PEC3. Análisis de datos Ómicos.

Mª de la Vega Rodrigálvarez Chamarro

19 de enero de 2025

Tabla de contenido

1. Resumen ejecutivo

El siguiente estudio ha tenido como objetivo la identificación de variantes minoritarias tales como *SNV (Single Nucleotide Variants)* e *indels* utilizando herramientas bioinformáticas y flujos de trabajo automatizados dentro de la plataforma *Galaxy.* El conjunto de datos analizados es un subconjunto de datos de la muestra HG00128 que se encuentra dentro del proyecto 1000 genomas.

El flujo que se ha seguido para realizar la identificación y anotación de las variables ha sido: Carga y preprocesamiento de los datos donde se ha validado la calidad de los datos y su idoneidad para realizar el análisis; Alineación con el genoma de referencia GRCh38/hg38 donde se han obtenido un alto porcentaje de lecturas correctamente emparejadas y mapeadas; Identificación de variantes genéticas usando *FreeBayes* y *LoFreq* y donde 15 variantes coincidieron en ambas herramientas; Anotación de variables con *SnpEff* dando como resultado la mayoría de variantes de polimorfismo único.

Finalmente, los polimorfismos identificados tenían una frecuencia poblacional significativa por lo que será necesario revisar los criterios de filtrado previos a la identificación de secuencias con el objeto de identificar variables minoritarias. Se identificaron áreas para mejorar, como la integración de más bases de datos y herramientas para ampliar la validación y caracterización de variantes.

2. Objetivos del estudio

El objetivo del presente estudio es la identificación de variantes minoritarias pequeñas, tales como SNVs (*Single Nucleotide Variants)* e *indels (*Inserciones y Deleciones) a través del uso de herramientas bioinformáticas y flujos de trabajo automatizados.

Las variantes minoritarias se consideran aquellas que son detectadas con una baja frecuencia en la población y su frecuencia dentro de las lecturas de secuenciación se puede encontrar entre el 1% y el 5% (1% < MAF < 5%) [1]. Las variantes minoritarias pueden conllevar una resistencia a antivirales o antibióticos o, por ejemplo, en el cáncer pueden representar la existencia a una resistencia a la quimioterapia o lo inmunoterapia, es por ello, que mucho de los estudios referidos a este tipo de variantes están relacionados con la resistencia a diversos tipos de fármacos o el comportamiento de diferentes tratamientos [2] [3].

3. Materiales y métodos

Los datos del presente estudio se basan en una selección aleatoria de fragmentos cuya fuente original es la muestra HG00128[1] dentro del proyecto 1000 genomas. La muestra fue recogida entre una población femenina con origen británico y escocés y cuyos ancestros habían nacido todos en el Reino Unido. La muestra extraída es de ADN extraída de líneas celulares de linfoblastos[2] a partir de muestras de sangre extraída de los repositorios de *Coriell*[3]. Los ficheros utilizados para el presente estudio son los denominados *sampleDat6_1.fq* y *sampelDat6_2.fq*, descargados desde el repositorio[4] proporcionado para la práctica y las cuales se encuentran en formato *FASTQ* y consisten en lecturas apareadas ("*Paired end reads*").

Para el análisis de las muestras, se ha utilizado *Galaxy*[5]que es una plataforma web de código abierto que permite la investigación y el análisis de datos biomédicos. Esta plataforma permite la carga de ficheros bioinformáticos y ofrece una gran cantidad de herramientas para su procesamiento desde que la muestra es secuenciada hasta que la anotación de variantes es realizada pasando por diferentes etapas tales como el alineamiento con el genoma de referencia, el análisis de calidad de las diferentes fases o la detección e identificación de variantes.

Para la ejecución del presente análisis, se ha definido un flujo de trabajo con el objeto de estandarizar los pasos a seguir y que se sigan independientemente del conjunto de datos proporcionados. El flujo de trabajo (Figura 1) está disponible en la siguiente URL: https://usegalaxy.org.au/u/eowin/w/pec3-ado-wf-rodrigalvarezchamarro-vega-au

[1] Proyecto 1000 Genomas. Muestra HG00128
[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM649540
[3] https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=HG00128
[4] Repositorio UOC muestra aleatoria HG00128
[5] Galaxy

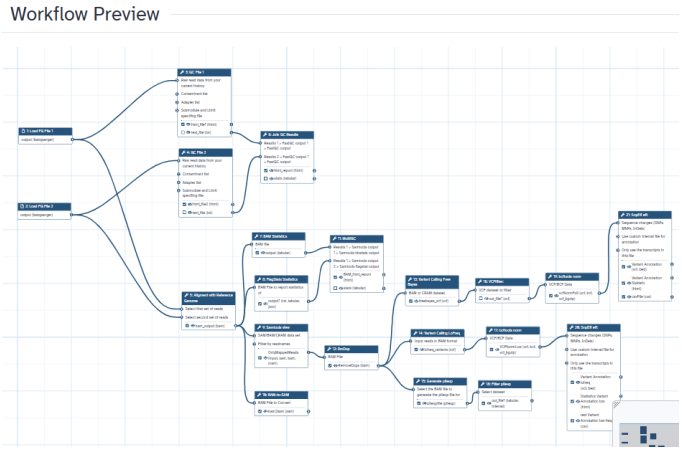

Figura 1. Flujo de trabajo PEC3.

Para comenzar con el análisis de datos se han creado dos puntos de entrada para poder **cargar los datos** en el momento en el que se inicia la ejecución del flujo de trabajo. Se define como formato de salida *fastqsanger* ya que los ficheros de entrada han sido secuenciados con *Ilumina* y se toma como base de datos de referencia *Human Dec. 2013 (GRCh38/hg38) (hg38)* tal y como se ha observado en la ficha de la muestra en el proyecto 1000 genomas.

Una vez que los ficheros han sido cargados, se ha **evaluado la calidad de las muestras**, de forma individual, usando la herramienta *FastQC* y se ha realizado una comparativa de las mismas a través de la herramienta *MultiQC.* El resultado de estas ejecuciones son varios informes en formato web donde se ve si las muestras son lo suficiente buenas para continuar con el análisis.

Después de que se ha efectuado la verificación de las muestras, se procede a la realizar el **alineamiento de las presentes muestras** con el genoma de referencia. Como genoma de referencia se ha utilizado *Human (Homo sapiens)(b38):hg38,* y como se indicaba en el enunciado se marca que las muestras a analizar están apareadas (*paired)*. Cómo en el análisis de calidad se ha visto que la codificación ha sido realizada con *Illumina* se selecciona como modo de análisis *Simple Illumina Mode* y, el resultado del alineamiento será un fichero en formato BAM ordenado según las coordenadas de los cromosomas.

Para **verificar la calidad del alineamiento**, se han utilizado varias herramientas que permiten visualizar estadísticas tales como *Samtools idxstats,* donde se indican el número de lecturas mapeadas, la longitud de las lecturas y su identificador así como las lecturas que no han sido mapeadas, *Samtools flagstats* para obtener información más descriptiva del fichero de alineamiento y *BAM-to-SAM* donde se muestra los nombre de los cromosomas y su identificación y otra información sobre la calidad de las diferentes lecturas. Finalmente se utiliza *Samtools View* con el objeto de filtrar las lecturas existentes en el BAM ya sea por región, calidad de la muestra o simplemente quedarse únicamente con aquellas lecturas que han sido mapeadas. Para eliminar todas aquellas lecturas duplicadas y reducir el tamaño de las lecturas se ha utilizado *RmpDup.*

Para obtener un conjunto de datos apilados a partir de los ficheros *BAM* proporcionados y así poder inspeccionar los datos de alineamiento de manera resumida, se utilizan las herramientas *Generate Pileup* y *Filter pileup.*

Una vez analizados los resultados generados por el alineamiento y revisada su calidad, se procede con la **identificación y detección de variantes**. En este caso se han analizado dos caminos:

- *FreeBayes* que es un detector de variantes genéticas bayesiano diseñado para encontrar polimorfismos pequeños, concretamente SNP (polimorfismos de un

solo nucleótido), *indels* (inserciones y deleciones), MNPs (polimorfismos de múltiples nucleótidos) y eventos complejos más pequeños que la longitud de una alineación de secuenciación de la lectura corta.

- *LoFreq* con el objeto de inferir SNVs e *indels* de baja frecuencia y acotar más los resultados a obtener a partir del conjunto inicial de datos.

Antes de realizar la anotación de las variantes, se procede a **normalizar** los resultados de las herramientas de identificación de variantes mencionadas con anterioridad, usando *bcf tools norm* y finalmente, se procede a realizar la **anotación de las variantes** usando *SnpEff eff* y tomando como base genómica de anotaciones el genoma *GRCh38.86*. Se obtiene un resultado para cada una de las ramas seguidas en la identificación de variantes.

4. Resultados

A continuación, se procede a mostrar los resultados obtenidos en cada uno de los pasos seguidos para obtener un listado de variantes minoritarias anotadas a partir de los ficheros proporcionados para el desarrollo del presente estudio.

