PEC3 - Análisis de Datos Ómicos

María Luisa Reyes Conde

Contents

1 Resumen

Este informe presenta un análisis bioinformático de una muestra paired-end FASTQ de datos de secuenciación genómica que corresponden a la muestra HG00128 del proyecto *1000 Genomes*. Para ello, se usó la plataforma Galaxy y se implementó un pipeline básico pero completo para realizar un control de calidad, un alineamiento contra el genoma humano (GFCh38 o gh38), identificación de variantes y anotación. Se usaron numerosas herramientas que están incluidas en la plataforma, como son `fastp`, `BWA-MEM` o `FreeBayes`, entre otras. Gracias a este pipeline, podemos analizar los datos de secuenciación de la muestra HG00128 para detectar variantes genéticas y conocer si existe alguna relevancia biológica en ellas.

2 Objetivos

El objetivo de este análisis bioinformático de la muestra HG00128 es, en general, realizar una identificación y anotación de las variates genéticas en una muestra humana y explorar su relevancia biológica. Además, de manera más específica, se pretende controlar las herramientas de la plataforma Galaxy:

- Evaluar las lecturas FASTQ mediante las herramientas de filtrado y de control de calidad,
- Alinear las lecturas al genoma humano de referencia y detectar variantes genéticas en la muestra,

- Anotar dichas variantes en función del impacto funcional para predecir su relevacia biológica y, por último,
- Interpretar de manera correcta los resutlados en un contexto biológico.

3 Métodos

Dividimos esta sección para poder comentar con claridad los métodos usdos en este análisis. ## Datos Los datos usados en este trabajo está sacada del proyecto *1000 Genomes* [8], más concretamente se trata de la muestra HG00128. Los datos de esta muestra se presentan en formato FASTQ (*paired-end reads*) pero para reducir la carga computacional se escogió aleatoriamente una pareja de las 10 muestras independientes que se sdeleccionaron. Para la selección aleatoria de la pareja, se usó el siguiente código en R:

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

Como el resultado obtenido es el número 9, hemos trabajado con los archivos 'sampleDat9_1.fq' (corresponde a las lecturas forward) y 'sampleDat9_2.fq' (lecturas reverse). Ambos archivos los subimos a la plataforma Galaxy para poder realizar el análisis. Además, creamos un archivo del tipo *paired* y o denotamos como 'sample9_paired' seleccionando ambos archivos subidos para poder comenzar con el preprocesamiento de los datos.

3.1 Preprocesamiento de los datos y estudio de su calidad

Para el preprocesamiento de los datos usamos la herramienta `fastp` [2] de la platafroma Galaxy [7]. Para ello, hemos usado seleccionado la opción 'Paired Collection' y seleccionamos como input nuestros datos emparejados 'sample9_paired'. Las opciones *adapter trimming*, *quality filtering* y *length filtering* las dejamos activadas (vienen así por defecto). Lo que sí modificamos son los valores para los filtros: dejamos un mínimo de *quality phred* de 20 (por defecto es 15, así que lo ponemos un poco más estricto) y una longitud mínima de 30 bases (aquellas lecturas menores de 30 bases son eliminadas) aunque sin máximo de longitud.
Una vez que se ha realizado el preprocesameinto de los datos, usamos la herramienta `Falco`, que es una alternativa más eficiente a la herramienta `FastQC` (y además recomendada por la propia plataforma). Para usar esta herramienta es tan fácil como aplicarla a cada uno de los archivos recortados que han salido del preprocesamiento.

