Taller 10 - Ensamble de genomas

Grupo 04 - sección 02

Integrantes:
- Nicolas Montoya Leon - 202310678
- John Anderson Acosta - 202212004
- Raquel Bautista Escobar – 202310296

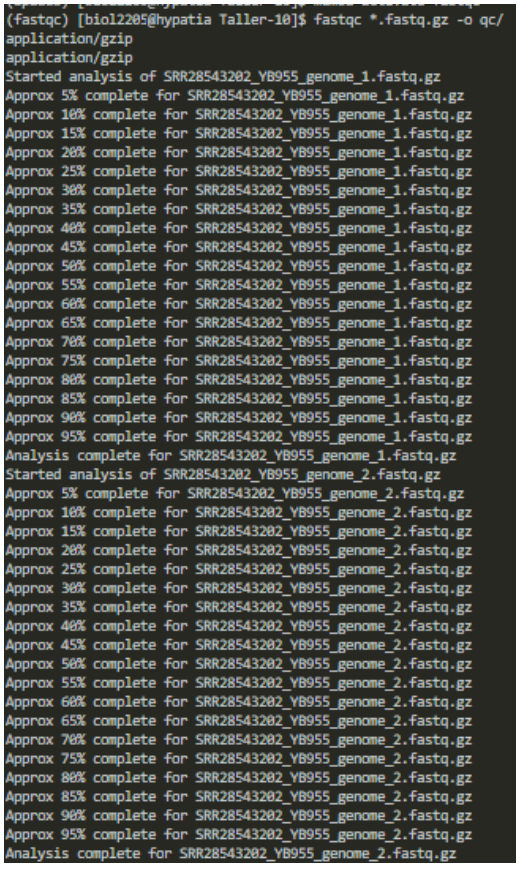
1. Revisando calidad de lecturas.

- Para revisar la calidad de las lecturas usadas en este taller se puede utilizar el programa FastQC. Primero asegúrese de copiar primero las lecturas desde la carpeta ~/Talleres/Taller-10 a su carpeta de grupo. Para esto último, se debe ejecutar el siguiente comando en la terminal:
    1. Activación del ambiente conda de fastqc:

        mamba activate fastqc

    2. Ejecución de FastQC:

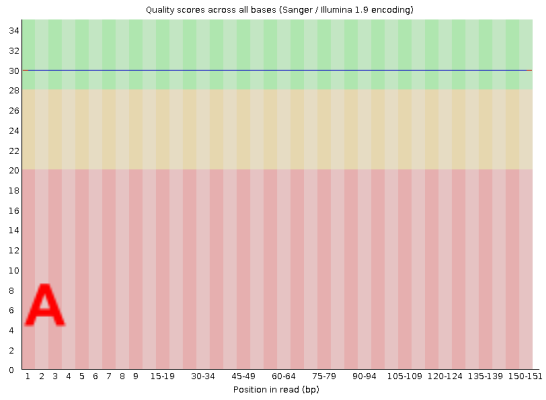
        mkdir qc
        fastqc *.fastq.gz -o qc/



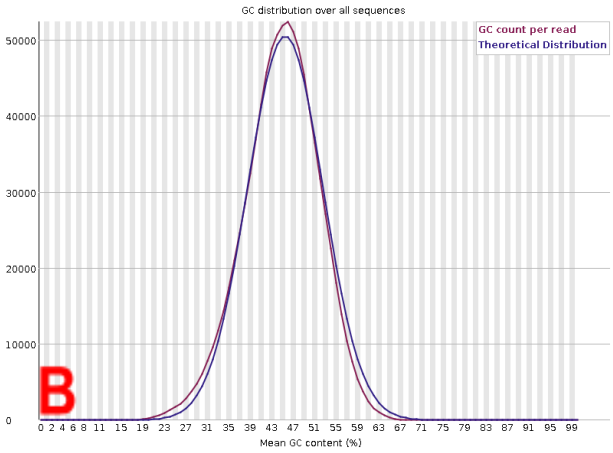
- Revise la salida de FastQC y determine si es necesario realizar algún tipo de filtrado o corrección de las lecturas.

Measure	Value
Filename	SRR28543202_YB955_genome_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	879071
Total Bases	129.8 Mbp
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	44