4.1. Carga y preprocesamiento de los datos

Inicialmente, se procede a cargar los datos en el sistema. El resultado son dos archivos en formato *fastqsanger* y cuya base de datos de referencia es *hg38*. Ambos archivos están compuestos de 1.000.000 de secuencias con una longitud de 101 nucleótidos cada secuencia.

4.2. Evaluación y control de calidad

Una vez analizados los datos con la herramienta *fastq* y *multiqc*, se puede afirmar que la secuenciación proporcionada es de buena calidad.

Como se puede observar en la Tabla 1, el porcentaje de lecturas repetidas es muy baja y el porcentaje de nucleótidos GC está alrededor del 47%, aunque tal y como se puede observar en la Figura 2, no ajusta en su totalidad a curva de Gauss teórica, aunque resulta bastante similar.

Tabla 1. Calidad de las muestras secuenciadas.

Sample Name	% Dups	% GC	M Seqs
sampleDat6_1_fq	1.2%	47%	1.0M
sampleDat6_2_fq	0.8%	47%	1.0M

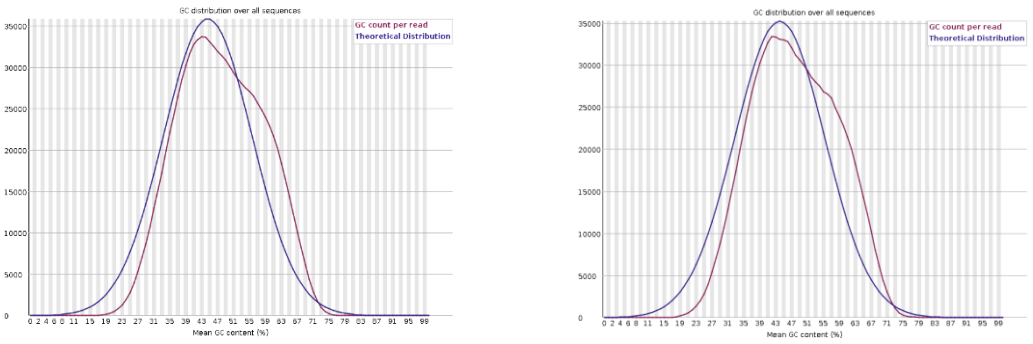

Figura 2. Contenido GC por secuencia.

Se observa que la calidad de la secuencia es muy buena, teniendo en todas las posiciones un valor mayor de 28 y situándose la mayoría de las lecturas en la zona verdes, excepto aquellas que se encuentran en los extremos donde la calidad empeora, aunque esto suele ser algo normal (Figura 3).

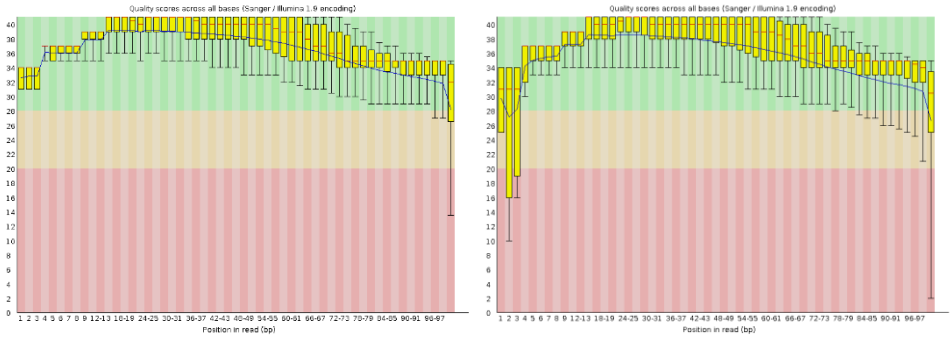

Figura 3. Calidad de la secuencia por base.

Como se puede observar, la calidad en las lecturas en la segunda muestra desciende al principio de la secuencia (Figura 4), lo que habría que tener en cuenta en los siguientes pasos. Para ver más detalle consultar los Anexos relativos a los informes de calidad (Anexo I, Anexo II, Anexo III). También se pueden ver los gráficos de calidad con más detalle en 9.1 Informe calidad de la primera secuencia, 0

Informe calidad de la segunda secuencia y 9.3 Informe calidad de ambas secuencias.

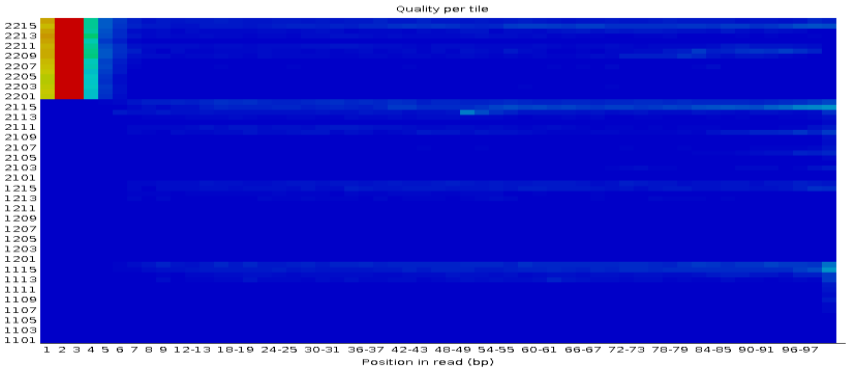

Figura 4. Calidad de la lectura por nucleótido.

4.3. Alineación secuencial al genoma de referencia

El siguiente paso consiste en realizar la alineación con el genoma de referencia *hg38*. A continuación, se muestran una serie de estadísticas una vez realizado el alineamiento.

En la Tabla 2 se puede observar que hay un total de 1764 lecturas que no han sido mapeadas. Según esta tabla se puede observar que aquellos cromosomas que más lecturas mapeadas tienen son los cromosomas *chr1, chr2, chrX, chr3* y *chr7*, pero si se compara el número de lecturas mapeadas frente a la longitud del cromosoma son los cromosomas *19, 17, 1, X* y *16* los que tienen mayores mapeos (se puede ver una tabla más completa en el apéndice 9.4 Calidad de los alineamientos o en el Anexo IV.

Tabla 2. Lecturas mapeadas por cromosoma.

Id. Cromosoma	Longitud	Nº mapeos	Nº lecturas no mapeadas	% chr mapeado
chr1	**248956422**	**215983**	294	0,0868
chr2	242193529	169850	204	0,0701
chr3	198295559	114137	135	0,0576
chr7	159345973	97581	124	0,0612
chr16	90338345	72839	93	0,0806
chr17	83257441	99533	126	0,1195
chr19	58617616	81883	120	**0,1397**
chrX	156040895	129339	146	0,0829
*	0	0	**1764**	

En la Figura 5 se puede observar el número de lecturas mapeadas por cada cromosoma donde se puede observar que los cromosomas *chr19* y *chr17* son los que tienen más lecturas mapeadas.

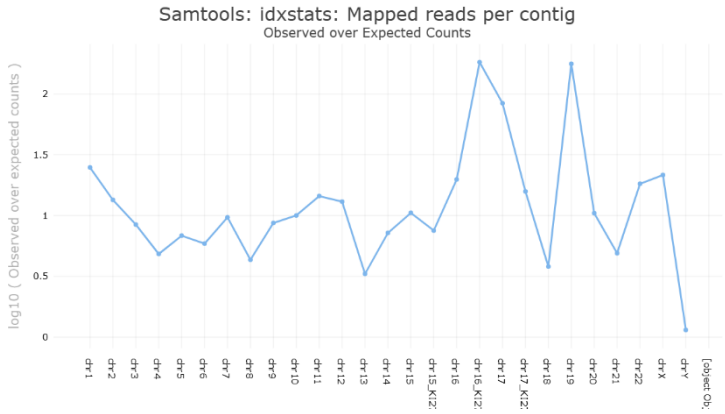

Figura 5. Lecturas mapeadas por cromosoma.

Al revisar las estadísticas relativas a los *flags,* se muestra que un 98,96% (1.979.252) de las lecturas han sido emparejadas correctamente y hay 2498 lecturas no emparejadas y 2530 que se han mapeado con un cromosoma diferente tal y como se muestra en la Tabla 3. A la vista de estos resultados, se llega a la conclusión de que el alineamiento con respecto al genoma de referencia es muy bueno.

Tabla 3. Calidad alineamiento BAM.

Sample	**Total Reads**	**Total Passed QC**	**Mapped**	**Supplementary Alignments**	**Duplicates**	**Paired in Sequencing**
flagstatistics	2000438	2000438	1996176	438	0	2000000
Sample	**Properly Paired**	**Self and mate mapped**	**Singletons**	**Mate mapped to diff chr**	**Diff chr (mapQ >= 5)**	
flagstatistics	1979252	1993240	2498	2530	1506	

En la Figura 30 se puede observar de forma gráfica y en porcentaje estas estadísticas, pero para ver el informe completo, consultar el Anexo IV.