3.2 Alineamiento al genoma de referencia

Para esta parte del análisis, usamos la herramienta `BWA-MEM` [4] y usamos como genoma de referencia el humano (GRCh38 o hg38). Tenemos que seleccionar los dos archivos recortados que nos dio como outputs la herramienta `fastp`. En este caso, no hace falta modificar los parámetros: el modo análisis lo dejamos en *Simple Illumina mode* y el modo de orden BAM (*BAM sorting mode*) se realiza por cromosomas ('*Sort by chromosomal coordinates*'). Ordenamos el archivo `.bam` que hemos btenido con la herramienta con `samtools sort`, dejando el orden en *coordinate* (por defecto) y posteriormente, lo indexamos con `bamCoverage`. Esto nos será de utilidad a la hroa de visualizar los resultados. Por último, para visualizar las estadísticas del alineamiento usaremos `Samtools Flagstat`.

3.3 Identificación de variantes

Para detectar variantes entre las lecturas alineadas y el genoma humano, usamos la herramienta `FreeBayes` [5]. Para ello, seleccionamos como *BAM input* el output generado por la herramienta anterior (alnieamiento, la BWA-MEM) y como *reference genome* el `hg38` (genoma humano). El resto de parámetros también los dejamos como vienen por defecto. Esta herramienta nos genera un archivo de variantes que viene dado en formato `.vcf`. Además, para quedarnos con las variantes de mejor calidad, hacemos un filtrado con la herramienta `VCFfilter` que posean calidad superior a 30 y con mas de 10 lecturas de soporte.

3.4 Visualización de resultados

Para tener una imagen y observar los resutlados que vamos obteniendo, basta con hacer click en el símbolo del gráfico que aparece en el output que queramos consultar de nuestro historial y visualizarlo con la herramienta `Trackster`, que es el *genome browser* que viene integrdo en la plataforma de Galaxy.

3.5 Anotación

Usamos la herramienta `SnpEff` para obtener información sobre el contexto biologico de las variantes. De nuevo, usamos como genoma de referencia el humano, denotado en esta herramienta por `GRCh38.86`. Con esta herramienta obtendremos como resultado un archivo `.vcf` anotado y una tabla a modo resumen con los efectos genéticos encontrados.

4 Resultados

Como hemos comentado en el apartado "Métodos", de la muestra HG00128 se seleccionó de forma aleatoria un par de las 10 muestras independientes que se habían preseleccionado. Se realizó un preprocesamiento para poder analizar su calidad y obtuvimos la siguiente tabla con sus estadístics básicas:

Basic Statistics: pass

Measure	Value
Filename	fastp on data 4 and data 3
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	960933
Sequences Flagged As Poor Quality	0
Sequence length	31 - 101
%GC:	47

Figure 1: Estadísticas Falco

Como vemos, se han analizado un total de 960933 secuencias y ninguna se clasifica por tener mala

calidad, esto nos indica que el preprocesamiento ha sido bueno. Además, el contendio de GC es cercano al 50%, que es su valor óptimo y no existen secuencias sobrerepresentadas (lo cual es beuno). .Een la siguente gráfica osbervamos que los nucleótidos poseen una calidad > 30, que estén todas en verde nos indica que las lecturas son de alta calidad:

