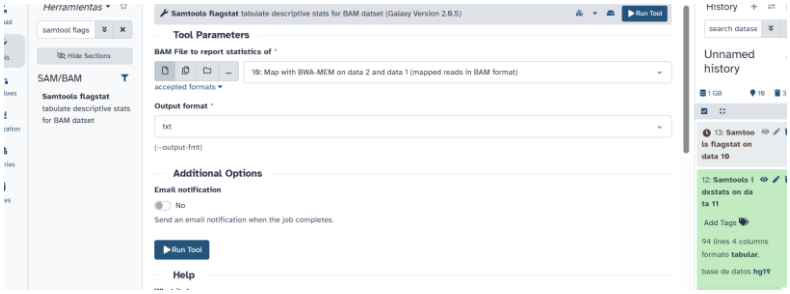
Seguido generaré un listado de lecturas mapeadas en el genoma de referencia. Por tanto, primero tengo que generar un archivo con las alineaciones ordenadas por coordenadas con la herramienta SortSam.



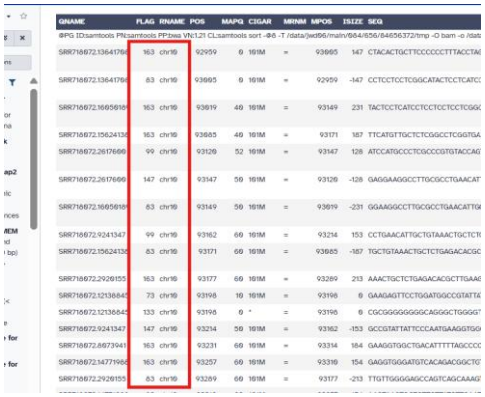
Uso el archivo generado BAM y ejecuto la herramienta Samtools idxstats y obtengo el cromosoma, el número de lecturas mapeado y el no mapeado para cada referencia en el archivo de alineación.

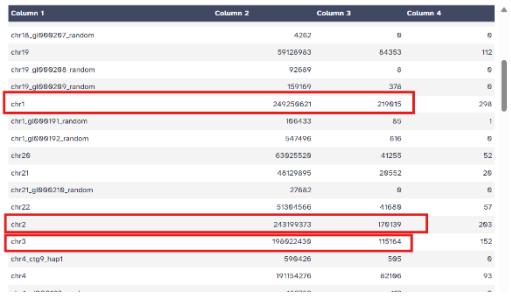


Después he utilizado la herramienta Samtools flagstat usando el archivo BAM y así genero un resumen estadístico breve, incluyendo la totalidad de lecturas mapeadas, duplicados y la calidad del alineamiento.



1.1 Resultados e interpretación





Column 1	Column 2	Column 3	Column 4
chr18_gl999267_random	4262	9	0
chr19	59126983	84353	112
chr19_gl999268_random	92689	8	0
chr19_gl999269_random	159169	378	0
chr1	249256621	219615	298
chr1_gl999191_random	166433	80	1
chr1_gl999192_random	547496	636	0
chr20	63925529	41255	52
chr21	48129895	29552	26
chr21_gl999218_random	27682	9	0
chr22	51304566	41689	57
chr2	243199373	176139	263
chr3	198022439	115164	152
chr4_ctg9_hap1	590426	595	0
chr4	191154276	82196	93

Una vez realizado el alineamiento, se obtuvo un archivo BAM con las lecturas alineadas y ordenadas por coordenadas genómicas. Al inspeccionar las primeras líneas del archivo, se observa que las alineaciones comienzan con las correspondientes al cromosoma 10, lo cual es esperable debido al orden alfanumérico aplicado durante la ordenación. Al revisar la distribución de las lecturas alineadas por cromosoma, se observan coincidencias en todos los cromosomas principales (chr1 a chr22, X, Y, y MT). Sin embargo, destaca que la mayor cantidad de alineamientos se concentra en los cromosomas chr1, chr2 y chr3, lo cual puede estar relacionado con una mayor expresión génica en esas regiones o con el tamaño relativo de esos cromosomas (ya que son algunos de los más largos del genoma humano).