Los datos resultantes del alineamiento son mostrados en un fichero con formato BAM el cual cuenta con la siguiente información [4] [5] [6] [7]:

- **QNAME**: Nombre de la lectura proveniente del secuenciador.
- **FLAG**: Información sobre la alineación. Es un valor numérico que es necesario pasarlo a binario para identificar el significado de cada flag.
- **RNAME**: Nombre del cromosoma sobre el que se realiza la alineación.
- **POS**: Posición del inicio de la alineación sobre la secuencia de referencia.
- **MAPQ**: Puntaje de calidad de la alineación. Se utiliza para entender lo único que es el alineamiento en el genoma (valores mayores que 10 indican que el alineamiento tiene una probabilidad muy alta de ser único).
- **CIGAR**: El número de bases coincidentes con la referencia.
- **MRNM**: Nombre de la referencia donde se alinea el compañero. Si está en el mismo cromosoma (=), si no se encuentra (*).
- **MPOS**: Posición en la que se encuentra el compañero.
- **ISIZE**: Longitud del fragmento entre las dos lecturas de un par (positivo o negativo en función si está antes o después). 0 si no está alineado.
- **SEQ**: Secuencia de la lectura
- **QUAL**: Puntuación de la calidad.
- **OPT**: Campos opcionales. NM: Distancia a la referencia, MD: Cadena para posiciones no coincidentes, MC: Cadena CIGAR para el segmento compañero, MQ: Calidad del compañero, AS: Puntuación de alineación generada por el alineador, X?: Reservada para usuarios finales.

Obtenido el alineamiento y estudiada la calidad del mismo se ha realizado el filtrado de los datos para poder eliminar aquellas lecturas de baja calidad (MAPQ < 30, no superan los parámetros de calidad del proveedor) y que están correctamente mapeadas (excluir lecturas que no se encuentras mapeadas y aquellas que no tienen una pareja). También se ha procedido a eliminar las lecturas duplicadas a través de la herramienta *RmDup*.

Pileup se ha utilizado para identificar aquellas secuencias que son de calidad, en esta ocasión, se han eliminado del conjunto de datos aquellas lecturas cuya calidad era menor que 30 y que la cobertura de lectura era menor que 10, de esta forma, ha quedado un conjunto de datos de 844 líneas con el siguiente formato.

Para ver las lecturas seleccionadas, consultar la tabla adjunta en el Anexo V.

Tabla 4. Muestra resultados pileup.

Cr.	**Pos.**	**Ref**	**Cons**	**Q**	**SNPQ**	**# lecturas**	**# bases**	**Secuencia**	**Q secuencia**	**C11**	**#A**	**#C**	**#G**	**#T**	**#total**	**C17**
chr1	1233723	A	W	8	8	60	10	..T,,.t.,,	DD><JHECDD	]]]]]]]]]	7	0	0	1	8	1
chr1	1637371	A	A	22	0	60	12	..,....,.T..	@DE8DBDCIJJD	]]]]]]]]]]]]	10	0	0	1	11	1
chr1	1657220	G	R	3	3	59	15	,$,..,,..,aa,.,.	CFFJJIE7JGJHJDF	U]]]]]]]]]]]]]]	2	0	12	0	14	2

4.4. Visualización de los resultados intermedios

Antes de proceder con el filtrado y anotación de variables, se muestra gráficamente el resultado de los alineamientos realizado usando IGV[6].

El alineamiento aquí realizado ha sido de genoma completo, por lo que resulta muy difícil explorar el archivo completo, por lo que se va a mostrar un cromosoma que se ha detectado que tiene una mayor proporción de lecturas y, además, en el fichero generado con *pileup* se ha localizado una región con muchos alineamientos. Como se ha visto con anterioridad, uno de los cromosomas donde más lecturas había era el cromosoma *chr19* y, consultando el archivo generado en el pileup, se observa que la posición *8.966.182* tiene alrededor de 21 lecturas. En el visor de variantes se carga el genoma de referencia el resultado del alineamiento, en el buscador se selecciona la región *chr19:8,965,958-8,966,407.* En la Figura 6, en la parte superior se puede visualizar la cobertura del alineamiento y se observa como en la zona central, el número de lecturas es mucho mayor. Marcado en diferente color, se puede ver las diferentes variantes existentes, donde en el cuadro superior se observa que en la posición *8.966.182* hay 11 C y 9 T. Al pulsar sobre una variante se puede ver la información más detallada.

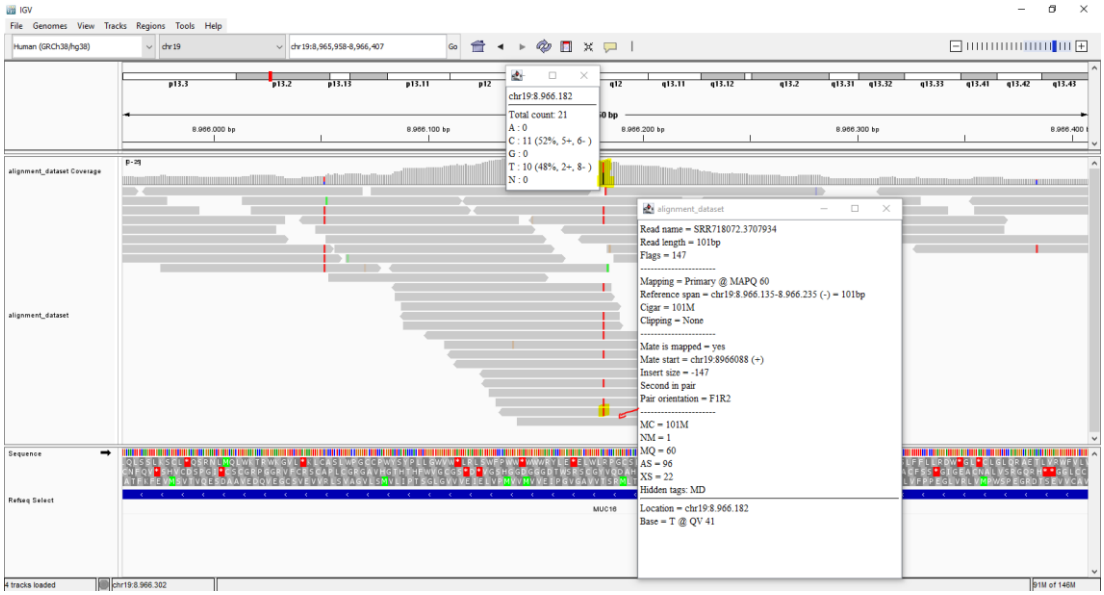

Figura 6. Visor de variantes IGV.

Si se accede al visor de la UCSC[7] y se accede a la misma región, con los filtros que aparecen por defecto (Figura 7) se pueden observar las variantes encontrada en las diferentes lecturas y, además, en la parte inferior, se puede ver que la variante está identificada como *rs2547074.*

[6] Integrated Genomic Viewer.
[7] UCSC Genome Browser

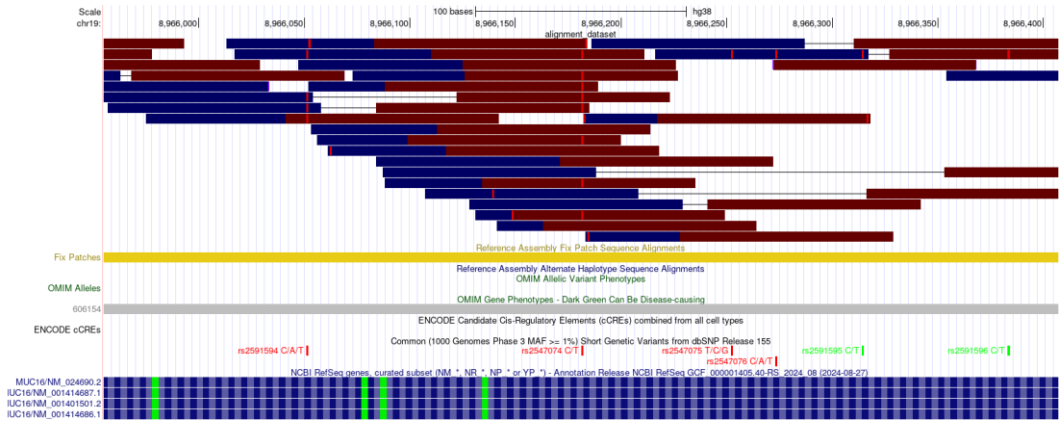

Figura 7. Visor de variantes UCSC.

Dado que la variante corresponde a un polimorfismo conocido, se puede consultar en la base de datos de *Ensemble*[8] y allí se puede ver que esta variante corresponde a una *missense variant*, como resultado la cadena de aminoácidos se modificará. El alelo de referencia es C y la variante observada es el alelo T. Es una variante observada en un 35% de la población y está clasificada como *benigna* y se encuentra asociada al fenotipo MUC16.

Al igual que se ha realizado el estudio de un área específica del alineamiento resultante, se podrían explorar otras áreas de la secuencia.

4.5. Identificación de diferencias genéticas

Una vez que se ha estudiado la calidad del alineamiento se ha realizado la identificación y anotado de variantes utilizando diferentes herramientas proporcionadas por *Galaxy*. A continuación, se muestran los dos estudios que se han llevado a cabo.