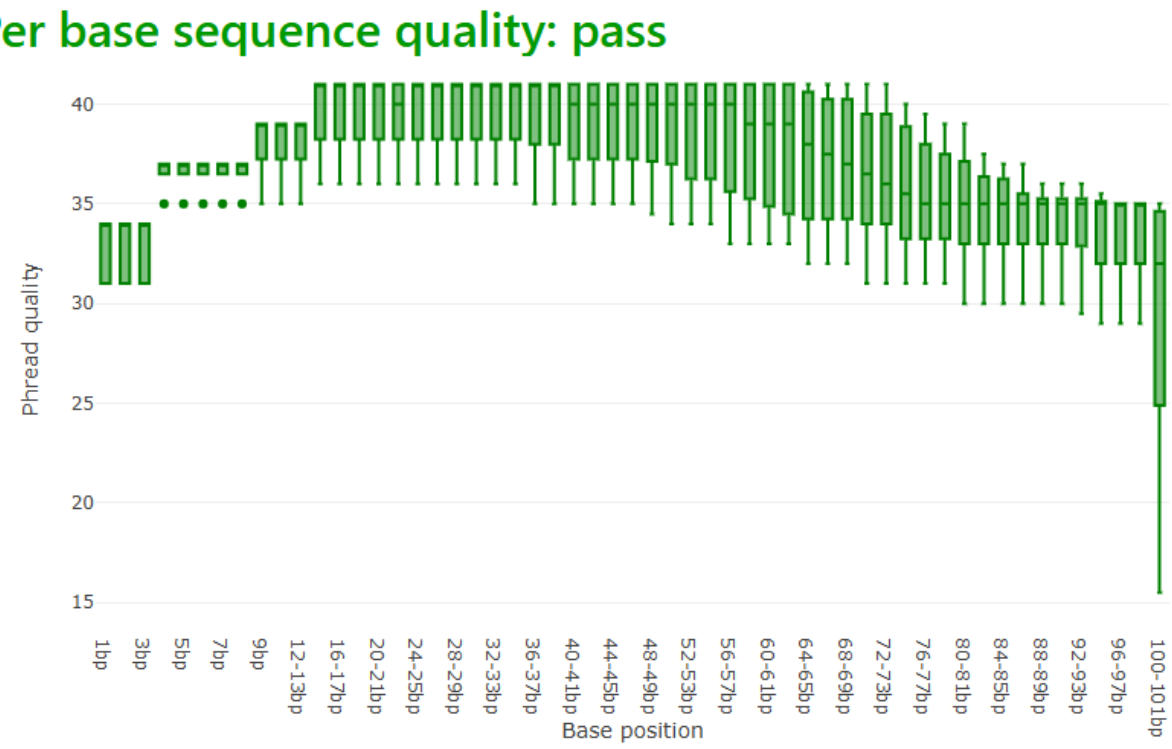

Figure 2: Calidad por bases de seq

Además, como hemos comentado, el porcentaje de GCs (presencia de bases G y C) es del 47%, lo cual es un valor bastante bueno. Podemos corroborarlo con la siguiente gráfica:

Vemos que la distributción de GCs es muy similar a la distribución teórica que sigue una campana de Gauss. Si la distribución tuviera desviaciones "raras", nos podría indicar que existe algún tipo de contaminación en los datos.
Por último, vemos que en el 'Adapter Content', todos los items están muy cerca del 0%, incluso el `PolyA`, que aunque parezca que se desvía mucho, llega al 0.12%, lo cual es buen señal ya que nos indica que el recorte del preprocesamiento se realizó de manera correcta:

A tener en cuenta: los gráficos presentados son los obtenidos para el primer archivo al que se le aplicó la herramienta `Falco`. Para el segundo, los resultados son muy similares. Hemos obtenido unos datos con buena calidad tras el trimming de la herramienta `fastp`.
Con nuestros archivos recortados realizamos el alineamiento contra el genoma de referecia: el humano, hg38. Gracias al uso de `Samtools Flagstat` sobre nuestro archuvo .bam ordenado, encontramos un total de 1922195 lecturas siendo mapeadas un 99.94% de ellas: 1921048 y con un 99.33% de estas estando emparejadas de forma correcta. Por último, hemos obtenido que hay 0 lecturas duplicadas, lo cual es un valor vastante óptimo, y sólo 7 singletons, que son aquellas secuencias que no pudieron ser emparejadas después del trimming y filtrado de las lectutas [9], qu eno hace ni el 0.1% de las lecturas de nuestro archivo, por tanto es un dato poco relevantw. Por lo tanto,