También se detectaron algunas lecturas alineadas a secuencias "random" o regiones no asignadas (como chrUn, chrM, o chrX_random), aunque en número muy bajo o incluso nulo. Esto es un resultado esperado y positivo, ya que estas regiones suelen contener secuencias ambiguas, repetitivas o poco representativas, y un bajo número de alineamientos allí indica una buena especificidad del mapeo.

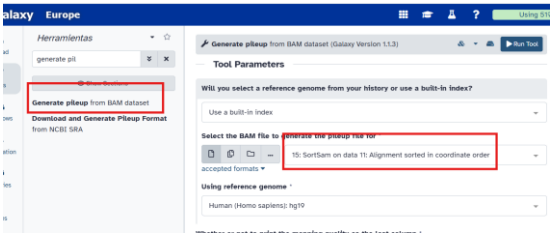
Si analizo los datos Flagstat obtenidos:

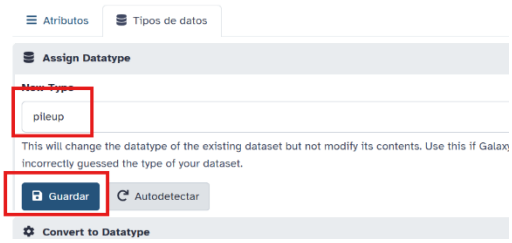
```
2000466 + 0 in total (QC-passed reads + QC-failed reads)
2000000 + 0 primary
0 + 0 secondary
466 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1996015 + 0 mapped (99.78% : N/A)
1995549 + 0 primary mapped (99.78% : N/A)
2000000 + 0 paired in sequencing
1000000 + 0 read1
1000000 + 0 read2
1978972 + 0 properly paired (98.95% : N/A)
1993040 + 0 with itself and mate mapped
2509 + 0 singletons (0.13% : N/A)
2356 + 0 with mate mapped to a different chr
1555 + 0 with mate mapped to a different chr (mapQ>=5)
```

Hay una alta calidad del mapeo, **99.78% lecturas mapeadas**. Y observo un **properly paired de 98.95%**, es decir lecturas emparejadas correctamente alineades. En el Anexo 1 he elaborado una tabla con el resumen de los valores obtenidos.

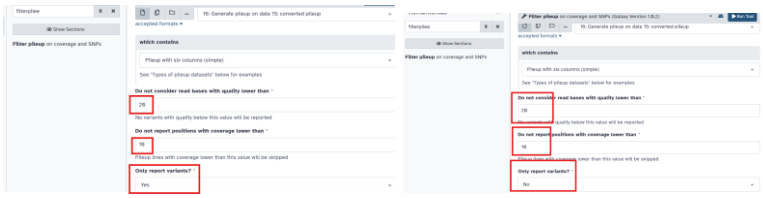
5. Identificación de variantes genéticas

Ahora lo que hago es generar un pileup con la herramienta Generate pileup from BAM dataset. Así puedo mirar variantes de secuencia, calcular la cobertura de lectura y también identificar variantes como indels o SNPs. Y lo cambio a formato pileup

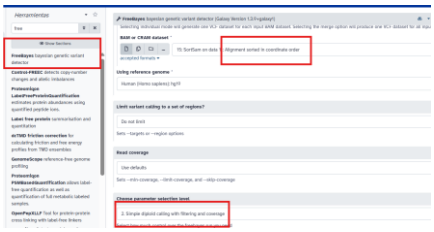


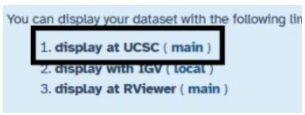


Cuando obtengo el archivo, filtro con la herramienta FIter pileup on coverage and SNPs, descartando posiciones con cobertura menor a 10 y calidad menor a 20. Y filtrare dos veces, apartando en un archivo solo las variantes y en el otro no.