Variant calling Free Bayes

FreeBayes es un identificador de variantes que evalúa la probabilidad de cada genotipo posible para cada posición en el genoma de referencia, dadas las lecturas observadas en esa posición y devuelve la lista de posibles variantes. Como el objetivo del estudio es detectar variantes minoritarias, se crea un filtro para que sólo tenga en cuenta aquellas variantes cuya profundidad sea mayor o igual a 30 y la calidad mayor a *30,* usando *VCFFilter.* Finalmente, se normalizará el VCF para poder alinear las *indels* a la izquierda.

Se obtiene un total de 1381 líneas después de realizar la identificación de variantes con *FreeBayes* ya que se realizó un primer filtrado donde la cobertura no podía ser menor de 10 y, al aplicar *VCF Filter,* el resulta ha sido un total de 25 variantes identificadas, como se puede ver en la siguiente tabla.

Tabla 5. Identificación de variables con FreeBayes después de filtrar.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	FORMAT	VALUES
chr1	13369246	.	T G	C A	130,64	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:49:38,11:38:1346:11:407:-21.4362,0,-105.373
chr1	13369358	.	G	A	1,71E-04	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:37:33,4:33:1252:4:155:-1.32057,0,-101.761

[8] Ensemble

chr1	133693 84	.	G	A	4,89E-03	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:35:31,4:31:1157:4:154:-2.2769,0,-93.4255
chr1	109690506	.	T	C	1,57E-09	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:38:36,2:36:1328:2:82:0,-3.65946,-111.757
chr1	109690516	.	G	C	302	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:36:21,15:21:772:15:529:-36.8008,0,-58.9594
chr1	120763696	.	A	T	3,29E-09	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:31:29,2:29:1106:2:51:0,-4.48722,-94.5738
chr1	146989687	.	A	G	0,00028421	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:30:26,4:26:992:4:146:-4.465,0,-79.8195
chr10	87358368	.	C	T	542,29	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:45:23,22:23:858:22:817:-59.6729,0,-63.9811
chr11	49871813	.	T	C	1,01E-08	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:33:31,2:31:1205:2:71:0,-3.19361,-101.735
chr17	18676396	.	T	C	3,22E-06	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:30:28,2:28:1062:2:79:0,-1.53327,-88.3792
chr17	20357898	.	A	G	4,16E-08	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:30:28,2:28:1097:2:59:0,-3.42689,-93.3909
chr21	10464942	.	T	C	0,0323061	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:35:30,5:30:1142:5:193:-6.41322,0,-92.531
chr5	176106582	.	C	T	8	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:36:30,6:30:1150:6:226:-9.41334,0,-91.0732
chr7	75024408	.	A	G	353	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:31:16,15:16:574:15:561:-41.4937,0,-42.6573
chr7	75057560	.	T	A	5,13E-08	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:32:30,2:30:1110:2:72:0,-2.79705,-93.3557
chr7	76440312	.	G	A	186	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:40:28,12:28:1024:12:426:-26.0758,0,-79.8773
chrM	723	.	A	G	2161,55	.	GT:DP:AD:RO:QR:AO:QA:GL	1/1:66:0,66:0:0:66:2456:-221.003,-19.868,0
chrM	750	.	A	G	1944,81	.	GT:DP:AD:RO:QR:AO:QA:GL	1/1:61:0,61:0:0:61:2237:-201.55,-18.3628,0
chrUn_KI270746v1	35656	.	A	T	6,19E-08	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:33:31,2:31:1179:2:80:0,-2.434,-96.5844
chrUn_KI270746v1	35708	.	C	T	422	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:38:19,19:19:737:19:671:-47.3777,0,-54.7626
chrX	24207756	.	G	T	2,08E-06	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:30:28,2:28:1092:2:77:0,-1.7227,-91.262
chrX	52652882	.	T	A	7,84E-10	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:39:37,2:37:1394:2:81:0,-4.09925,-117.798
chrX	85108126	.	T	C	1,55E-09	.	GT:DP:AD:RO:QR:AO:QA:GL	0/0:38:36,2:36:1366:2:68:0,-5.05547,-113.891
chrX	85108134	.	A	G	1234,15	.	GT:DP:AD:RO:QR:AO:QA:GL	1/1:38:0,38:0:0:38:1425:-126.333,-11.4391,0

chrX	141003560	.	C	G	549	.	GT:DP:AD:RO:QR:AO:QA:GL	0/1:34:12,22:12:433:22:820:-62.7379,0,-29.0305

Call Variants with LoFreq

Como el objetivo del presente estudio es detectar variantes minoritarias, es decir, que tienen una frecuencia baja de aparición en la población se ha seleccionado esta herramienta para detectar este tipo de variantes específicas. Seguidamente, igual que el flujo anterior se ha normalizado el alineamiento de los *indels.*

Para realizar el filtrado de variantes se ha indicado en la configuración que sólo tenga en cuenta variantes de tipo *SNP* e *indels* y que la cobertura sea mayor de 30. El resultado de este filtrado es un total 16 variantes como se puede ver en la Tabla 6.

Tabla 6. Identificación de variantes usando lofreq.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr1	13369246	.	T	C	262	PASS	DP=50;AF=0.220000;SB=0;DP4=25,14,7,4
chr1	13369247	.	G	A	262	PASS	DP=49;AF=0.224490;SB=0;DP4=24,14,7,4
chr1	13369358	.	G	A	83	PASS	DP=38;AF=0.105263;SB=2;DP4=17,17,3,1
chr1	13369384	.	G	A	94	PASS	DP=36;AF=0.111111;SB=9;DP4=16,15,4,0
chr1	109690516	.	G	C	424	PASS	DP=36;AF=0.416667;SB=1;DP4=7,14,4,11
chr1	146989687	.	A	G	79	PASS	DP=32;AF=0.125000;SB=0;DP4=13,13,2,2
chr5	176106582	.	C	T	157	PASS	DP=36;AF=0.166667;SB=0;DP4=11,19,2,4
chr7	75024408	.	A	G	452	PASS	DP=34;AF=0.470588;SB=3;DP4=10,7,12,4
chr7	76440312	.	G	A	286	PASS	DP=43;AF=0.279070;SB=0;DP4=12,18,5,7
chr10	87358368	.	C	T	673	PASS	DP=46;AF=0.500000;SB=6;DP4=7,16,12,11
chr21	10464942	.	T	C	101	PASS	DP=34;AF=0.117647;SB=2;DP4=11,19,2,2
chrX	85108134	.	A	G	1384	PASS	DP=38;AF=1.000000;SB=0;DP4=0,0,20,18
chrX	141003560	.	C	G	653	PASS	DP=37;AF=0.594595;SB=9;DP4=6,6,4,18
chrM	723	.	A	G	2427	PASS	DP=67;AF=0.985075;SB=0;DP4=0,0,26,40
chrM	750	.	A	G	2188	PASS	DP=61;AF=1.000000;SB=0;DP4=0,0,22,39
chrUn_KI270746v1	35708	.	C	T	567	PASS	DP=38;AF=0.500000;SB=2;DP4=12,7,9,10

Se puede observar que, en ambos procesos hay un total de 16 variantes coincidentes. La única variante que no tiene en cuenta es el *chr1* posición *13369247,* aunque se puede observar, que en *FreeBayes* está considerada como una secuencia de 2 bases en lugar de un SNPs.

Las tablas completas de ambos estudios pueden visualizarse en el Anexo VI.

La identificación de variantes también puede ser visualizada y estudiada desde el explorado de genomas como se puede ver en la Figura 8.

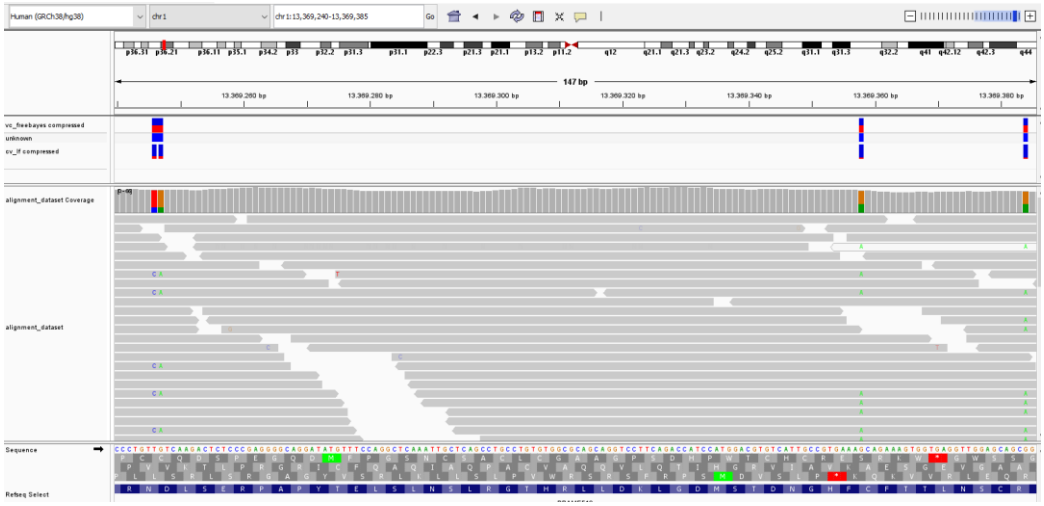

Figura 8. Visor de genomas. Variantes identificadas.

