Presentado por: Andrea Arredondo Restrepo, Sofía Betancur Álvarez, Mariana Gutiérrez Tamayo, María José Montoya Cuartas.
Grupo: Bioamigas

ENTREGA 1: Control de calidad, ensamblaje y mapeo

1. **Resumen QC:** Se realizo el análisis de control de calidad de las lecturas ancestrales y evolucionadas con *FastQC* y *MultiQC*, con el fin de evaluar su calidad antes y después de realizar el trimming con *fastp*. A continuación, se explicará los resultados obtenidos en las dos etapas, y los criterios considerados para realizar el trimming.

Análisis Pre-trimming (lecturas crudas):

Sample Name	% Dups	% GC	M Seqs
anc_R1	6.7 %	50 %	0.3 M
anc_R2	6.9 %	50 %	0.3 M
evol1_R1	11.7 %	50 %	1.0 M
evol1_R2	14.1 %	50 %	1.0 M
evol2_R1	11.4 %	51 %	0.9 M
evol2_R2	13.5 %	50 %	0.9 M

Figura 1. *MultiQC – General statistics.*

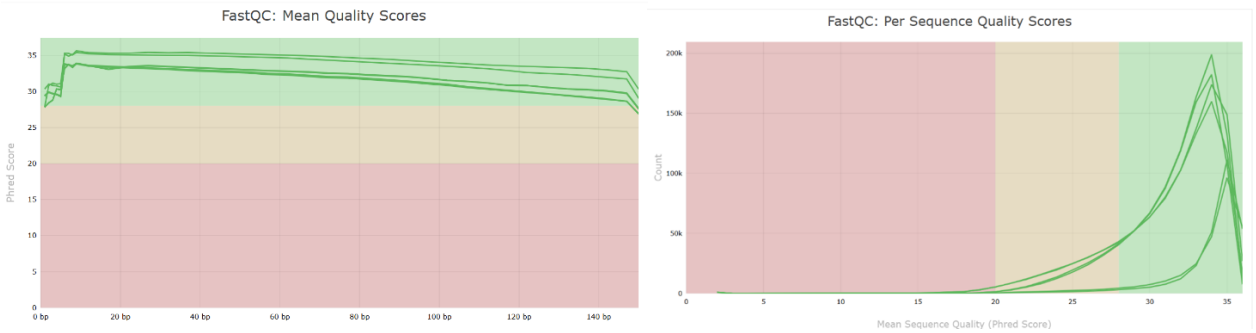


Figura 2 y Figura 3 (respectivamente). *MultiQC – Mean quality scores. MultiQC: Per sequence quality scores.*

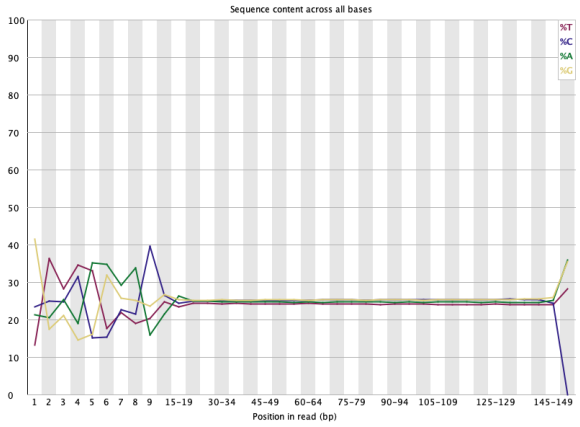


Figura 4. *FastQC – Per base sequence content del ancestro 1.*

ENTREGA 1: Control de calidad, ensamblaje y mapeo

Al analizar los archivos obtenidos por *FastQC* y *MultiQC* antes del trimming, se observa que todas las lecturas tienen una longitud de 35-150 pb, lo que significa que la secuenciación fue homogénea y limpia; sumado a esto, la calidad de las bases en su mayoría supera el valor Phred de 30 (Figura 2), teniendo una ligera caída en las primeras y últimas bases, que corresponde a un patrón típico de Illumina, pero a lo largo de la lectura la calidad se estabiliza, por lo que, es posible concluir que las lecturas son de alta calidad y confiables, teniendo un error aproximado del 0.1% (Illumina, s.f.).

Complementario a esto, en la Figura 3 se muestra la calidad promedio de cada lectura, confirmando una vez más que las 4 muestras tienen una calidad alta y son consistentes en la secuenciación; por su parte, las ancestrales tienen una fracción de lecturas de menor calidad más pequeña (Q20-25) que las evolucionadas. Sin embargo, según las métricas principales de cada lectura (Figura 1) las secuencias evolucionadas tienen una mayor profundidad que las ancestrales, lo que permite una mejor cobertura al tener mayor cantidad de lecturas.

A su vez, no se observan contaminaciones importantes ya que él %GC está entre el 50-51% lo que corresponde a lo esperado para *E. coli* según la literatura, que es de aproximadamente 50.79% (Sacristán et al., 2021). En cuanto a el porcentaje de duplicados, la Figura 1 muestra también que es mayor para los evolucionados, esto se debe a la profundidad, que como se explicó anteriormente es mayor que en los ancestrales. Los duplicados corresponden a las lecturas que son copias idénticas de la molécula de ADN que se dan por problemas en la PCR o la preparación de librerías, y pueden generar problemas si el porcentaje es mayor o igual al 50%, debido a que se sobreestima la cantidad de información útil de la secuenciación (Brennan, 2016). No obstante, el porcentaje de los datos analizados no supera el 13.5%, lo cual no corresponde a un problema y se puede trabajar con ellos perfectamente.