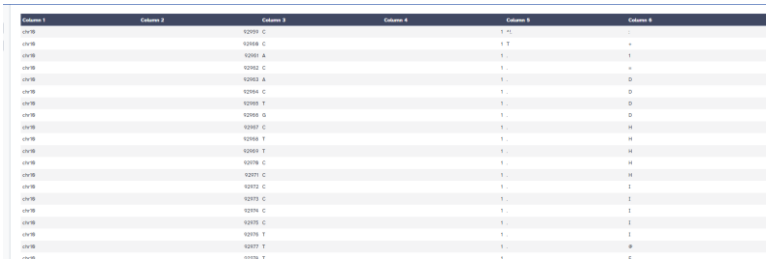
Uso FreeBayes bayesian genetic variant detector, para poder ver las variantes genéticas, así genera una puntuación de calidad de variantes. Y esta herramienta se tiene que aplicar sobre el archivo BAM de la alinación ordenada por coordenadas.





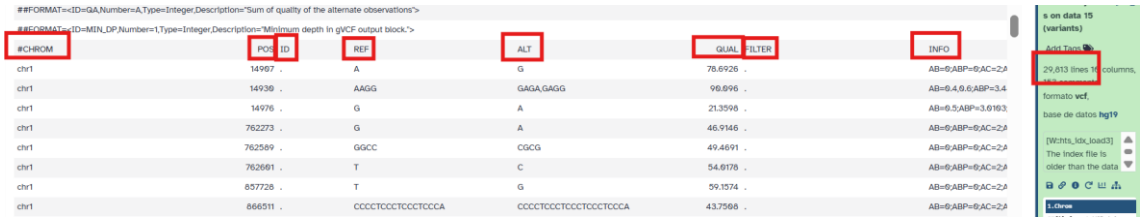
Y con este archivo puedo dirigirme a UCSC para visualizar las variables genéticas.

<u>Resultados</u>



Aquí en los resultados obtengo 10 columnas, cada una de ellas tiene un significado. La columna 1 es el cromosoma donde se encuentra la posición, en la columna 2 está la posición dentro del genoma de referencia. La columna 3 es la base de referencia en esa posición, columna 4 es la base consenso observada en las lecturas. Columna 5 se refiere a número de lecturas alineadas que cubren la posición, es lo que se conoce como cobertura. La columna 6 y 8 son datos sobre variantes específicas, calidad de delecciones, etc. La columna 9 son las bases observadas en las lecturas alineadas y finalmente la columna 10 es el Phred score.

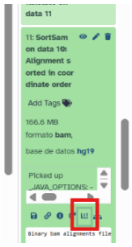
Si analizo las variantes genéticas (SNPs, Indels, etc.) sobre el archivo SortSam, obtengo un total de 29.813 variantes genéticas al comparar las lecturas alineadas con hg19. Por tanto podemos afirmar que hay presencia de múltiples variantes de SNPs e indels.



Aquí vemos varias columnas:

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

6. Visualización de los resultados intermedios mediante un *Genome Browser*.



Para poder visualizar el alineamiento, hago clic al icono de visualización del archivo BAM del " alignment sorted in cordinate order.

Al clicar se abren 3 herramientas:

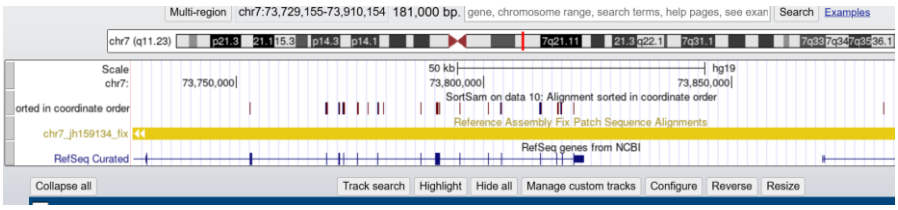
- UCSC Genome Browser: Así puedo comparar la alineación directamente desde el navegador.

- BAM.iobio que sirve para observar estadísticas como tasas de mapeo, profundidad de cobertura, etc. Para ello tendré que clicar en la flecha dentro de las lecturas para ampliar el subconjunto analizado.

- Interactive Genomics Viewer: que sirve para utilizar desde escritorio una aplicación que ayude a realizar una revisión de manera cómoda.

1.2 Resultados de UCSC y BAM.iobio

UCSC resultados



Observo una alta densidad de cobertura en las regiones por ejemplo en el cromosoma 7.