Per base sequence quality



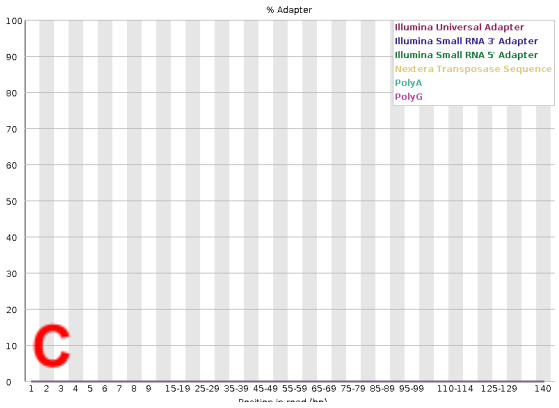
Per sequence GC content



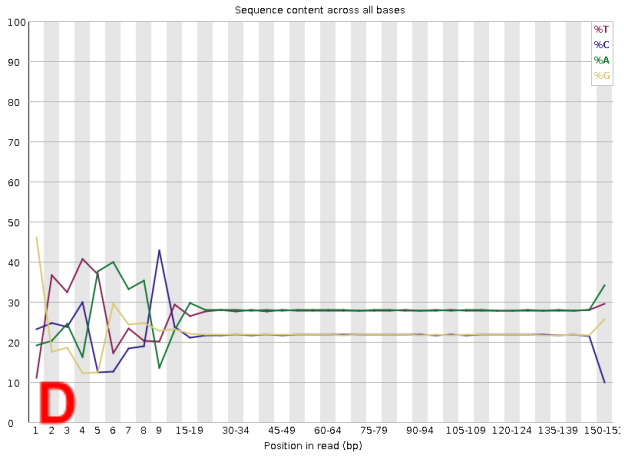
Overrepresented sequences

No overrepresented sequences

Adapter Content



Per base sequence content



Se observan en las ***Figura A, Figura B, Figura C y Figura D*** los resultados generados por el archivo html SRR28543202_YB955_genome_1_fastqc.html en el que pertenece al programa de https://www.bioinformatics.babraham.ac.uk/services.html En este se observa que tanto para los resultados del score usado, como para el adaptador y contenido CG son consistentens y correctos. No obstante una modificación o corrección sería el contenido de las primeras bases en el read, algo que se observa ***Figura D,*** en la cual las primeras posiciones de los read se encuentran ambiguos o confuso para su análisis, por lo que se sugiere hacer un filtrado en el comienzo de los reads para su posterior ensamblaje y evitar regiones inespecificas en las que no sea capaz su ensamblaje.

2. Ensamble de genomas.

- Para ensamblar los genomas se puede utilizar el programa SPAdes. Usando las lecturas de la carpeta tal como se tienen:
    1. Activación del ambiente conda de spades:

       mamba activate spades

    2. Ejecución de SPAdes:

       spades.py -t 8 --isolate -1 <forward> -2 <reverse> -o <outdir>

```
(spades) [biol2205@hypatia Taller-10]$ spades.py -t 8 --isolate -1 SRR28543202_YB955_genome_1.fastq.gz -2  SRR28543202_YB955_genome_2.fastq.gz -o Ensamblaje
```

3. Calidad del ensamble.

- Para evaluar la calidad del ensamble se puede utilizar el programa SeqFu. Usando los archivos de salida de SPAdes:
    1. Activación del ambiente conda de seqfu:

       mamba activate seqfu

    2. Ejecución de SeqFu:

       seqfu stats -n <contigs> <scaffolds>

File	#Seq	Total bp	Avg	N50	N75	N90	auN	Min	Max
contigs.fasta	141	4034382	28612.64	1016502	439016	241166	704422.11	78	1042555
scaffolds.fasta	139	4034555	29025.58	1016502	459922	306265	736309.77	78	1042555

- Analice las estadísticas obtenidas y describa la calidad del ensamble en términos de la longitud de los contigs y su N50.

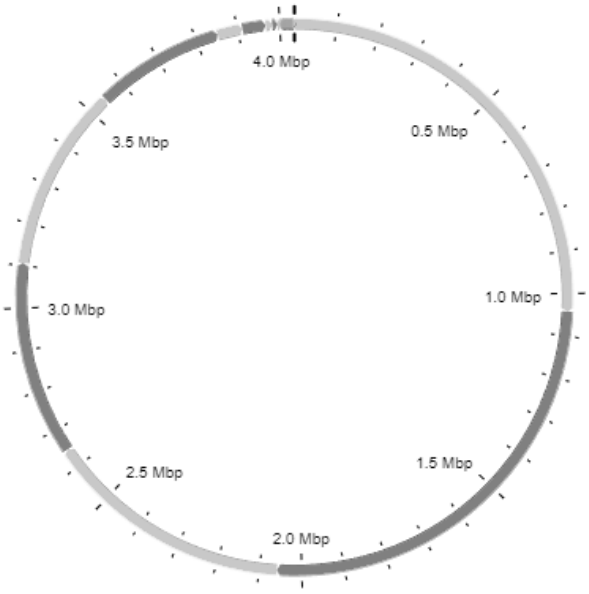
Se observa que las estadísticas se dividen en 2 grupos; uno para Contigs y otro para Scaffolds, Entre estos dos grupos resalta inicialmente la gran similitud entre sus resultados, ya que poseen el mismo mínimo, promedio, longitud en pb y una cantidad de secuencias muy cercanas. Esto podría ser causa de que el ensamblaje no se llevó como se esperaba, como argumenta Mandric et al. (2018), los scaffolds están constituidos por contigs adyacentes que fueron reconstruidos por regiones conservadas o sobrelapadas, sí el algoritmo no encuentra contigs adyacentes o algún otros contigs, no se formara y tomará la secuencia como un scaffolds. En nuestro caso, se formaron únicamente 2 scaffolds reduciendo el número de secuencias de 141 a 139. No obstante,