4.6. Filtrado y anotación de variables genéticas

Una vez que las variantes han sido identificadas, se procede a realizar su anotación para ver cuál es el posible efecto de esa variante sobre el genoma o incluso sobre los fenotipos de la persona.

Anotación de las variantes identificadas por FreeBayes

De las 25 anotaciones existentes en la identificación realizada por *FreeBayes,* se puede ver que 24 son polimorfismos de nucleótido único (SNP) y 1 polimorfismo múltiple y que no hay ni inserciones ni deleciones, así como ningún cambio estructural (Tabla 7).

Tabla 7. Número de variantes por tipo.

Type	Total
SNP	24
MNP	1
INS	0
DEL	0
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	25

Tabla 8.Nº efectos por impacto.

Type	#	%t
HIGH	1	4.348 %
LOW	4	17.391%
MODERATE	11	47.826%
MODIFIER	7	30.435%

Tabla 9. Impactos funcionales.

Type	#	%
MISSENSE	10	66.667%
NONSENSE	1	6.667 %
SILENT	4	26.667%

Revisando el impacto que las diferentes variantes pueden tener sobre la secuencia, se puede ver que hay 7 variantes que tienen un efecto modificador, 11 tienen un efecto

moderado, 1 tienen un impacto alto y 4 bajo. Funcionalmente, hay 10 variantes que producen un cambio en la cadena de proteínas (*missense*), 1 que producen una parada (*nosense)* y 4 silenciosas.

Anotación de las variantes identificadas por loFreq

Dado que la identificación de variables realizada con *loFreq* es un subconjunto de la realizada con *FreeBayes*, los resultados obtenidos son muy similares.

Tabla 10. Número de variantes por tipo.

Type	Total
SNP	16
MNP	0
INS	0
DEL	0
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	25

Tabla 11.Nº efectos por impacto.

Type	#	%t
HIGH	0	0
LOW	3	21.429%
MODERATE	8	57.143%
MODIFIER	3	21.429%

Tabla 12. Impactos funcionales.

Type	#	%
MISSENSE	8	72.727 %
NONSENSE	0	-
SILENT	3	27.273 %

Se pueden ver informes más detallados en el Anexo VII.

Además, inspeccionando el visor de genomas, se ha observado que únicamente hay 2 polimorfismos identificados como se puede ver en la Figura 9.

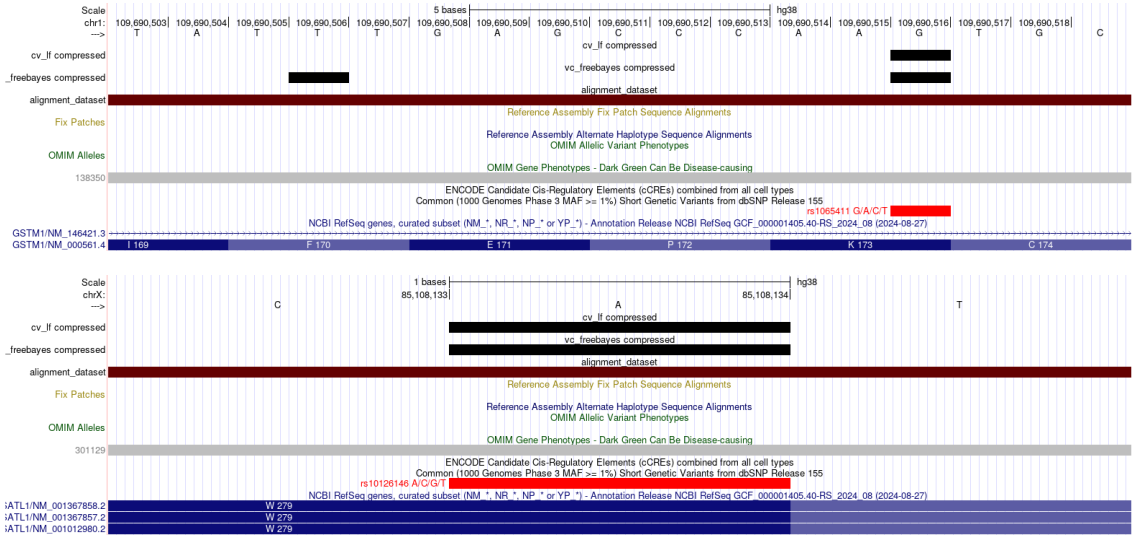

Figura 9. Variantes anotadas.

El polimorfismo *rs1065411*[9] corresponde a una variante de tipo *missense*, está localizado el cromosoma 1 en la posición *109.690.516*. Tiene una frecuencia en la población de 0.48 y puede tomar cualquier valor y no parece tener ningún fenotipo asociado.

[9] Polimorfismo rs1065411

El polimorfismo *rs10126146*[10] corresponde a una variante de tipo *missense*, está localizado el cromosoma X en la posición *85.108.134*. Tiene una frecuencia en la población de 0.42 y puede tomar cualquier valor siendo el ancestro el nucleótido G. Este polimorfismo está asociado con la condición SATL-1 pero no parece provocar un gran impacto sobre la condición humana.

5. Discusión, limitaciones y conclusiones del estudio

El objetivo del presente estudio ha sido detectar variantes minoritarias dentro de la secuencia proporcionada sobre el genoma humano de referencia (hg38). Se partía de una secuenciación de genoma completo y se ha realizado un filtrado de variables manejable en número aplicando criterios de calidad muy altos. En próximos estudios habría que verificar los criterios utilizados y estudiar otras posibles herramientas que permitan detectar variantes minoritarias.

Por otra parte, se ha observado que sólo había dos polimorfismos etiquetados de los 25 detectados, por lo que sería interesante consultar otras bases de datos para verificar si las variantes identificadas están documentadas y tienen algún efecto o no sobre los fenotipos de las personas. Una vez confirmado que se ha encontrado alguna variante no documentada, se podrían subir a repositorios públicos que sirviera para alimentar estudios poblacionales.

Otra forma de centrar el estudio podría ser definir un problema objetivo como el "efecto del fármaco X en la cura de la enfermedad Y", para ello, sería necesario identificar diferentes grupos de estudio (casos-control) para poder identificar las variantes que hacer que unos pacientes sean resistentes a un determinado fármaco.

Finalmente, otras herramientas comerciales como, por ejemplo *(https://franklin.genoox.com/analysis-tool/join-cta)*, podrían ser utilizadas para la secuenciación, alineación e identificación de variantes.

También sería útil poder contar con personas expertas en el campo de la medicina, genética o farmacológico que permitan analizar y obtener conclusiones veraces de los resultados obtenidos.

6. Notas

Todos los anexos generados para el presente documento se encuentran disponibles en el repositorio de Github: https://github.com/VegaUOC/Rodrigalvarez-Chamarro-MariadelaVega-PEC3.git

[10] Polimorfismo rs10126146

7. Referencias

[1] I. Galán Chilet, *Identificación de variantes genéticas poco frecuentes y raras en diabetes mellitus tipo 2 mediante secuenciación de exoma,* Valencia, 2015.

[2] N. Lozano, V. Lanza, J. Suárez-González, M. Herranz, P. Sola-Campoy, C. Rodríguez-Grande, S. Buenestado-Serrano, M. Ruiz-Serrano, G. Tudó, F. Alcaide, P. Muñoz, D. García de Viedma y L. Pérez-Lago, «Detection of Minority Variants and Mixed Infections in Mycobacterium tuberculosis by Direct Whole-Genome Sequencing on Noncultured Specimens Using a Specific-DNA Capture Strategy,» *mSphere,* 2021.

[3] S. Gianella y D. Richman, «Minority variants of drug-resistant HIV,» *Infect Dis,* vol. 202, nº 5, pp. 657-66, 2010.

[4] C. Benner, «Homer. Understanding and manipulating SAM/BAM alignment files,» UCSD, [En línea]. Available: http://homer.ucsd.edu/homer/basicTutorial/samfiles.html. [Último acceso: enero 2025].

[5] A. Rougon, «6.2 Formatos de Datos Biológicos,» Plants & Python, [En línea]. Available: https://plantsandpython.github.io/PlantsAndPython/L_6_MINERIA_DE_DATOS_EN_LA_LINEA_DE_COMANDOS_DE_UNIX/0_Lecciones/6.2_Formatos_Biologicos.html. [Último acceso: enero 2025].

[6] The SAM/BAM Format Specification Working Group, «Sequence Alignment/Map Optional Fields Specification,» 2024.

[7] Wikipedia, «SAM (file format),» [En línea]. Available: https://en.wikipedia.org/wiki/SAM_(file_format). [Último acceso: Enero 2025].

8. Apéndices

9.1. Informe calidad de la primera secuencia

Estadísticas básicas

Tabla 13. Estadísticas básicas secuencia 1.