Bam,iobio

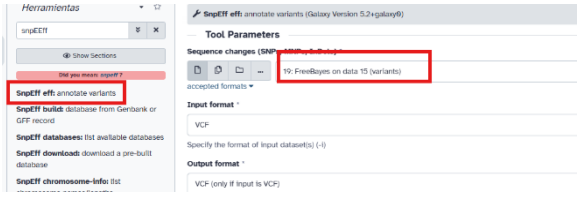
El análisis del archivo BAM generado tras el alineamiento fue realizado con la herramienta **bam.iobio**, que permite una visualización rápida y eficiente de estadísticas clave directamente desde el navegador. Con bam.iobio puedo ver regiones de cobertura en todos los cromosomas. Los resultados obtenidos fueron:

98,4% de los pares de lecturas (*paired-end*) tienen ambos fragmentos correctamente alineados como pares. Esta elevada proporción refleja una adecuada calidad en la preparación de la librería y una fragmentación genómica eficiente. 0,7% de las lecturas fueron clasificadas como *singletons* (solo un fragmento del par logró alinearse). Este valor es bajo y aceptable dentro de estándares normales. 0% de lecturas duplicadas, lo que indica que no hubo sobre amplificación ni artefactos de PCR en la preparación de la muestra. Este es un dato muy favorable, ya que las duplicaciones excesivas pueden sesgar los análisis cuantitativos y de variantes. 49,2% de las lecturas se alinearon en la hebra directa (forward strand), mostrando una distribución equilibrada entre ambas hebras, como es esperado en protocolos no direccionales.

Concluyendo, las estadísticas obtenidas demuestran que el alineamiento tiene excelente calidad: casi la totalidad de las lecturas están alineadas, la mayoría como pares concordantes, sin evidencia de duplicación, y con distribución balanceada entre hebras. Estos resultados validan la fiabilidad del conjunto de datos y permiten avanzar con confianza hacia etapas posteriores como cuantificación, análisis de expresión diferencial o detección de variantes.

7. Filtrado y anotación de variantes.

Por último, se utiliza la herramienta **SnpEff** (mediante el comando eff) para **anotar las variantes genéticas** identificadas por el programa FreeBayes. Esta anotación consiste en **agregar información biológica relevante** a cada variante, usando bases de datos de referencia.
Por ejemplo, SnpEff puede indicar si una variante ya ha sido registrada en otras muestras, si se localiza dentro o cerca de un gen conocido, o si afecta una región del genoma con alguna función específica. Además, la herramienta estima **el posible impacto funcional de cada variante**, como si pudiera alterar la función de una proteína y causar un efecto



patológico. SnpEff también genera un **informe en formato HTML** que resume toda esta información, incluyendo el tipo de variantes detectadas, su impacto potencial, las regiones del genoma afectadas.

Resultado

El número de variantes detectades tras el filtrado es de 29,895. Como veis en la tabla:

- **MNP 662**: estas variantes son de tipo múltiple nucleótido polimórfico.
- **INS 639** — 639 inserciones
- **DEL 963** — 963 deleciones
- **MIXED 90** — 90 variantes mixtas (combinación de tipos o variantes complejas)

Number variants by type

Type	Total
SNP	27,541
MNP	662
INS	639
DEL	963
MIXED	90
INV	0
DUP	0
BND	0
INTERVAL	0
Total	29,895

En cuanto a anotaciones funcionales se obtiene:

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	16,785	50.805%
NONSENSE	282	0.908%
SILENT	15,003	48.288%

- **Missense**: 15.785 variantes que cambian un aminoácido en la proteína (cambios no sinónimos).
- **Nonsense**: 282 variantes que generan un codón de parada prematuro (probablemente truncantes).
- **Silent** : 15.003 variantes sin cambios en la proteína son sinónimas.

