Informe de Ensamblaje y Análisis de Lecturas Genómicas Grupo 2

Santiago Berty Muñoz
Tomás Gonzaléz Franco
Sofía Cabezas Cruz
Escuela de Ingenieros de Antioquia
Pregrado en Ingeniería Biotecnológica

1 de octubre de 2025

1. Introducción

Se evaluó la calidad de lecturas (ANC, EVOL1, EVOL2) utilizando `FastQC`/`MultiQC`, se aplicó `fastp` para el trimming, se realizaron ensamblajes *de novo* (RAW vs. TRIMMED) con `SPAdes`, se compararon con `QUAST` y finalmente se mapearon las lecturas evolutivas contra la referencia seleccionada.

2. Control de Calidad (QC)

QC pre-trimming

Los reportes de `FastQC`/`MultiQC` sobre datos crudos mostraron en general una calidad *Phred* alta (valores >28), con una caída típica hacia el extremo 3′, presencia de adaptadores y una distribución de longitudes heterogénea. El contenido de GC se mentuvo cercano al 50 %.

Parámetros de trimming y justificación

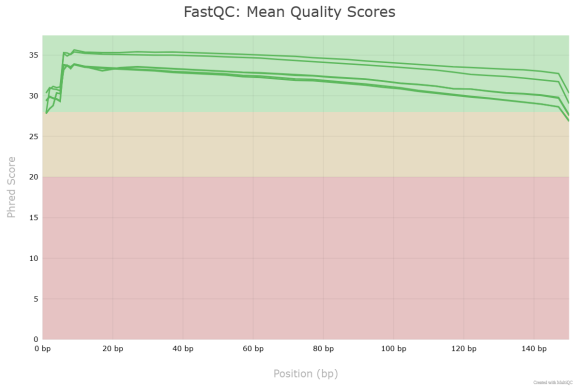
Se empleó `fastp` con:

- `-detect_adapter_for_pe`: Para la detección y eliminación automática de adaptadores en pair ends.
- `-qualified_quality_phred 25` y `-cut_mean_quality 25`: Para conservar únicamente bases de alta fidelidad.
- `-length_required 50`: Para evitar fragmentos muy cortos.
- `-cut_front` y `-cut_tail`: Corte adaptativo en extremos.
- `-correction`: Para hacer corrección por solapamiento en pair ends.

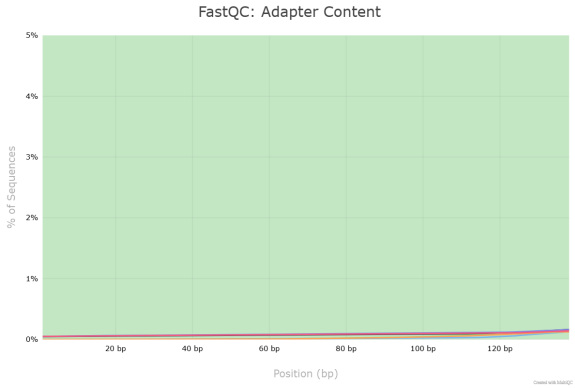
Este conjunto de parámetros nos permite equilibrar la limpieza (menos adaptadores y regiones de baja calidad) con la retención de info suficiente para ensamblar y mapear.

QC post-trimming

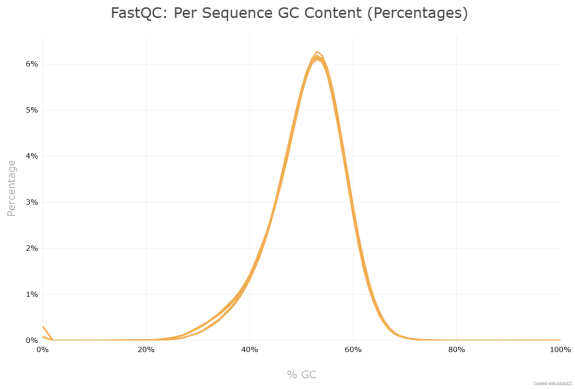
Tras el trimming, los reportes de MultiQC mostraron mejoras claras: En síntesis la calidad por base se mantuvo estable a lo largo de toda la longitud de las lecturas, los adaptadores prácticamente desaparecieron casi por completo y la distribución de longitudes mostró mayor homogeneidad. (Ver figura 2)