Measure	Value
Filename	sampleDat6_1_fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000000
Total Bases	101 Mbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	47

Calidad de la secuencia por base

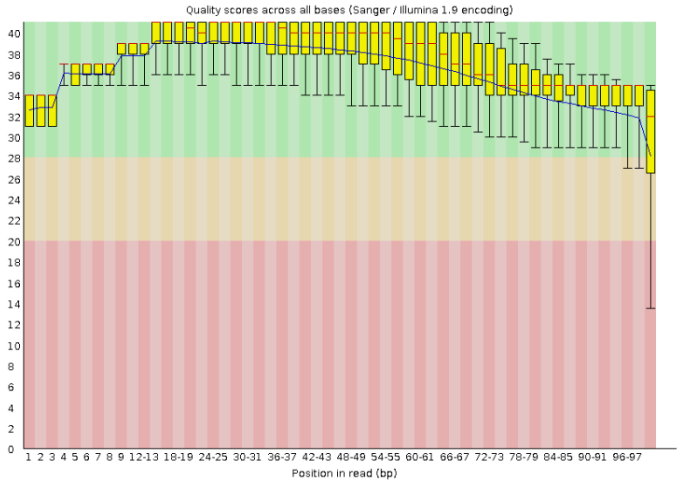

Figura 10. Calidad de la secuencia 1 por base.

Calidad de la secuencia por área

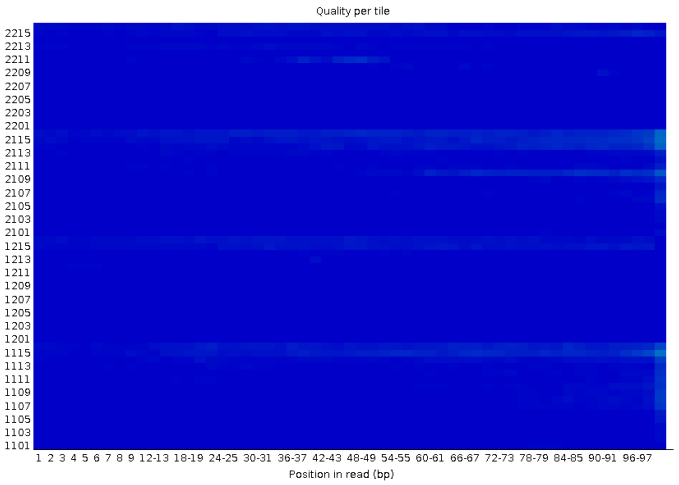

Figura 11. Calidad de la secuencia 1 por área.

Puntuaciones de calidad por secuencia

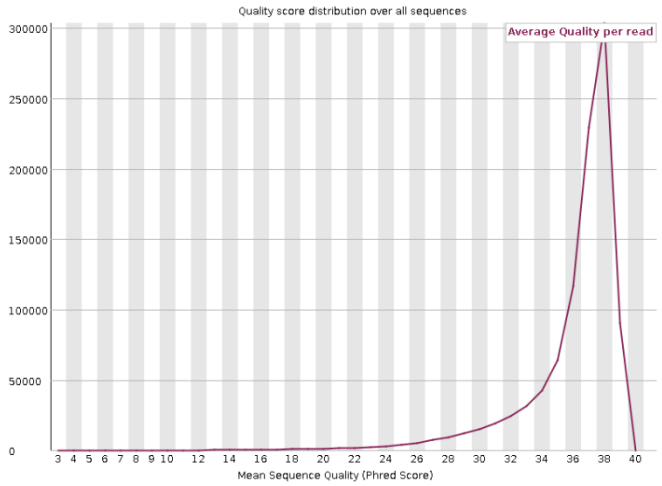

Figura 12. Puntuación de calidad secuencia 1.

Contenido por secuencia de bases

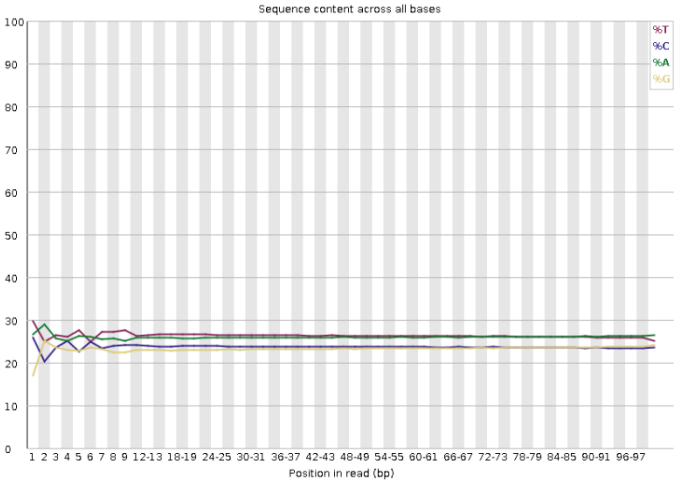

Figura 13. Contenido por secuencia de bases muestra 1.

Contenido GC por secuencia

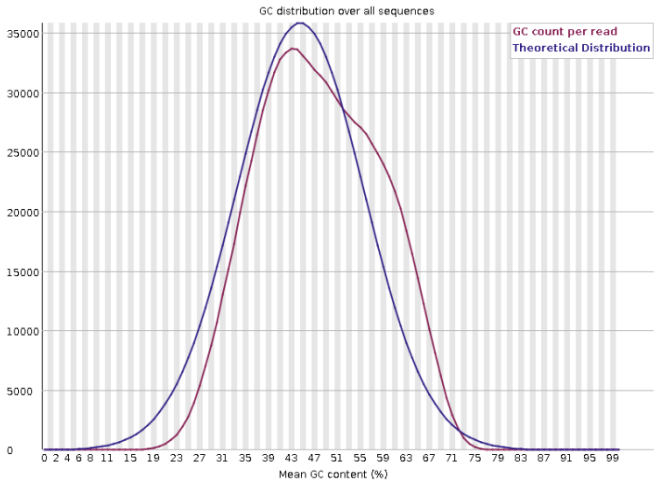

Figura 14. Contenido GC por secuencia muestra 1.

Distribución de la longitud de las secuencias

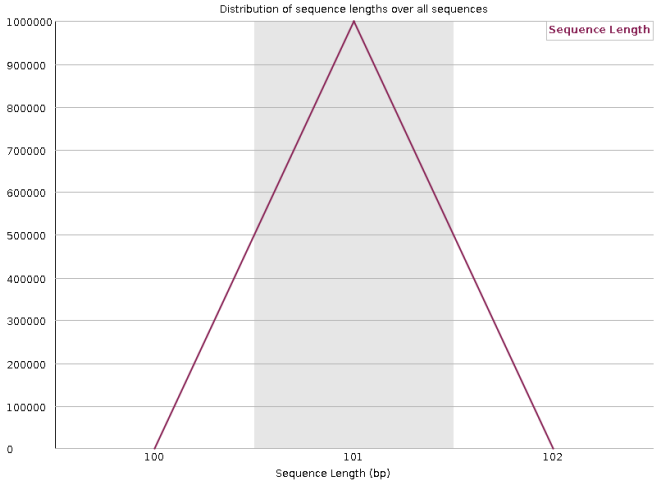

Figura 15. Distribución longitud de las secuencias muestra 1.

Niveles de duplicación de secuencias

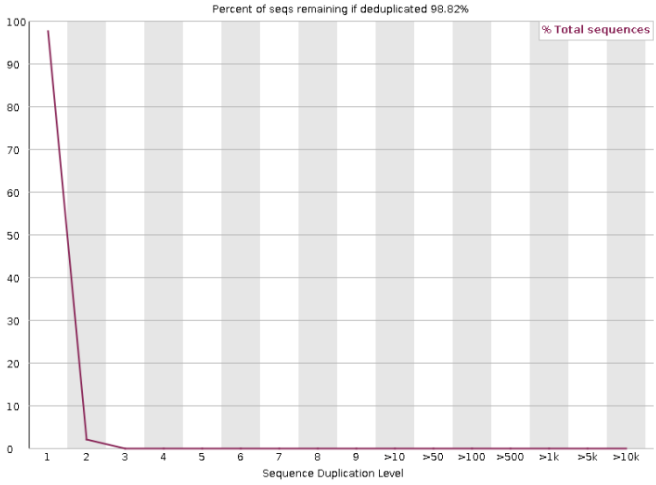

Figura 16. Niveles de duplicación secuencias muestra 1.

9.2. Informe calidad de la segunda secuencia

Estadísticas básicas

Tabla 14. Estadísticas básicas secuencia 2.

Measure	Value
Filename	sampleDat6_2_fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000000
Total Bases	101 Mbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	47

Calidad de la secuencia por base

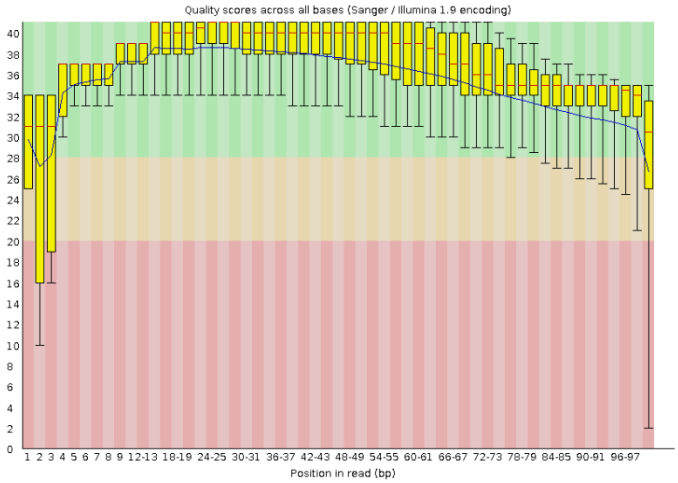

Figura 17. Calidad de la secuencia 2 por base.