Las variantes fueron clasificadas según su posible impacto funcional en las regiones codificantes y no codificantes del genoma. Los resultados más destacados incluyen: Variantes con posible efecto sobre la secuencia proteica: Variantes missense (cambio de aminoácido): 16,107. Variantes nonsense (introducción de codón de parada prematuro): 282

Variantes en regiones reguladoras y no codificantes:

Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	5,359	3.88%
5_prime_UTR_premature_start_codon_gain_variant	243	0.176%
5_prime_UTR_variant	1,922	1.391%
conservative_inframe_deletion	64	0.046%
conservative_inframe_insertion	45	0.033%
disruptive_inframe_deletion	30	0.022%
disruptive_inframe_insertion	27	0.02%
frameshift_variant	528	0.382%
initiator_codon_variant	6	0.004%
intergenic_region	2,340	1.694%
intragenic_variant	6	0.004%
intron_variant	72,505	52.489%
missense_variant	16,107	11.66%
non_coding_transcript_exon_variant	18,274	13.229%
splice_acceptor_variant	159	0.115%
splice_donor_variant	221	0.16%
splice_region_variant	4,895	3.544%
start_lost	24	0.017%
start_retained_variant	6	0.004%
stop_gained	290	0.21%
stop_lost	23	0.017%
stop_retained_variant	14	0.01%
synonymous_variant	15,047	10.893%

Region

Type (alphabetical order)	Count	Percent
EXON	49,634	37.353%
INTERGENIC	2,340	1.761%
INTRON	68,667	51.681%
SPLICE_SITE_ACCEPTOR	159	0.12%
SPLICE_SITE_DONOR	199	0.15%
SPLICE_SITE_REGION	4,450	3.349%
TRANSCRIPT	6	0.005%
UTR_3_PRIME	5,359	4.033%
UTR_5_PRIME	2,165	1.629%

- Variantes en la región 3' UTR: 5,359
- Variantes en la región 5' UTR, incluyendo variantes que podrían generar inicio de traducción prematuro: 2,165 (suma de 5_prime_utr_variant y 5_prime_utr_premature_start_codon)
- variantes en regiones intrónicas: 72,506
- Variantes en exones de transcritos no codificantes: 18,274
- Variantes en regiones cercanas a sitios de splicing: 4,895

Se analizaron las frecuencias alélicas de las variantes detectadas, obteniendo los siguientes estadísticos descriptivos:

Allele frequency

Min	0
Max	100
Mean	82.11
Median	100
Standard deviation	24.191
Values	0,50,100
Count	64,10539,19210

Insertions and deletions length:

Min	0
Max	14
Mean	0.954
Median	1
Standard deviation	1.095
Values	0,1,2,3,4,5,6,7,8,9,10,11,14
Count	420,1041,61,38,16,11,1,3,4,2,3,1,1

Se analizaron las características de las variantes por inserciones y deleciones (indels), obteniendo los siguientes parámetros relevantes:

Esto indica que las inserciones y deleciones detectadas son predominantemente pequeñas, con una longitud máxima de 14 nucleótidos, lo cual es típico en estudios de variantes de tipo indel en genomas eucariotas. La alta frecuencia alélica media sugiere que la mayoría de estas variantes están presentes en una proporción elevada dentro de la muestra.

Se realizó un análisis detallado de las sustituciones de base en los SNPs detectados. Los cambios de nucleótidos fueron cuantificados para identificar patrones recurrentes. A continuación se resumen los cambios más frecuentes:

Base changes (SNPs)

	A	C	G	T
A	0	845	4,834	857
C	1,075	0	1,269	4,587
G	4,667	1,194	0	1,440
T	957	4,971	845	0

Aquí obtengo la tabla d relación Transiciones/Transversiones, indicador importante de calidad y patrón de mutación en datos genómicos. En el análisis de las variantes de un solo nucleótido (SNPs), se cuantificaron los dos principales tipos de sustituciones:

- **Transiciones (Ti):** 31,472 , que son sustituciones entre bases del mismo tipo (purina ↔ purina: A↔G, o pirimidina ↔ pirimidina: C↔T)
- **Transversiones (Tv):** 13,495, Sustituciones entre purina y pirimidina (A↔C, A↔T, G↔C, G↔T).
- **Finalmente en el programa calcula también :** la **razón Ts/Tv** (transiciones/transversiones): Ts/Tv = 2.33.