Por último, la Figura 4 muestra el porcentaje de cada base nitrogenada en las diferentes posiciones de la lectura a lo largo de la secuenciación (Babraham Bioinformatics, s.f.), como se puede observar luego de la posición aproximada de 15 pb la tendencia de las bases nitrogenadas se estabiliza casi en un alinea recta, lo que significa que la incorporación de los nucleótidos a la secuencia se de manera uniforme y corresponde a la composición real del ADN diana. Esto no ocurre en las posiciones 1-15 pb y 145-150 pb, en donde el porcentaje de cada base es variable. Este sesgo no se debe adaptadores (ya que no fueron detectados en los reportes), sino a errores técnicos propios de la plataforma Illumina.

Análisis del trimming:

Partiendo del análisis de la Figura 4, fue necesario hacer un trimming con el fin de corregir el error arrojado por el Per base sequence content, ya que se esperaba que esta distribución fuera uniforme y sin sesgos significativos; además, que el porcentaje de G-C y A-T fueran aproximadamente iguales. Como se observó los extremos de la secuenciación no cumplieron esta consigna; el mayor pronunciamiento se dio en el extremo 5´, lo cual puede indicar problemas con el cebador aleatorio de hexámeros utilizado en la preparación de las librerías,

lo que genera sesgos en la composición, debido a que su unión no es completamente aleatoria (Biostate AI, 2024) lo cual explica el ruido inicial en las primeras 15 bases.

Esto también es esperado ya que partiendo de lo descrito en la literatura se espera que las primeras 13 posiciones del extremo 5´ en las lecturas correspondan a este sesgo (Hansen et al., 2010). Por su parte, el sesgo que se observa en las últimas 5 bases del extremo 3´ (145-150 pb) corresponde a la degradación de la calidad de la secuencia con la progresión del ciclo (Biostate AI, 2024).

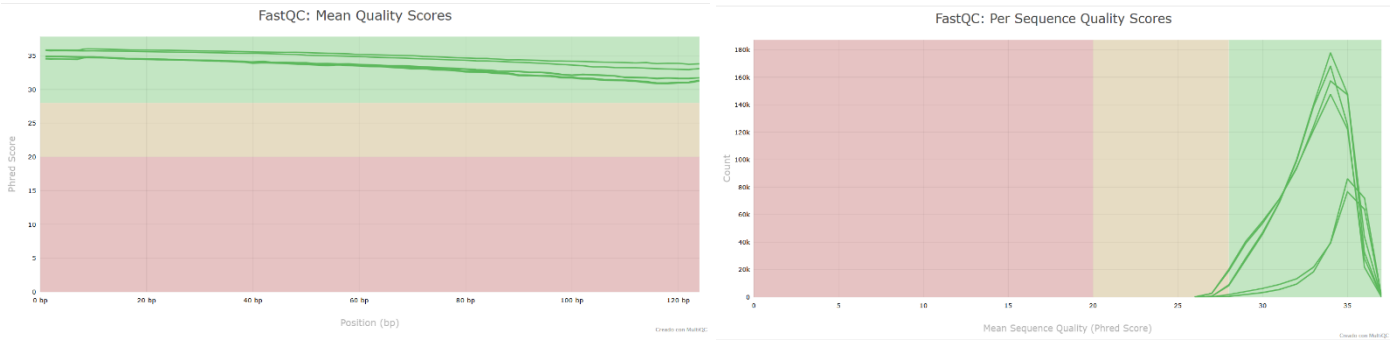
Es por esto, que se decidió realizar un trimming con el fin de reducir las regiones que generan el problema en la secuenciación y obteniendo extremos con una calidad más alta, Para esto, se cortaron las primeras 15 pb y las ultimas 5pb de cada lectura, estableciendo también una calidad de 28.

Análisis Post-trimming:

Luego de realizar el trimming, se generaron nuevamente los reportes *FastQC* y *MultiQC*.

Nombre de muestra	% Duplicación	M Lecturas después del filtrado	Contenido de GC	% PF	% Adaptador	% Dups	% GC	Secuencias M
anc_R1	0.6 %	0.5 METRO	51.0 %	84.3 %	27.8 %	6.2 %	51 %	0.2 METRO
anc_R2						6.4 %	50 %	0.2 METRO
evol1_R1	0.2 %	1.5 METRO	51.3 %	76.1 %	24.7 %	13.7 %	51 %	0.8 METRO
evol1_R2						14.5 %	51 %	0.8 METRO
evol2_R1	0.2 %	1.4 METRO	51.5 %	76.0 %	13.2 %	13.2 %	51 %	0.7 METRO
evol2_R2						13.8 %	51 %	0.7 METRO

Figura 5. *MultiQC – General statistics post trimming.*



Figuras 6 y 7 (respectivamente). *MultiQC – Mean quality scores post trimming. MultiQC: Per sequence quality scores post trimming.*