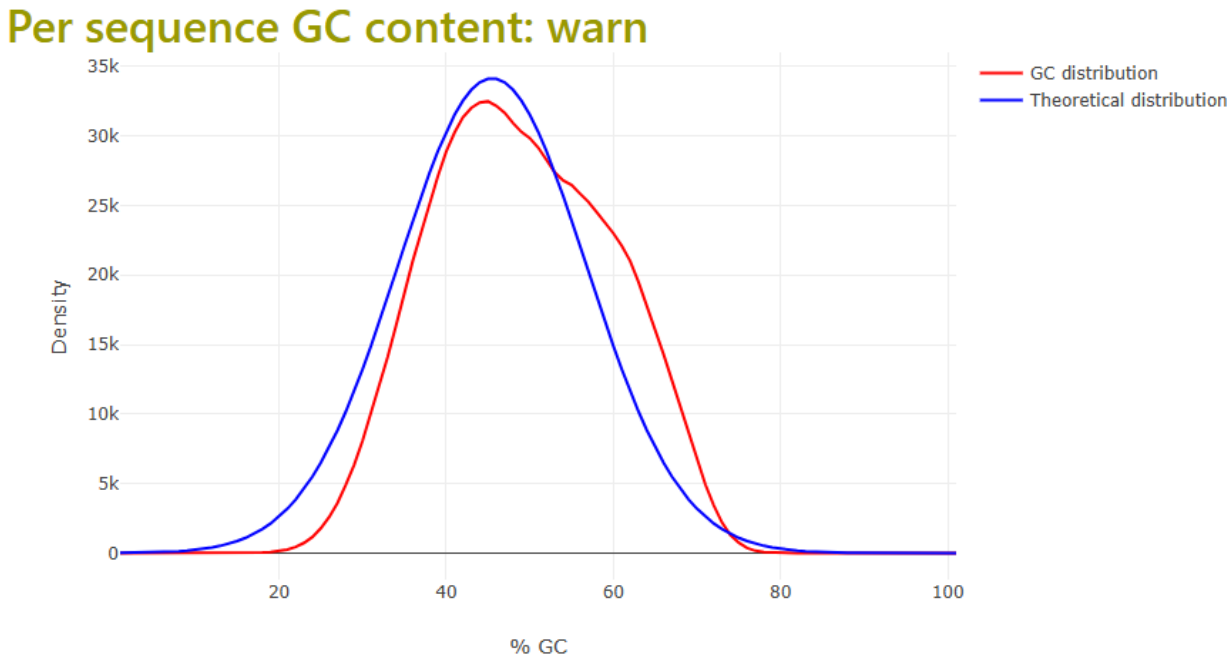

Figure 3: Contenido GC por seq

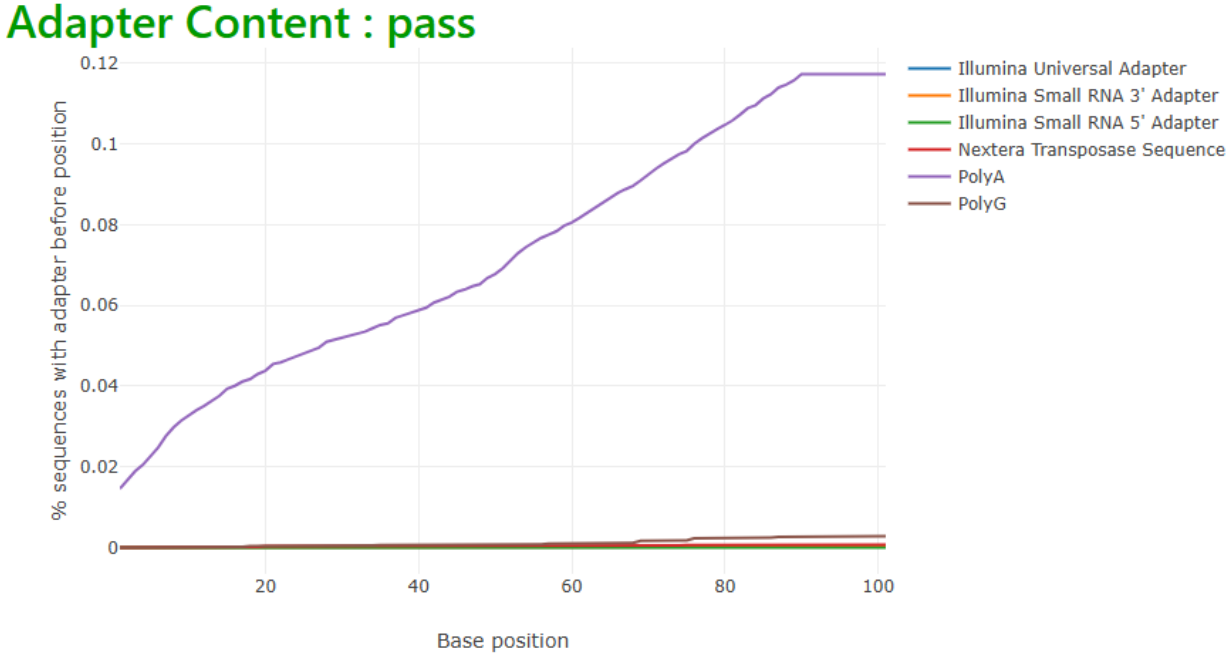

Figure 4: Adapter content

hemos obtenido un resultado muy bueno: prácticmente todas las lecturas han sido mapeadas y emparejadas correctamente y además no hay presencia de lecturas duplicadas. (Consultar el Anexo 1 para más información sobre las estadísticas de las lecturas ordenadas tras el alineamiento).

Una vez se ha realizado el alineamiento y sabemos que de manera exitosa, buscamos variantes entre las lecturas alineadas y el genoma humano. Observando al archivo .vcf sin filtrar, obtenemos que hay 31148 variantes, siendo el 89.5% variantes con una calidad que supera el umbral del 30%. Al ordenar e indexar nuestro archivo de variantes, obtenemos finalmente que, obligando a que las variantes superen un umbral del 30% de calidad y con más de 10 lecturas, tenemos 698 variantes en total. Para obtener más información sobre las variantes, se aplicó la herramienta `SnpEff`, que sirve para la anotación de estas. Gracias a esta herramienta, obtenemos un archivo .csv que podemos encontrar adjunto en el repositorio GitHub [1]. De aquí obtenemos que, en efecto, se han anotado 701 variantes en total y se dividen en 5 tipos: el 2.43% son de tipo DEL (deleciones: cambios genéticos en