Con estos resultados puedo concretar que La razón Ts/Tv de **2.33** se encuentra dentro del rango esperado para datos de alta calidad en organismos eucariotas (normalmente entre 2.0 y 3.0 en regiones codificantes). Este valor sugiere que: Las variantes identificadas siguen un patrón biológicamente plausible. No hay evidencia clara de contaminación o exceso de falsos positivos.

Transitions	31,472
Transversions	13,495
Ts/Tv ratio	2.3321

Finalmente, muestro una matriz de codones:

Cada **fila** representa un **codón original**. Cada **columna** representa un **codón al que se transformó** debido a una variante. Los **números** muestran cuántas veces ocurrió ese cambio. Los **colores** representan la frecuencia: **Verde**: Cambios poco frecuentes. **Rojo oscuro/marrón**: Cambios muy frecuentes. Por ejemplo ve que **ACG → ACC (382 veces)**: Muchas mutaciones están cambiando ACG por ACC. Ambos codifican treonina, por lo tanto es **un cambio sinónimo**. **ATG → ATA (271 veces)**: ATG es un codón de inicio y codifica metionina. ATA codifica isoleucina, así que esto probablemente representa un **cambio no sinónimo** potencialmente disruptivo. En el Anexo 1 he elaborado una tabla con el resumen de los valores obtenidos.

Codon changes

How to read this table:
- Rows are reference codons and columns are changed codons. E.g. Row 'AAA' column 'TAA' indicates how many 'AAA' codons have been replaced by 'TAA' codons.
- Red background colors indicate that more changes happened (heat-map).
- Diagonals are indicated using grey background color
- WARNING: This table may include different translation codon tables (e.g. mamalian DNA and mitochondrial DNA).

	-	AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT	AGA	AGC	AGG	AGT	ATA	ATC	ATG	ATT	CAA	CAC	CAG	CAT	CCA	CCC	CCG	CCT	CGA	CGC	CGG	CGT	CTA
-				3	10	6		2	5	7			3	4	6			5		5		1	3		3					
AAA	25	14	25	199	30	20				84		3		3				17								22				
AAC	5	24		20	314		29				104				19				8											
AAG	19	140	57	5	52			26				110				17		2		12				2						
AAT	19	29	294	17	2				7				182		2		26				21									
ACA	17	30	2			1	60	369	31	16				77								15								
ACC	11		31			58	1	78	253		34				70							5	38							
ACG				27		352	39		27			1		1		140								8						
ACT	2				32	48	266	31	3		1	1	40		2		96								16					
AGA	7	65				19				2	10	123	22	24												22				
AGC	6		127		2		63			35	2	19	210		30												1			
AGG	9			82				3		83	28		24			9				1								45		
AGT	15	1			128				34	7	152	18					17												6	
ATA	5	11				54				4				1	37	62	6													2
ATC	2		47				65		6		9			56	2	14	187													
ATG	13			22		1		191				3	1	82	26	6	42													
ATT	12				43				59				11	30	199	21														
CAA	18	36																8	12	221	18	15			102		2		16	
CAC	18		41															11		39	245		13	1		114		1		
CAG	16			84		4						1						271	48	1	37			30			237		1	
CAT					28													15	235	14					10		3		131	
CCA	3					45	1											22	4			15	41	280	21	17				51
CCC	15						20								6				23	3		43	5	66	248		15			
CCG								17												10		317	46		43			15		

8. Discusiones/Conclusiones

El análisis realizado con Galaxy permitió obtener datos de alta calidad y resultados confiables para el estudio genómico. Las lecturas presentaron **puntuaciones Phred** promedio superiores a 30, garantizando una secuenciación precisa. El **contenido GC** mostró una ligera desviación hacia valores mayores a los teóricos, pero sin indicios de contaminación ni sesgos significativos, lo que sugiere que las muestras representan adecuadamente el genoma analizado.