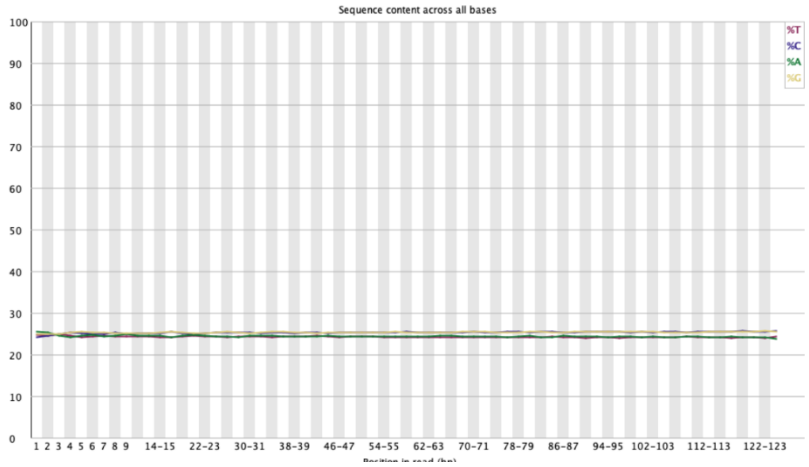


Figura 8. *FastQC – Per base sequence content del ancestro 1 post trimming.*

Como se pueden observar y al comparar los resultados post trimming con el pre trimming, la longitud de las lecturas disminuyó (50-120 bp) un efecto esperado tras el corte de las bases nitrogenadas que generaban el problema; también la calidad de las bases aumentó y ya no se ven caídas en los extremos (Figura 6), mantenido todas un valor Phred mayor a 30%. A su vez, como se ve en la Figura 6, ya no hay presencia de colas en la calidad de las secuencias, y siguen teniendo picos altos y bien definidos.

En cuanto a la Figura 1, las métricas también tuvieron cambios luego de realizar el corte, el %GC aumentó muy levemente y sigue manteniéndose en el rango esperado para *E. coli* lo cual era de esperarse puesto que con el trimming se eliminaron las bases nitrogenadas que no pertenecían al genoma real. Por su parte, el número de duplicados también creció un poco de 6.7 %– 13.5% (pre trimming) a 6.2% a 14.8%, debido a que, al realizar los cortes, se fragmenta más la secuencia y pueden coincidir lecturas entre sí, no obstante, es importante precisar que es porcentaje sigue siendo bajo y no interfiere negativamente en el proceso.

Al analizar el número de lecturas se encuentra que estas bajan con respecto a los datos crudos, lo cual es esperado y corresponde al corte del trimming; y en cuanto a la profundidad se puede ver que varió según las lecturas, en algunas aumentó con respecto al pre trimming y en otras disminuyó, esto se debe por que se eliminaron las secuencias con calidad menor y a su vez hay más lecturas confiables; también es fundamental precisar que la profundidad es más confiable porque proviene de lecturas tratadas con el control de calidad.

Para finalizar, la Figura 8 muestra una distribución uniforme de los porcentajes de las pares de bases a lo largo de la secuenciación, representando correctamente el ADN que se quiso secuenciar, y permitiendo concluir que el trimming fue efectivo y no solo logró mejorar métricas como la de la calidad, si no que corrigió el error generado por el *Per base sequence content* del pre trimming para todas las lecturas.

2. **¿Por qué se ensambló el genoma a partir de la línea ancestral y no de las líneas evolucionadas?**

Las lecturas ancestrales se ensamblarán de Novo y representaran el genoma de referencia, este tendrá las secuencias originales antes de la evolución de las bacterias, por lo que, se obtiene un genoma base que será la clave para identificar las posibles mutaciones, inserciones, deleciones, etc., que tendrán las lecturas evolucionadas. Si fuera al revés, se dificulta la identificación de los cambios adquiridos durante el experimento, porque quedan incorporados en el genoma de referencia.

3. **Resultados de *QUAST***

En este caso, para realizar el ensamblaje del genoma, se hizo uso de la herramienta *QUAST* (Quality Assesment Tool, por sus siglas en inglés), ya que es útil para comparar el ensamblaje de novo de diferentes genomas. Además, la evaluación de la calidad de un ensamblaje genómico se tiene en cuenta las 3C, conocidas como las tres dimensiones fundamentales: contigüidad, completitud y correctitud (Pac Bio, 2022).

Antes de iniciar con el análisis del ensamblaje, es necesario definir dos elementos claves dentro de la evaluación de la calidad por contigüidad, los contigs y los scaffolds. Los primeros, son secuencias contiguas de ADN que se reconstruyen por solapamiento y unión de las lecturas que comparten similitud de secuencia, mientras que los scaffolds son estructuras construidas para evitar las interrupciones en la lectura, las cuales están formadas por contigs enlazados en el orden y la orientación correcta (Muñoz et al., 2010). De este modo, la contigüidad es el grado en que él se producen secuencias largas y sin cortes en el ensamblaje y dependen tanto del tamaño de los contigs como de la longitud de los scaffolds y, las métricas de QUAST que se centran en la contigüidad son las de la rama Nx y Lx, al medir el grado de fragmentación del genoma (Salzberg et al., 2011)

Las métricas Nx (con $0 < x < 100$), cuantifican la longitud de los fragmentos de secuencia necesarios para cubrir un porcentaje dado de la longitud total del ensamblaje. Es decir, el N50 se define como la longitud mínima de contigs en pb, tal que la suma de las longitudes de todos los contigs de ese tamaño abarcan al menos el 50% de la longitud total del genoma ensamblado. Un valor alto implica que una mayor parte del genoma se encuentra en fragmentos más grandes, lo cual es altamente deseable para evitar discontinuidad. Análogamente, el N90 es la longitud de contigs necesaria para cubrir el 90% de la longitud total del ensamblaje. Esta última métrica, cubre la mayor parte del genoma, entonces es altamente sensible a la fragmentación de los contigs medianos y grandes, así, si se tiene un valor bajo, sugiere que, aunque hay algunos contigs muy grandes, el resto se encuentra muy fragmentado (Gurevich et al., 2013).