Calidad de la secuencia por área

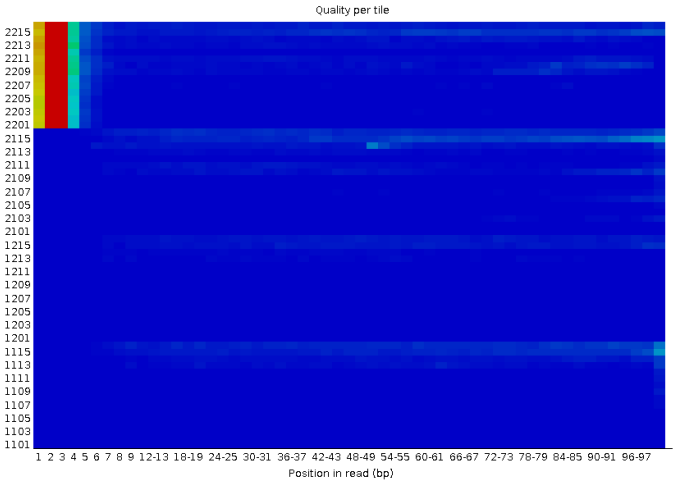

Figura 18. Calidad de la secuencia 2 por área.

Puntuaciones de calidad por secuencia

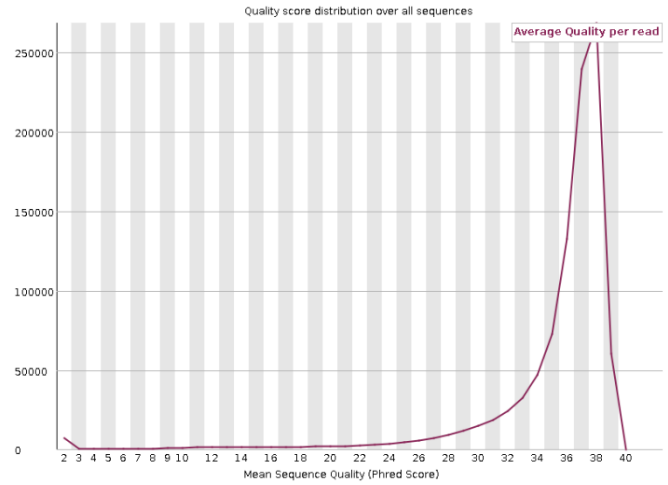

Figura 19. Puntuación de calidad secuencia 2.

Contenido por secuencia de bases

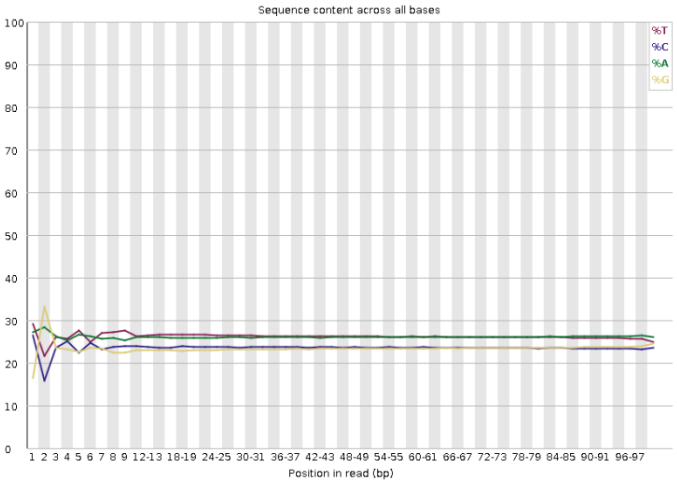

Figura 20. Contenido por secuencia de bases muestra 2.

Contenido GC por secuencia

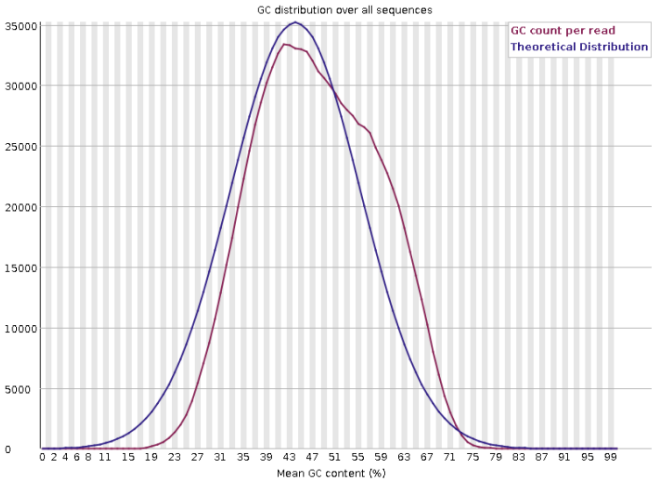

Figura 21. Contenido GC por secuencia muestra 2.

Distribución de la longitud de las secuencias

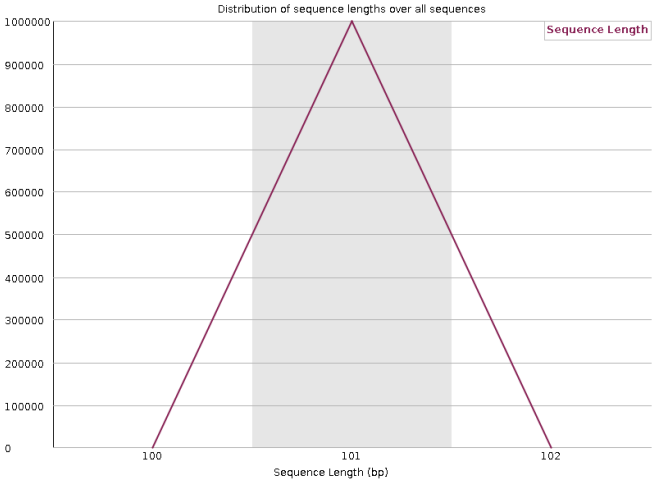

Figura 22. Distribución longitud de las secuencias muestra 2.

Niveles de duplicación de secuencias

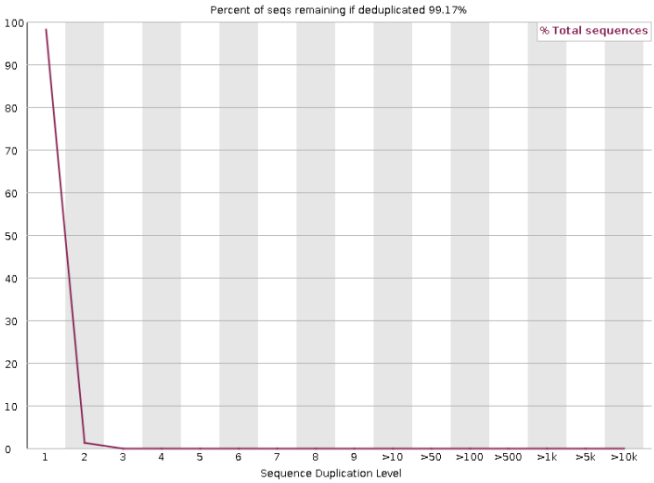

Figura 23. Niveles de duplicación secuencias muestra 2.

9.3. Informe calidad de ambas secuencias

Estadísticas generales

Tabla 15. Estadísticas generales ambas muestras.

Sample Name	% Dups	% GC	M Seqs
sampleDat6_1_fq	1.2%	47%	1.0M
sampleDat6_2_fq	0.8%	47%	1.0M

Recuento de secuencias

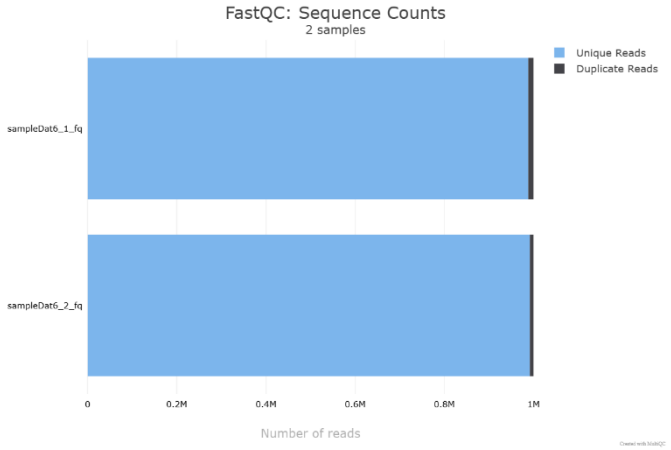

Figura 24. Recuento de secuencias por muestra.