los que se pierde una parte del ADN), el 0.57% del tipo INS (inserciones: un segmento de ADN se ha añadido a la secuencia original de un gen), el 0.29% del tipo MIXED, 1.71% MNPs (donde varios nucleótidos son diferentes entre individuos) y, la gran mayoría, el 95% son SNPs (un solo nucleótido en la secuencia de ADN difiere entre individuos, es un tipo específico de MNPs): son 666 variantes de este tipo. Este es resultado es normal, ya que los SNPs son el tipo más frecuente de variación genética en humanos. Encontramos que hay numerosos effects (consecuencias funcionales de una variante en la expresión o función de un gen o proteína) que podemos dividir por el impacto que tienen:

- Alto impacto: afectan a proteínas. Encontramos solo 7 de alto impacto.
- Impacto modeardo: por ejemplo, cambios de aminoácidos. Encontramos 514.
- Bajo impacto: sinónimos, encontramos 510.
- Modifier: efectos intrafónicos, intergénicos... Son de los que más encontramos: 1845.

También podemos clasificarlos por clase funcional:

- Missense: mutación en la cual un codón muta de forma que dirige la incorporación de un aminoácido diferente. Encontramos 487 de este tipo.
- Silent: variantes genéticas que no cambian la secuencia de aminoácidos de una proteína, aunque sí alteran la secuencia de ADN. Encontramos 422.