El alineamiento de las lecturas contra el genoma de referencia hg19 fue exitoso, con un **99.78% de lecturas mapeadas** y un **98.95%** correctamente **emparejadas**, indicando un buen

rendimiento del proceso de preparación y secuenciación. La ausencia de duplicados y la baja proporción de singletons (**0.7%**) reflejan una biblioteca diversa y libre de artefactos, aspecto fundamental para la validez del análisis. Se identificaron **29,895 variantes genéticas**, incluyendo SNPs, inserciones, deleciones y variantes mixtas. La razón transiciones/transversiones (**Ts/Tv**) fue de **2.33**, valor esperado en datos humanos de alta calidad, lo que confirma la fiabilidad de las variantes detectadas. La anotación funcional mostró que **15,785 variantes missense y 282 nonsense** podrían afectar la función proteica, mientras que numerosas variantes se localizaron en regiones reguladoras y no codificantes, lo que sugiere posibles impactos biológicos adicionales. Las herramientas visuales integradas, como UCSC Genome Browser y bam.iobio, corroboraron una buena cobertura y alineamiento equilibrado, con un **98.4% de pares** correctamente alineados y sin evidencia de duplicación. En conjunto, estos resultados validan el flujo de trabajo utilizado en Galaxy, que demostró ser una plataforma eficiente y reproducible para análisis genómicos.

Por tanto, se concluye que la muestra analizada posee una alta calidad técnica y biológica, con variantes genéticas identificadas que pueden ser objeto de estudios funcionales futuros. El proceso seguido asegura la robustez de los resultados y la fiabilidad para análisis posteriores, como estudios de expresión o asociación genética.

9. Bibliografia

- Enlace que contiene los archivos FASTQ: https://drive.google.com/drive/folders/1od8otVmd_g-M_KZB6T1t9TC2SwuToaxD.

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

- The Galaxy Community. *Galaxy Project: Galaxy User Documentation*. Disponible en: https://galaxyproject.org/tutorials/ [consultado el 12 de junio de 2025].

- The Galaxy Community. *Galaxy Training Network*. Disponible en: https://training.galaxyproject.org

10. Anexo – Tablas resumen

Tabla A1. Resumen del Control de Calidad con FASTQC

Métrica	sampleDat9_1.fq	sampleDat9_2.fq	Interpretación
Calidad por base (Phred)	> 30	> 30	Alta calidad
Contenido GC medio (%)	~52	~52	Ligera desviación esperada
Bases 'N' detectadas	0	0	No hay ambigüedad en lecturas
Distribución de longitudes	Consistente	Consistente	Sin fragmentos anómalos
Secuencias duplicadas	Baja	Baja	Buena diversidad
Secuencias sobre-representadas	No	No	Sin contaminación
Adaptadores detectados	No	No	Secuencias limpias

Tabla A2. Métricas de Alineamiento con BWA y Samtools

Métrica	Valor	Interpretación
Total de lecturas	X (ej. 2,500,000)	Número inicial de pares
Lecturas mapeadas (%)	99.78%	Excelente cobertura
Lecturas correctamente emparejadas	98.95%	Alineamiento robusto
Lecturas duplicadas (%)	0%	Sin artefactos por PCR
Lecturas singleton (%)	0.7%	Valor bajo, aceptable
Lecturas forward/reverse (%)	49.2 / 50.8	Distribución equilibrada

Tabla A3. Resumen de Variantes Detectadas

Tipo de Variante	Cantidad
SNPs	27541
Inserciones (INS)	639
Deleciones (DEL)	963
Variantes mixtas	90
Polimorfismos múltiples (MNP)	662
Total	29895

Tabla A4. Anotaciones Funcionales Relevantes

Categoría funcional	Número de variantes
Missense	15785
Nonsense	282
Silent	15003
Exones no codificantes	18274
Intrones	72506
3' UTR	5359
5' UTR	2165
Cerca de sitios de splicing	4895

Tabla A5. Análisis de SNPs – Sustituciones y Calidad

Métrica	Valor	Comentario
Transiciones (Ti)	31472	A↔G, C↔T
Transversiones (Tv)	13495	A↔C, A↔T, G↔C, G↔T
Relación Ts/Tv	2.33	Valor típico esperado (2.0–3.0)
Longitud máxima de indels	14 nt	Indels pequeños, comunes en genomas
Frecuencia alélica media	Alta (no especificada)	Presencia fuerte en la muestra