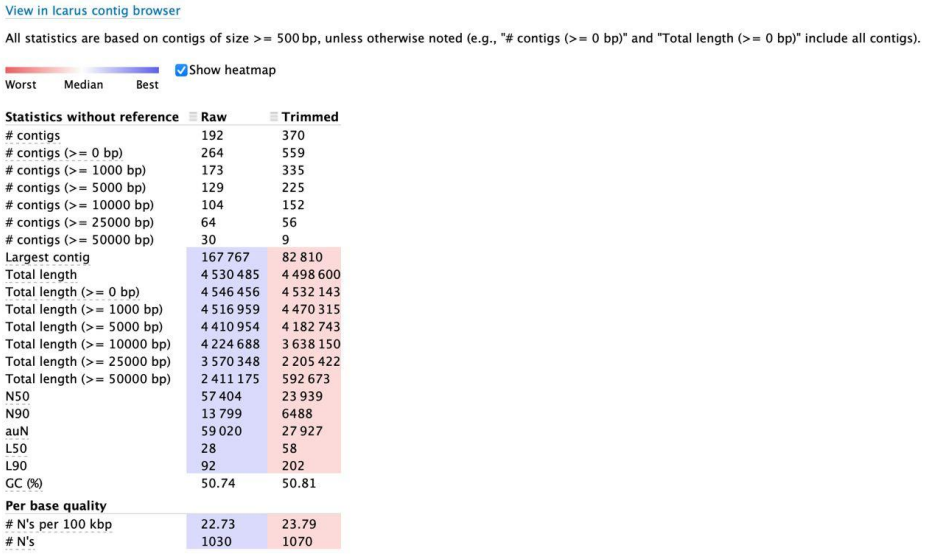
Por otra parte, las métricas Lx (con $0 < x < 100$), complementan las métricas Nx al proporcionar una medida del número de fragmentos que componen la estructura del ensamblaje. El L50 es el número mínimo de contigs que, al ordenarse de mayor a menor, suman al menos el 50% de la longitud total del ensamblaje. Un valor L50 bajo es preferible, ya que indica que se requiere una cantidad reducida de fragmentos largos para representar la mitad del genoma. De la misma manera, el L90 es la cantidad necesaria para cubrir el 90% del genoma e igualmente un valor bajo es indicativo de un ensamblaje robusto y poco fragmentado (Wang & Wang, 2023). Además, una de las métricas de QUAST más recientes, es la llamada auN, que es usada para capturar lo tendencia general de la contigüidad en un solo valor, debido a que considera toda la curva y no solamente un punto especifico x (Li, 2020). Es decir, el auN es el área bajo la curva Nx, y es más estable y se ve menos afectado por grandes saltos en longitudes de los contigs. De esta manera, un auN más alto indica una mejor contigüidad del ensamblaje.

Teniendo en cuenta lo anterior, mejores métricas no significan un mejor ensamblaje, porque, aunque N50 o un auN más altos representan una buena contigüidad, no son suficientes para declarar un ensamblaje mejor que otro, debido a que la calidad de un ensamblaje no depende solo de esta dimensión y se deben abarcar también los parámetros que representan las otras dos dimensiones. Por lo tanto, es fundamental interpretar las métricas de contigüidad conjuntamente con métricas que evalúan la fidelidad de la secuencia, es decir, la correctitud y completitud.

Así entonces, dentro del reporte QUAST, la métrica indicada para evaluar la Correctitud a nivel de base y la ambigüedad es el # N's per-100 kbp (QUAST 5.3.0 manual, s.f), la cual es una estadística que mide el numero promedio de bases ambiguas o sin llamar, representadas por ’N’, que generalmente indican regiones donde el ensamblador no pudo determinar un consenso de secuencia confiable debido a la falta de cobertura o solapamiento ambiguo. En este caso entonces, lo que se busca es un valor más bajo porque sugiere que el ensamblaje tiene alta confianza a nivel de bases, y un valor más alto podría indicar que al realizar el proceso de trimming se perdió cobertura de lectura Gurevich et al., 2013).

Relacionando lo anterior con el reporte QUAST obtenido en el desarrollo del proyecto, se presenta la Figura 9, que contiene las métricas arrojadas por la herramienta. En cuanto a las métricas relacionadas con la contigüidad se tiene que, el ensamblaje Trimmed resultó en casi el doble de fragmentos necesarios (370 contigs vs. 192 en Raw) para cubrir una longitud genómica esencialmente idéntica (aproximadamente 4.5 Mbp), mostrado así que los parámetros usados para realizar el trimming, fragmentaron el genoma. Además, N50 y auN del Raw superan en más del doble a los del Trimmed (57,404 vs. 23,939 para N50; 59,019.9 vs. 27,927.4 para auN). Esto se refuerza con la longitud del contig más grande, donde Raw produce un fragmento de 167,767 pb, más del doble que el contig más grande de Trimmed (82,810 pb). Asimismo, los valores L50 (28) y L90 (92) de Raw son marcadamente inferiores a los de Trimmed (58 y 202, respectivamente). Esto significa que Raw logró

ensamblar el 90% del genoma utilizando menos de la mitad de los contigs necesarios por Trimmed, demostrando unas mejores métricas de contigüidad. A demás, el ensamblaje Raw logró producir 30 contigs de longitud ≥50,000 pb, mientras que el ensamblaje Trimmed solo retuvo 9 contigs en este rango, esta alta diferencia demuestra que el proceso de depuración de calidad llevó a la ruptura de las regiones más largas y posiblemente críticas del genoma, lo que explica el valor mucho más bajo en el valor N50 del Trimmed.