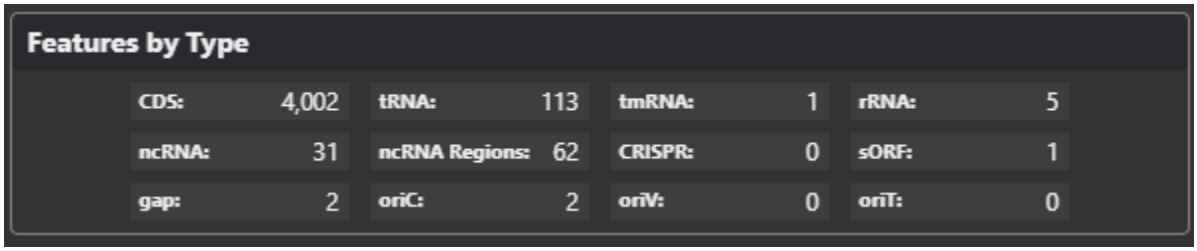
el hecho de que no se conformaron más scaffolds no representa que el ensamblaje sea un ensamblaje inexacto, debido a que como se señala en la imagen, la mayoría de las secuencias tienen una longitud superior a 1000000 pb y esto al considerar los reads iniciales, es un resultado acertado y específico en la reconstrucción del genoma

Por otra parte, el estadístico del N50 representa que el 50% de las secuencias reconstruidas tienen o superan esa longitud en pb. Asimismo, esto ocurre para N75 y N90, en el que cada estadístico toma un percentil de la longitud de la secuencia. Ahora bien, en nuestros resultados de ensamblaje se obtuvo un **N50** de **1016502pb** tanto para contigs como para Scaffolds, lo que significa que las secuencias tienen o son mayores a esa longitud de pares de bases. Al analizarlo con el promedio o "Avg" de los resultados, se determina que el ensamblaje es preciso, consistente y cumple con el objetivo principal de formar o reconstruir el genoma en base de las lecturas provistas, de manera que sean continúas los contigs, al tiempo, que se reduce los gaps entre las secuencias y disminuye el número de secuencias en el que partió. Ya que, al observar, el número promedio de los contigs es de **28612.64pb**, lo que es un resultado destacable tanto por su continuidad y tamaño notable en su secuencia. Igualmente, como se mencionó anteriormente si se examina el tamaño del **N50** este tiene un tamaño considerable de pb, puesto que al considerar que la secuencia tiene en total **4034382pb** y la mitad supera un ensamblaje de **1016502pb**, se concluye que el ensamblaje para los contigs es acertado y notorio en su capacidad para cubrir la mayoría del genoma. En este sentido, al observar los otros percentiles, estos tienen un tamaño de pb que no es despreciable y, en el caso de evaluar, finalmente, la calidad del ensamblaje, estos poseen una calidad aceptable y hasta favorable. Por lo que, Finalmente, se concluye que los contigs generados en el ensamblaje son resultados óptimos.

4. Visualizar y anotar el ensamble.

- Descargue los scaffolds resultados del ensamble y visualícelos con Proksee. Luego de cargar el genoma, busque la pestaña de herramientas (*tools*) en la interfaz y anote el genoma. Revise las estadísticas de anotación y descríbalas.





Por medio de la página Proksee, pudimos visualizar un genoma en específico, lo que nos permitió observar ciertas características importantes que nos ayudan a tener un mejor entendimiento de la secuencia.

Principalmente, encontramos 4,002 CDS, que es la región codificante, es decir, es la parte de un gen o ARN que contiene las instrucciones para la síntesis de una proteína. Para la secuencia utilizada, encontramos 31 moléculas de ncRNA en 62 regiones, la cual es una molécula de ADN que no se traduce en proteínas sino que realiza diferentes funciones dentro de la célula. Asimismo, hay presencia de 113 tRNA, que, a diferencia del ncRNA, esta se encarga de traducir información genérica del mRNA para la síntesis de proteínas, del que se encarga, específicamente en este caso, 5 moléculas de rRNA. Adicionalmente, se observa un tmRNA, que solo se presenta en bacterias, y es sistema de rescate para los ribosomas que se detienen durante la síntesis de proteínas debido a la presencia de codones de terminación prematuros en el ARNm.

Por otra parte, oriC se refiere al origen de replicación, por lo tanto, el hecho de que haya 2 oriC implica que hay mejor replicación, mayor estabilidad cromosómica y diversidad genética, aunque hay 2 regiones con bases ausentes (gap). Finalmente, se encontró solamente una sORF, lo que significa que solo hay un marco de lectura abierto corto, es decir, una secuencia de ADN o ARN que tiene las características de un gen, pero que es inusualmente corta.

Referencias

- Mandric, I., Knyazev, S. P., & Zelikovsky, A. (2018). Repeat-aware evaluation of scaffolding tools. *Bioinformatics*, *34*(15), 2530-2537. https://doi.org/10.1093/bioinformatics/bty131