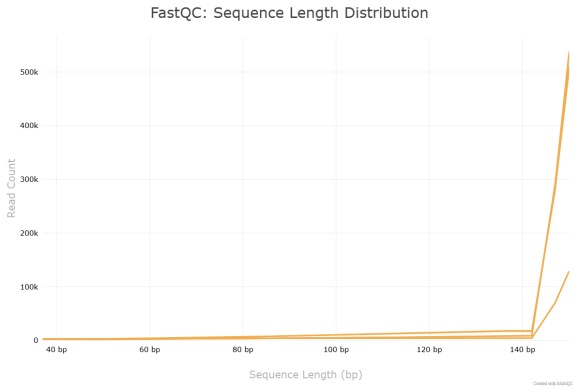
(a) Pre: per-base sequence quality



(b) Pre: adapter content



(c) Pre: per-sequence GC content



(d) Pre: sequence length distribution

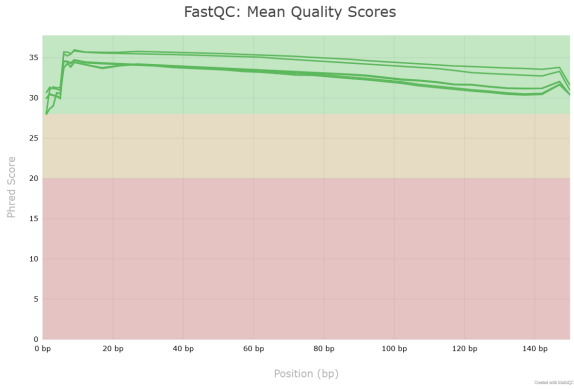
Figura 1: Resumen QC **pre-trimming** exportado de MultiQC.

3. Ensamblaje y Evaluación

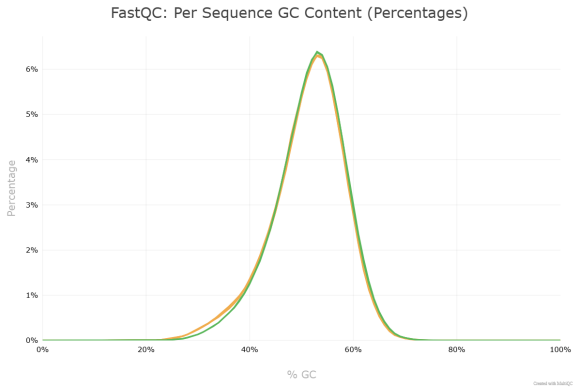
El ensamblaje del genoma se llevó a cabo usando `SPAdes`, utilizando las lecturas de la línea ancestral (ANC). La elección de esta línea se justifica porque proporciona una referencia común y estable sobre la cual se pueden contrastar las líneas evolucionadas. De esta forma se evita el sesgo que implicaría ensamblar cada evolución de manera independiente, garantizando consistencia en las comparaciones de más adelante.

Resultados de QUAST

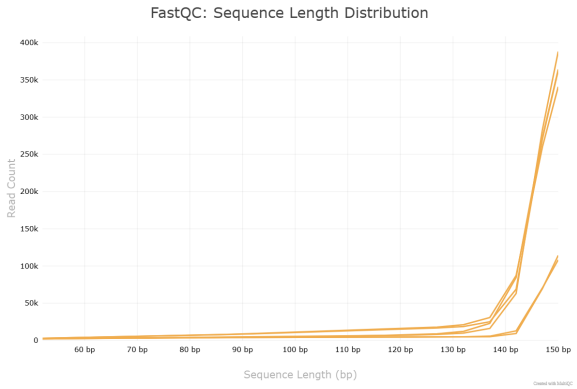
Los ensamblajes obtenidos a partir de reads crudos (RAW) y reads depurados (TRIM) se compararon con `QUAST`. Se pueden ver los resultados en el cuadro 1.