View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	Raw	Trimmed
# contigs	192	370
# contigs (>= 0 bp)	264	559
# contigs (>= 1000 bp)	173	335
# contigs (>= 5000 bp)	129	225
# contigs (>= 10000 bp)	104	152
# contigs (>= 25000 bp)	64	56
# contigs (>= 50000 bp)	30	9
Largest contig	167 767	82 810
Total length	4 530 485	4 498 600
Total length (>= 0 bp)	4 546 456	4 532 143
Total length (>= 1000 bp)	4 516 959	4 470 315
Total length (>= 5000 bp)	4 410 954	4 182 743
Total length (>= 10000 bp)	4 224 688	3 638 150
Total length (>= 25000 bp)	3 570 348	2 205 422
Total length (>= 50000 bp)	2 411 175	592 673
N50	57 404	23 939
N90	13 799	6488
auN	59 020	27 927
L50	28	58
L90	92	202
GC (%)	50.74	50.81
Per base quality		
# N's per 100 kbp	22.73	23.79
# N's	1030	1070

Figura 9. *Reporte QUAST – Comparación de métricas para los dos genomas ensamblados (Raw y Trimmed).*

Como se mencionó antes, otro aspecto clave de la evaluación es si la pérdida de contigüidad fue compensada por una ganancia en correctitud. Sin embargo, el ensamblaje Raw muestra un promedio de 22.73 N's por 100 kbp, mientras que el ensamblaje Trimmed muestra 23.79 N's por 100 kbp. Este resultado muestra que, al acortar las lecturas, se debilitó la evidencia de consenso en ciertas regiones, obligando al ensamblador a introducir más 'N's para mantener la integridad del consenso local.

En cuanto a la calidad del ensamblaje en el Trimmed, la disminución de contigüidad no fue compensada por una mejora en la correctitud. Ya con estas dos dimensiones en desventaja en comparación al raw data, es posible concluir que el trimming realizado, si bien buscaba eliminar los errores de secuenciación encontrados en el reporte *multi-qc* para los primeros ciclos de Illumina, fue muy agresivo y afecto las métricas del ensamblaje. Entonces, contrario a la expectativa de que el trimming puede mejorar la calidad del ensamblaje, en este caso, la depuración agresiva de las primeras lecturas pudo eliminar información crítica de solapamiento, afectando así el ensamblaje de Novo del genoma.

Respecto a la información sobre la cobertura y profundidad, en las Figura 10 se presentan los histogramas arrojados por el reporte QUAST, debido a que estas métricas también son importantes para determinar la calidad de un ensamblaje de Novo. El histograma de cobertura muestra entonces la longitud total del ensamblaje (eje Y, en kbp) que alcanzó una determinada

profundidad de lectura (eje X, en X veces). En el ensamblaje de Novo a partir de lecturas cortas como Illumina, la baja profundidad implica que el ensamblador tiene menos lecturas para construir cada región, por eso, la mayor parte del genoma ensamblado (la "longitud total" en el eje Y) está concentrada en un rango bajo entre 5x y 7x. La similitud de las líneas de tendencia, indican que el proceso de trimming no disminuyó, ni aumentó la profundidad de secuenciación de las regiones que lograron ensamblarse.

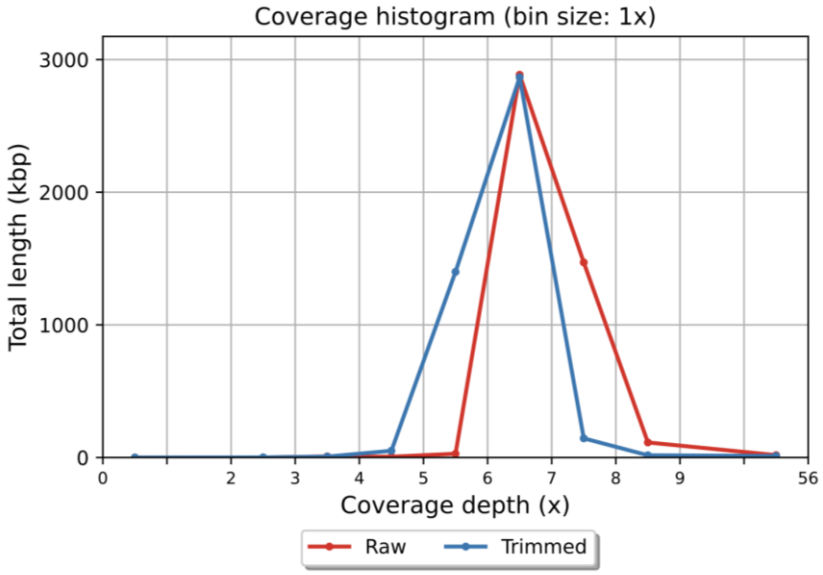


Figura 10. *Histograma de cobertura de profundidad para los datos crudos (Raw) y los datos depurados (Trimmed).*

El ensamblaje Raw presentó métricas de contigüidad superiores al ensamblaje Trimmed, evidenciado por valores de N50, auN, longitud del contig más grande, número total de contigs y valores de L50 y L90 más favorables. Esto demuestra que el genoma pudo ser reconstruido en fragmentos más largos y con menor fragmentación. En cuanto a la correctitud, el número promedio de bases ambiguas (N's por 100 kbp) fue ligeramente menor en Raw (22.73) que en Trimmed (23.79), lo que indica que el trimming no mejoró la calidad y, por el contrario, redujo la confianza en el ensamblaje. La completitud de ambos ensamblajes fue muy similar, con longitudes totales aproximadas a 4.5 Mb y porcentajes de GC muy similares.