Calidad de las secuencias

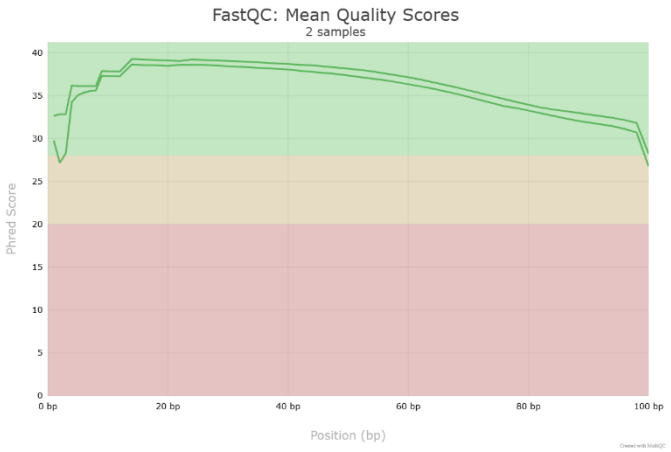

Figura 25. Calidad de las secuencias.

Puntuaciones de calidad por secuencia

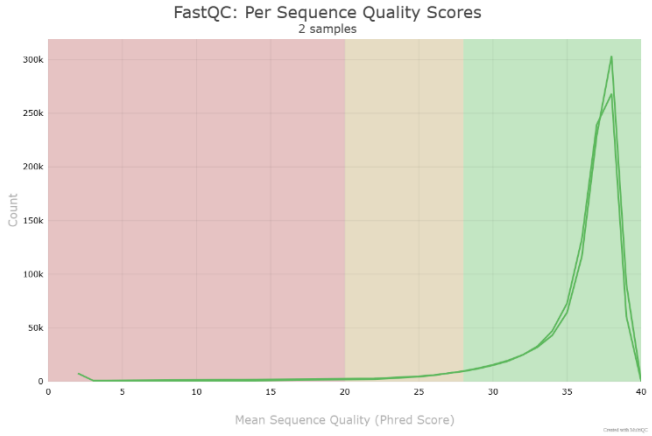

Figura 26. Puntuación de calidad.

Contenido base N

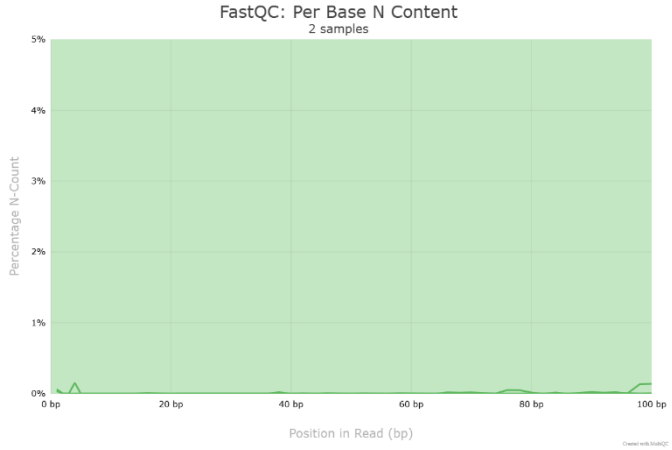

Figura 27. Contenido base N.

Contenido de adaptadores

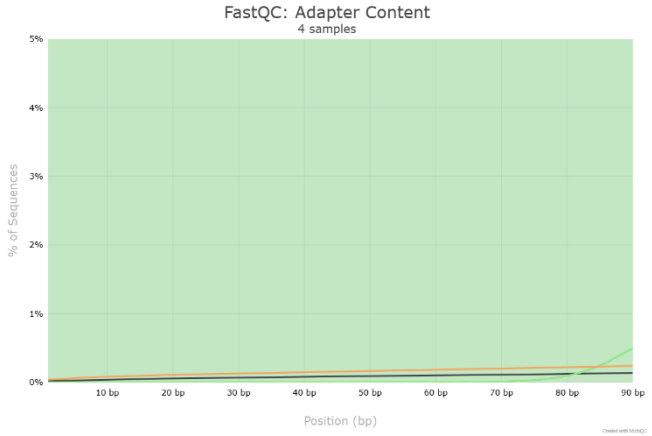

Figura 28. Contenido de adaptadores.

Comprobación de calidad global

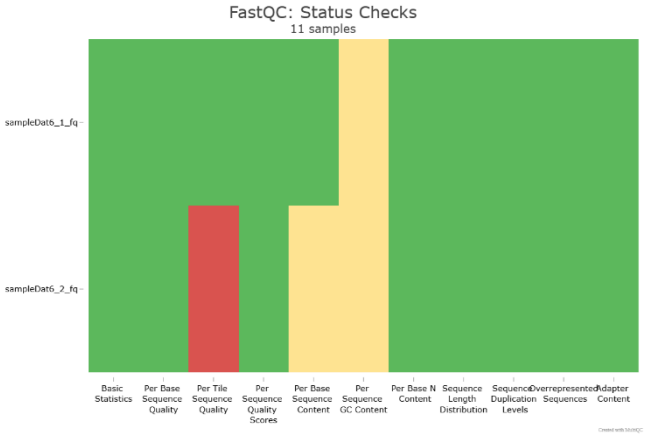

Figura 29. Comprobación de calidad global.

9.4. Calidad de los alineamientos

Lecturas mapeadas por cromosoma

Tabla 16. Lecturas mapeadas por cromosoma.

Id. Cromosoma	Longitud	Nº mapeos	Nº lecturas no mapeadas	% chr mapeado
chr1	248956422	215983	294	0,0868
chr2	242193529	169850	204	0,0701
chr3	198295559	114137	135	0,0576
chr4	190214555	80918	100	0,0425
chr5	181538259	94201	109	0,0519
chr6	170805979	81717	85	0,0478
chr7	159345973	97581	124	0,0612
chr8	145138636	57398	63	0,0395
chr9	138394717	80783	105	0,0584
chr10	133797422	83242	104	0,0622
chr11	135086622	97398	134	0,0721
chr12	133275309	92345	105	0,0693
chr13	114364328	37016	50	0,0324
chr14	107043718	57044	59	0,0533
chr15	101991189	64786	86	0,0635
chr16	90338345	72839	93	0,0806
chr17	83257441	99533	126	0,1195
chr18	80373285	29013	33	0,0361
chr19	58617616	81883	120	0,1397
chr20	64444167	40846	47	0,0634
chr21	46709983	20046	28	0,0429
chr22	50818468	39842	55	0,0784
chrX	156040895	129339	146	0,0829

chrY	57227415	2129	1	0,0037
*	0	0	1764	

Flag statistics. Calidad de las lecturas

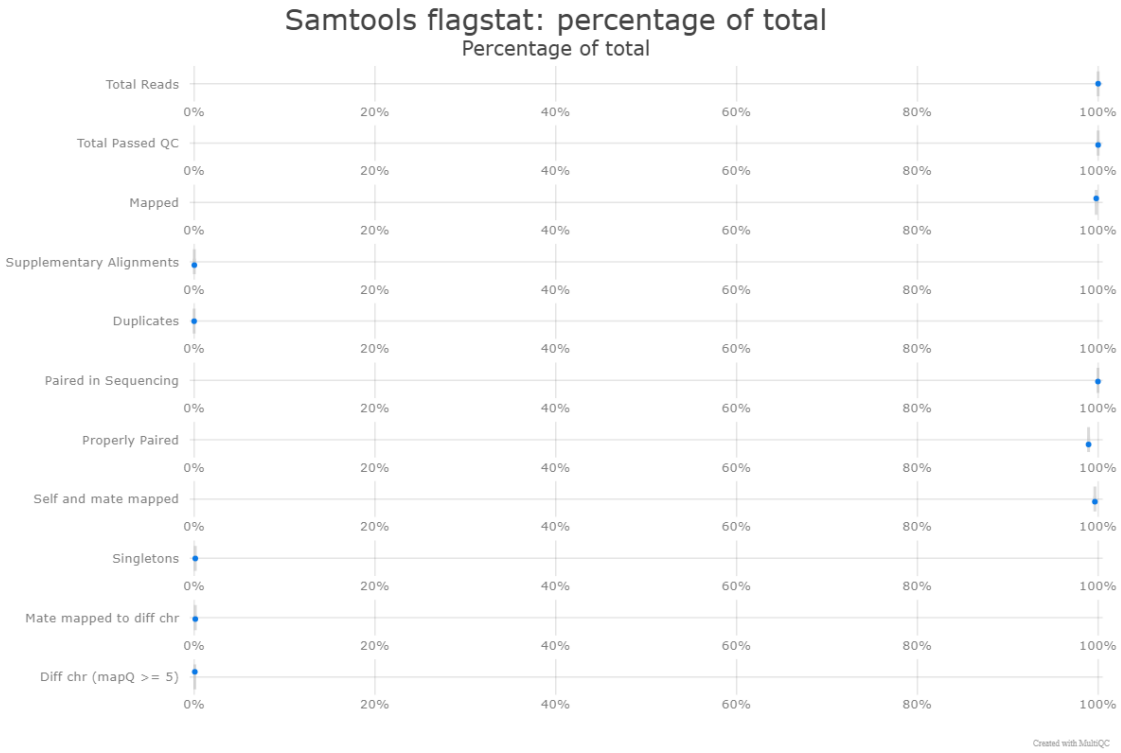

Figura 30. Flag stat. Percentage.