(a) Post: per-base sequence quality



(b) Post: per-sequence GC content



(c) Post: sequence length distribution

Figura 2: Resumen QC **post-trimming** exportado de MultiQC.

Métrica	RAW	TRIM
# Contigs ($\geq$ 0 bp)	262	264
# Contigs ($\geq$ 1000 bp)	173	182
# Contigs ($\geq$ 5000 bp)	129	135
# Contigs ($\geq$ 10000 bp)	104	106
N50	57404	50016
N90	13799	13213
L50	28	29
L90	92	97
auN	59020.0	55835.8
Largest contig	167767	136463
GC (%)	50.74	50.74

Cuadro 1: Resultados de QUAST para los ensamblajes RAW y TRIM.

Discusión de métricas

Los resultados de QUAST muestran que el ensamblaje RAW presenta una mayor continuidad en comparación con el de TRIM. Este resultado se refleja en un valor superior de N50 (57,404 vs 50,016), un contig máximo más largo (167,767 vs 136,463) y un L50 más bajo (28 vs 29). En general, estas métricas nos indican que el ensamblaje RAW está menos fragmentado y ofrece una representación más estable del genoma.

Si bien el ensamblaje TRIM se construyó a partir de lecturas depuradas (reduciendo el riesgo de errores que tengan que ver con adaptadores o bases de baja calidad), la pérdida de continuidad

que se observa limita su utilidad como referencia. En este caso, las métricas de continuidad son más determinantes, ya que lo que se busca es contar con un ensamblaje ancestral lo más completo y representativo posible para poder mapear las líneas evolucionadas.

Por estas razones, el ensamblaje RAW fue seleccionado como referencia para los análisis posteriores, priorizando la integridad y la estabilidad estructural del genoma frente a pequeñas ganancias en depuración.

4. Mapeo y Profundidad de Cobertura

Los reads de las líneas evolucionadas (evol1 y evol2) se mapearon contra el ensamblaje ancestral usando `BWA` y `SAMtools`.

Cobertura

La profundidad promedio de cobertura fue:

- **Evol1:** 54.53×
- **Evol2:** 51.25×

Estos valores son adecuados para el análisis posterior de variantes, ya que superan el umbral mínimo recomendado (20–30×).

Indexación del genoma y mapeo

En bioinformática, el **indexar** significa crear una estructura de datos auxiliar que actúa como un "índice" o mapa de acceso rápido al contenido de un archivo. Al igual que un índice en un libro, este proceso nos permite localizar información sin tener que leer y recorrer TODO el contenido del archivo de principio a fin.

En el caso del genoma de referencia, la indexación genera tablas que permiten al alineador (`BWA`) identificar de forma eficiente los lugares exactos donde cada read puede encajar. Esto reduce significativamente el tiempo de búsqueda y nos hace posible trabajar con genomas de gran tamaño.

Por otro lado, la indexación de los archivos BAM produce ficheros `.bai` que permiten rápidamente el acceso aleatorio a regiones específicas del alineamiento. Esta indexación es fundamental para la visualización en herramientas como IGV, ya que evita cargar en la memoria todo el archivo completo y permite explorar de forma interactiva los segmentos del genoma ensamblado y las lecturas asociadas.

Interpretación de flagstat

- **Evol1:** 99.92 % de reads mapeados, 99.38 % correctamente apareados.
- **Evol2:** 97.80 % de reads mapeados, 97.07 % correctamente apareados.

Estos valores reflejan un alineamiento bueno, con perdidas casi mínimas de información y una baja tasa de lecturas descartadas.

5. Conclusiones

- El trimming optimizado mejoró la calidad general de los reads sin perder excesiva información.

- El ensamblaje RAW mostró métricas superiores en continuidad, mientras que los datos TRIM garantizan un menor sesgo y mayor fiabilidad.
- El mapeo de evol1 y evol2 contra el genoma ancestral mostró alta cobertura y porcentajes de alineamiento adecuados, validando la estrategia experimental.
- Las métricas de QUAST y las estadísticas de mapeo confirman que los datos son adecuados para futuros análisis de variantes y evolución.