Finalmente, el trimming aplicado resultó ser demasiado agresivo, ya que incrementó la fragmentación y redujo la contigüidad, afectando el resultado final arrojado por *QUAST*. Pero, a pesar de que el ensamblaje usando los datos crudos obtuvo mejores métricas; se optó por utilizar el ensamblaje con las secuencias del Trimmed para la etapa de mapeo. Esta decisión se basa en que el trimming eliminó las bases de baja calidad y las lecturas con errores de secuenciación propios de Illumina, lo cual pese a que se reduce la continuidad, aporta mayor confianza en la correctitud de las secuencias individuales.

Es por esto que, aunque los contigs del ensamblaje Trimmed estén más fragmentados, las bases de estos vienen de secuencias con mejor calidad, lo cual reduce la probabilidad de errores de consenso en las posiciones críticas del mapeo. Entonces, es preferible optar por contigs de menor longitud, puesto que el ensamblaje con el trimming seguirá siendo representativo del genoma y tendrá menos sesgos de calidad

4. ¿Qué significa y por qué se debe indexar el genoma?

Indexar el genoma es fundamental antes de realizar el alineamiento y mapeo de las lecturas, en este se prepara el archivo de secuencia de referencia, usualmente en formato FASTA, para que pueda ser utilizado por algoritmos de alineamiento de lecturas de manera eficiente. En este proceso, el genoma no se modifica, sino que se generan archivos adicionales que representan su contenido en un formato optimizado para búsquedas rápidas. De esta manera, con BWA, en lugar de comparar cada lectura directamente contra la secuencia completa del genoma, lo que sería un proceso extremadamente lento, la indexación organiza la información en un formato que permite búsquedas rápidas y precisa (Li & Durbin, 2009). En este caso, se indexó el archivo *scaffolds.fasta*, correspondiente a los scaffolds del ensamblaje del genoma del ancestro con los reads Trimmed, con el fin de utilizarlo como genoma de referencia en el mapeo de las lecturas evolucionadas.

Durante la indexación con BWA se generan archivos auxiliares (.amb, .ann, .bwt, .pac, .sa), los cuales corresponden a estructuras de datos derivadas del algoritmo Burrows–Wheeler Transform (BWT) que permiten que se realicen búsquedas rápidas y eficientes dentro de la secuencia para el alineamiento. Estos archivos permiten comprimir el genoma, manejar regiones ambiguas y, sobre todo, realizar búsquedas rápidas de subsecuencias durante el alineamiento. Sin ellos, el programa tendría que comparar cada lectura contra la secuencia completa de manera lineal, lo que sería computacionalmente inviable. Dichos archivos no son resultados finales, sino productos intermedios necesarios para que el mapeo de lecturas se realice, reduciendo el tiempo de procesamiento y el uso de memoria.

Es por esto que, la indexación del genoma no solo es importante desde el punto de vista computacional, sino también práctico, ya que permite que los análisis de mapeo sean realizables en un tiempo razonable. Sin un índice, el alineamiento de lecturas cortas o largas contra genomas grandes, como en este caso, un genoma bacteriano, sería prácticamente imposible en términos de tiempo y recursos. Por esta razón, indexar el genoma es un paso fundamental en todo flujo de análisis de datos de secuenciación.

5. Si quiero ver en IGV el resultado de mi mapeo, ¿qué significa y por qué debo indexar el mapeo?

Indexar un archivo de mapeo en formato BAM consiste en generar un archivo auxiliar *.bai* mediante herramientas como *samtools*. Este índice funciona como una tabla de accesos directos que indica en qué posición del archivo BAM se encuentran las lecturas alineadas a cada región del genoma. La indexación es necesaria porque los archivos BAM pueden ser muy grandes, y sin el índice los programas tendrían que recorrer todo el archivo para acceder a un segmento específico, lo que haría el análisis y la visualización inviable. Con el archivo *.bai*, IGV puede cargar rápidamente solo las secciones del genoma que se desean visualizar (Li et al., 2009).

IGV requiere acceder de forma interactiva a segmentos particulares del genoma, por ende, al cargar el archivo BAM, busca el archivo *.bai* en la misma carpeta para mostrar rápidamente los alineamientos, la cobertura y las posibles variantes en una región determinada, observando los resultados del mapeo. Para visualizar de manera general el mapeo, se cargó como genoma de referencia el ensamblaje obtenido a partir de las lecturas trimmed del ancestro (*scaffolds.fasta*), junto con los archivos de mapeo resultantes (*evol1.filtered.bam* y *evol2.filtered.bam*), teniendo en cuenta que los archivos indexados se encuentren en la misma carpeta de mapeo. En la Figura 11, se muestra la visualización de las líneas evolucionadas mapeadas respecto al genoma de referencia, en la posición 42.173 a 42.206 bp del nodo 1.