Por último, podemos observar que efectivamente el alineamiento es de calidad y que las lecturas, efectivamente, se encuentran bien alineadas al genoma humano. Por ejemplo, podemos ver en la siguiente región del cromosoma 1 entre las posiciones 16576063 al 16576531:

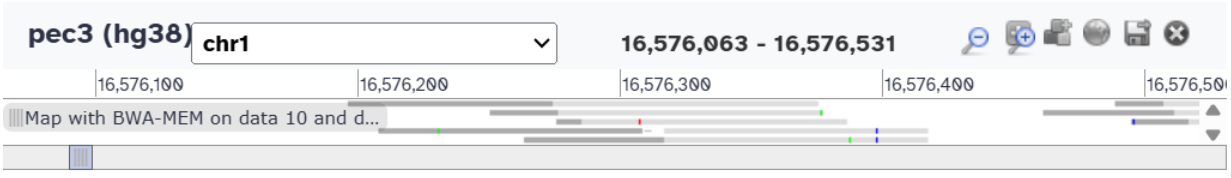

Figure 5: Chromosoma 1 vs hg38

Las barras que vemos son lecturas de secuenciación alineadas en esa región del genoma. Todo lo que se visualiza como gris son alineaciones comunes y los pequeños trozos de colores que se ven indican

mismatches, inserciones, deleciones u otras variaciones respecto al genoma de referencia. De hecho, si consultamos el barplot de cobertura o *coverage plot* del archivo `bigWig` que obtuvimos al aplicar la herramienta `bamCoverage` en las posiciones 16576003-16576603 (que contienen las posiciones que vemos en el alineamiento de la gráfica anterior), observamos que hay toda una región con cobertura total:

Figure 6: Cobertura chr1 en hg38

Las regiones que se observan con cobertura muy baja o nula, podríaindicar que tenemos errores de alineamiento o regiones que no están secuenciadas.
Por ejemplo, podemosobservar la cobertura general de nuestras lecturas con los cromosomas de sexo: X e Y. En el barplot de cobertura con el chromosoma X observamos que a lo largo de todo el cromosoma, hay regiones con picos de coberturas y otras en las que esta es menos o incluso nula en algunos sitios concretos, pero podemos decir que nuestras lecturas se encuentran representadas en el cromosoma X, aunque la cobertura no sea total y continua:

Figure 7: Cobertura chrX en hg38

Sin embargo, en el cromosoma Y solo encontramos cobertura al inicio y va disminuyendo. Por lo tanto, nuestras el cromosoma Y no está apenas cubierto por las lecturas que hemos analizado en nuestra muestra:

Figure 8: Cobertura chrY en hg38

Esto se puede deber a que nuestra muestra corresponda a una mujer, ya que normalmente en una muestra femenina (XX), lógicamente no se tiene un cromosoma Y completo. El que encontremos picos en algunas regiones específicas del Y se puede deber a regiones que son compartidas entre X e Y y pueden dar lecturas en ambas o a posibles artefactos de alineamiento.
En el caso de las variantes, en el archivo *.vcf* observamos cómo las variantes anotadas se encuentran en contigs y no en los cromosomas principales. Por ello, si visualizáramos el cromosoma 1 como hemos hecho con las lecutras, no observaríamos nada. Sin embargo, por ejemplo, en el cromosoma 4 tenemos un contig alternativo (representa regiones variantes del genoma humano) en el que está representado una variante (la línea vertial azul):

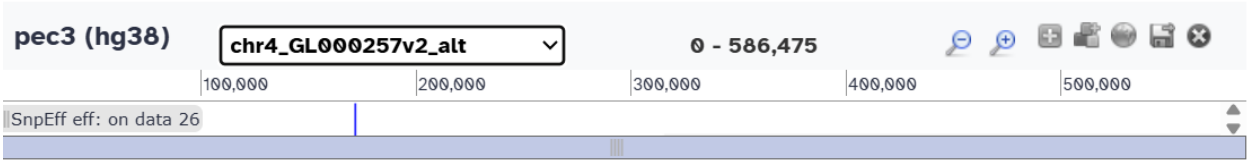

Figure 9: Variante en contig del chr4

O este caso de otra variante anotada en un contig alternativo del cromosoma 11:

Figure 10: Variante en contig del chr11

En conclusión, las variantes no están representadas en los cromosomas principales, sino en ciertas regiones de contigs.