Figura 11. *Visualización en IGV del mapeo respecto al genoma de referencia, mostrando todos los nucleótidos de cada lectura.*

Figura 12. *Visualización en IGV del mapeo respecto al genoma de referencia.*

En el panel superior se observa el histograma de profundidad de cobertura, el cual representa el número total de lecturas de secuenciación que se han alineado en cada posición del genoma. Debajo, se visualiza la pila de lecturas mapeadas para cada línea evolucionada con respecto al genoma de referencia. Cada línea horizontal corresponde a una única lectura, en la cual se comparan las bases obtenidas experimentalmente con las del genoma ancestral. En esta sección del genoma, en general, las bases coinciden con la referencia, debido a que se muestran en color gris (Figura 12), no obstante, también aparecen nucleótidos coloreados, los cuales indican la presencia de mismatches.

El color e intensidad diferente del nucleótido indican que existe una diferencia entre las bases observadas en las lecturas y las bases del genoma de referencia, representando mutaciones reales, posibles errores de secuenciación o la lectura de la hebra complementaria. La intensidad del color refleja la calidad del nucleótido en la lectura: colores más intensos y definidos corresponden a bases con alta calidad, mientras que colores más claros o traslucidos indican baja confianza en esa posición del nucleótido. Por ello, tanto la consistencia de un mismatch en múltiples lecturas, como la calidad de las bases, son criterios para identificar si se trata de un error o de una variante (Thorvaldsdóttir et al., 2013).

En la Figura 11, en la región señalada (center line) se observa un cambio de nucleótido en una de las lecturas (C en vez de A), el cual posee una coloración traslucida, debiéndose posiblemente a un error de secuenciación; además, a lo largo de este segmento genómico se identifican otros mismatches con diferente intensidad de color, lo que sugiere, también la presencia de posibles polimorfismos de un solo nucleótido (SNPs) en las líneas evolucionadas. Finalmente, en la parte interior del panel se observa la secuencia de referencia y la traducción de esta a aminoácidos, con el respectivo codón de parada (rojo) y de inicio (verde).

6. Interpretación *Qualimap*:

Luego del proceso de calidad, donde se depuraron los datos crudos para el organismo evolucionado 1, se obtuvo que el número de reads tras el filtrado de calidad fue 1,505,688. Sin embargo, después del mapeo contra el genoma ancestral y la generación del archivo *filtered.bam*, el número final de reads considerados para el mapeo fue de 1,468,073. Esta diferencia se debe al filtrado que se aplica sobre el archivo BAM, ya que durante el alineamiento se descartan lecturas que no se logran mapear o lecturas de baja calidad, por lo que el archivo *filtered.bam* contiene únicamente alineamientos más robustos.

En ese orden de ideas, según el reporte de calidad, el porcentaje obtenido de reads mapeados fue del 100 %, de los cuales 1,467,946 fueron pares mapeados juntos. Esto indica que ambos extremos (R1 y R2) lograron alinearse correctamente y que, por ende, biológicamente el fragmento de ADN está bien representado en el genoma de referencia. Los 127 reads restantes corresponden a *singletons*, lo que indica la presencia de lecturas en las que uno de los extremos logró alinearse correctamente y el otro no. Esto puede suceder ya sea porque el extremo que no se pudo alinear no tenía una calidad muy buena o debido a que el corte del *trimming* fue agresivo, provocando que el extremo quedara demasiado corto. Por todo lo anterior, se puede decir que el ensamblaje del genoma ancestral es representativo respecto al genoma evolucionado, ya que casi no hay secuencias que queden sin mapear.

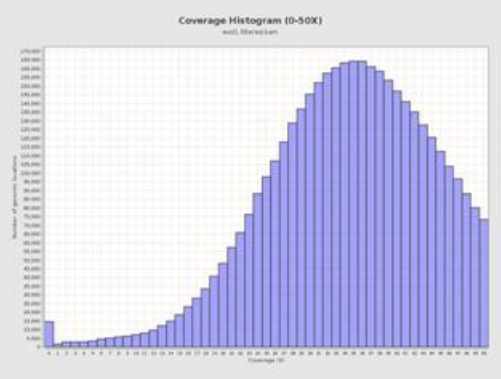


Figura 13. *Histograma de cobertura 0-50X organismo evolucionado 1.*

En cuanto a la cobertura, es importante tener en cuenta que el tamaño del genoma de referencia es de 4,53 Mb y las bases mapeadas fueron 170 Mb, lo que permitió obtener una cobertura promedio de 37,5758X. Esto indica que cada base fue leída aproximadamente 37 veces, reduciendo así el riesgo de errores y mejorando la calidad del ensamblaje. Asimismo, la cobertura también se puede interpretar a partir de la Figura 13, donde el eje X representa la profundidad de cobertura y el eje Y el número de bases con esa cobertura. Como se puede observar, a la izquierda del gráfico hay pocas bases con una cobertura entre 0–5X y, a partir de ahí, se evidencia un aumento progresivo que indica un incremento en la cantidad de bases con cobertura media. Luego, se alcanza el pico central, donde la cobertura es aproximadamente de 35–40X, lo cual indica que la mayoría de las bases del genoma fueron leídas en promedio entre 35 y 40 veces. Posteriormente, la gráfica comienza a descender, mostrando que hay muy pocas regiones con una cobertura tan alta.

De esta manera, el patrón en el histograma forma una campana centrada, indicando así que las lecturas se distribuyeron de forma uniforme y que la cobertura media es confiable para procedimientos que se quieran llevar a cabo a futuro, como el análisis de variantes. Por otra parte, se obtuvo que la calidad de mapeo fue de 55,72, lo cual es bastante bueno, puesto que este parámetro indica la confianza de que una lectura fue correctamente alineada a una posición específica del genoma, y valores altos (>30) reflejan una alta calidad en el alineamiento. Este mismo análisis se realizó para el organismo evolucionado 2, donde se debe tener en cuenta que, después de la depuración de los datos crudos, se obtuvieron 1,408,676 reads, pero luego del mapeo, el número de reads fue de 1,347,762. Esta diferencia puede deberse a las mismas razones explicadas anteriormente para el organismo evolucionado 1. De acuerdo con los resultados de *Qualimap,* se mapearon el 100 % de los reads, de los cuales 1,347,679 fueron mapeados juntos y 83 correspondieron a *singletons*, indicando que hay pocas secuencias sin mapear y que la mayoría de los reads encontraron su lugar en el genoma ancestral.

En cuanto a la cobertura, la cantidad de bases mapeadas fue de 161 Mb, lo que permitió obtener una cobertura media de 35,5426X, indicando así que cada base fue leída en promedio 35 veces.

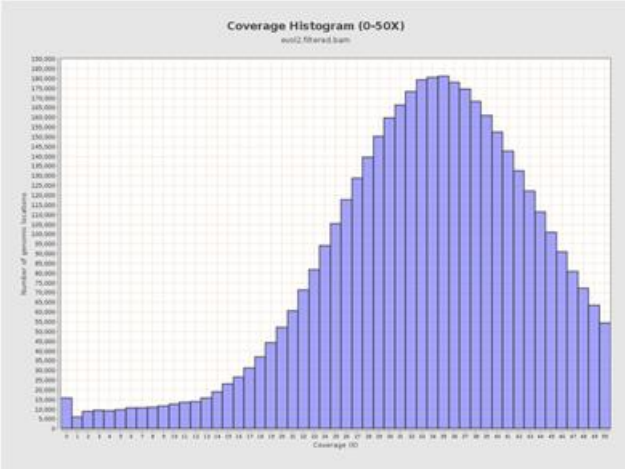


Figura 14. *Histograma de cobertura 0-50X organismo evolucionado 2.*

Como se mencionó anteriormente, la cobertura también se puede analizar mediante el histograma de la Figura 14, donde se observa que la mayor parte del genoma tiene una cobertura entre 30–40X, con un pico pronunciado en 35X, lo cual coincide con el dato reportado en el informe de *Qualimap*. Además, la curva es simétrica y está distribuida uniformemente a lo largo del genoma, lo que puede indicar que no hay regiones sobrerrepresentadas.

Finalmente, si comparamos ambos organismos evolutivos, se puede decir que ambos presentan coberturas muy similares y distribuciones uniformes a lo largo del genoma, indicando que, tras aplicar los filtros de calidad, los reads que permanecen en los archivos *filtered.bam* representan de forma robusta el genoma de cada organismo evolucionado. La similitud en cobertura, la distribución uniforme a lo largo del genoma y la alta calidad de

mapeo en ambos casos sugiere que el genoma ensamblado del organismo ancestral es una buena referencia y que las diferencias que se pueden detectar entre los organismos evolucionados pueden corresponder a variaciones biológicas.

REFERENCIAS.

Babraham bioinformatics. (s. f.). *Per base sequence content. En FastQC Help — Analysis Modules*. Babraham bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html

Beyond contiguity - assessing the quality of Genome Assemblies with the 3 CS. PacBio. (2022, April 11). https://www.pacb.com/blog/beyond-contiguity/

Bioestate AI. (s.f.). *Assessing FastQC Results for Per Base Sequence Content in RNA-Seq*. Bioestate AI. https://biostate.ai/blogs/assessing-fastqc-per-base-sequence-content-rna-seq/

Brennan, R. (2016). *The trouble with PCR duplicates.* The molecular ecologist. https://www.molecularecologist.com/2016/08/25/the-trouble-with-pcr-duplicates/

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). Quast: Quality Assessment Tool for Genome Assemblies. Bioinformatics, 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hansen, K., Brenner, S. & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research, 38*(12), 2-7. https://doi.org/10.1093/nar/gkq224

Illumina. (s.f.). *What is a Quality Score in Sequencing?*. Illumina. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html#:~:text=Cuando%20la%20calidad%20de%20la,de%20nueva%20generaci%C3%B3n%20(NGS)

Li, H. (2020). auN: A new metric to measure assembly contiguity. Sitewide ATOM. https://lh3.github.io/2020/04/08/aun-a-new-metric-to-measure-assembly-contiguity

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Muñoz, A., Zheng, C., Zhu, Q. et al. (2010). Scaffold filling, contig fusion and comparative gene order inference. BMC Bioinformatics 11, 304 https://doi.org/10.1186/1471-2105-11-304

QUAST 5.3.0 manual. Quast 5.3.0 Manual. (n.d.). https://quast.sourceforge.net/docs/manual.html#sec3

Sacristán, E., Gonzales, S., Peiró, R., Carrasco, F., Amils, R., Requena, J., Berenger, J. & Aguado, B. (2021). ARAMIS: From systematic errors of NGS long reads to accurate assemblies. *Briefings in Bioinformatic, 22*(6), 1-14. https://doi.org/10.1093/bib/bbab170

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marçais, G., Pop, M., & Yorke, J. A. (2011). Gage: A critical evaluation of genome assemblies and assembly algorithms. Genome Research, 22(3), 557–567. https://doi.org/10.1101/gr.131383.111

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Wang, P., & Wang, F. (2023). A proposed metric set for evaluation of Genome Assembly Quality. Trends in Genetics, 39(3), 175–186. https://doi.org/10.1016/j.tig.2022.10.005