5 Discusión

Aunque el pipeline implementado es bastante completo a la par de sencillo de usar, encontramos alguna que otra limitación. Como la muestra que hemos usado se ha seleccionado aleatoriamente de fragmentos de lecturas reales, esto limita la posibilidad d detectar variantes estructurales o eventos más complejos, además de que no se evalúa la cobertura real.
Una gran idea es usar la plataforma Galaxy para realizar este análisis. Esta plataforma es sencilla de usar incluso para personas sin experiencia en programación y sin estar familiarizado con Galaxy es fácil implementar las herramientas bioinformáticas requeridas para cada paso del pipeline. No obstante, hay ciertas restricciones en cuanto a la personalización de parámetros avanzados, sobre todo si no se ha estudiado con profundida el uso de cada herramienta.
Aún asì, con este pipeline hemos visto que es muy útil aplicar pipelines de análisis de variantes genéticas a datos reales de secuenciación, permitiendo obtener información relevante sobre ciertas variantes que pueden ser potencialmente funcionales en una muestra humana.
En resumidas cuentas, Galaxy es una herramienta sencilla pero posee ciertas limitciones si no se conoce a fondo. Además, sería interesante realizar un estudio de secuenciación más extenso, sobre la muestra completa o sobre más lecturas y además, realizarlo de nuevo cuando se implemente el genoma T2T-CHM13, que será el “nuevo” genoma de referencia humano, aún no disponible en Galaxy.

6 Conclusiones

Gracias a la implementación en la plataforma Galaxy de este sencillo pero completo pipeline bioinformático, hemos conseguido realizar un análisis de variantes genéticas en una pequeña parte de una muestra del proyecto *1000 Genomes*. Hemos hecho un análisis de calidad en las lecturas, alineado la secuencia al genoma humano (nuestro genoma de referencia) y hemos identificado y anotado las variantes genéticas que tienen posibilidad de poseer una implicación funcional. Aunque

nos encontramos limitaciones con respecto al tamaño de los datos con los que hemos trabajado, los resultados obtenidos son biológicamente interpetablels, lo que nos hace confiar en la firmeza del enfoque dado al análisis.
Gracias a este estudio, hemos podido familiarizarnos con el flujo de trabajo de análisis genómico, así como con la plataforma Galaxy y sus principales herramientas, que utilizadas en numerosos estudios de genómica computacional.

7 Referencias

- Repositorio de GitHub: [PEC3-Análisis de Datos Ómicos - María Luisa Reyes Conde](https://github.com/marisireyes/Reyes-Conde-MariaLuisa-PEC3) [1]
- FastQC. Babraham Bioinformatics. [2]
- Chen, S. et al. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. [3]
- Vasimuddin, M. et al. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. IEEE TPDS. [4]
- Garrison, E., Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907. [5]
- Cingolani, P. et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. [6]
- Afgan, E. et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses. Nucleic Acids Research. [7]
- The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature. [8]

8 Anexo

1. Estadísticas sobre las lecturas ordenadas (archivo .bam) tras el alineamiento contra el genoma humano (hg38): 1922195 + 0 in total (QC-passed reads + QC-failed reads)
   0 + 0 secondary
   329 + 0 supplementary
   0 + 0 duplicates
   1921048 + 0 mapped (99.94%:N/A)
   1921866 + 0 paired in sequencing
   960933 + 0 read1
   960933 + 0 read2
   1908922 + 0 properly paired (99.33%:N/A)
   1920712 + 0 with itself and mate mapped
   7 + 0 singletons (0.00%:N/A)
   1492 + 0 with mate mapped to a different chr
   872 + 0 with mate mapped to a different chr (mapQ>=5)