Proyecto integrador en Bioinformática

Laura del Sol González
Maria Camila Gómez Villegas
Sara Floréz Hernández
Sergio

Universidad EIA
Envigado – Antioquia
Septiembre 2025

1. Resumen QC

La secuenciación de nueva generación (NGS), como la plataforma illumina , permite obtener millones de lecturas cortas a partir de fragmentos de ADN mediante ciclos de síntesis y detección por fluorescencia amplificados por PCR window) y se establecen longitudes mínimas de lectura. Estas etapas aseguran que las secuencias utilizadas en los análisis bioinformáticos posteriores sean de mayor. Por esta razón, antes de realizar ensamblajes o análisis posteriores, es fundamental aplicar un control de calidad con herramientas como FastQC, que permiten evaluar parámetros clave de las lecturas crudas y determinar su confiabilidad. Posteriormente, los datos deben ser sometidos a un proceso de depuración o trimming, en el que se eliminan adaptadores, se recortan regiones de baja calidad (por ejemplo, mediante sliding window) y se establecen longitudes mínimas de lectura. Estas etapas aseguran que las secuencias utilizadas en los análisis bioinformáticos posteriores sean de mayor calidad y representatividad, lo que incrementa la exactitud de los resultados obtenidos.

1.1.Ancestral

Los reportes de FastQC para la muestra ancestral (R1 y R2) evidenciaron una calidad promedio global cercana a Q35, lo cual indica lecturas confiables. En ambos reads se observó un sesgo en la composición de bases durante los primeros nucleótidos, patrón común asociado al bias de secuenciación en Illumina. La longitud de las secuencias crudas osciló entre 35 y 150 pb, y se detectaron secuencias sobre-representadas, principalmente compuestas por nucleótidos “N”. Para mitigar estos problemas se aplicó un recorte en los primeros 10 pb, junto con un filtrado por calidad (sliding window con $Q \geq 30$) y una longitud mínima de 80 pb, con el fin de conservar únicamente lecturas de alta calidad. Tras la depuración, la calidad global se mantuvo estable en Q35, el sesgo de contenido de bases se corrigió, la longitud de las lecturas se estandarizó en el rango de 80–130 pb y se eliminaron las secuencias sobre-representadas, mejorando así la confiabilidad de los datos para los análisis posteriores.

1.2.Evol1

Los datos crudos de la muestra Evol1 presentaron una calidad promedio global de Q35, con un sesgo evidente en la composición de bases en los primeros nucleótidos. La longitud de las lecturas se distribuyó entre 35 y 150 pb, con un porcentaje de lecturas únicas cercano al 85%, lo que refleja un nivel bajo de duplicación. Para la depuración se aplicó un recorte de las primeras 20 bases, un filtrado con sliding window $Q \geq 30$ y una longitud mínima de 80 pb. Tras el procesamiento, se obtuvo un conjunto de lecturas limpias con calidad promedio de Q34, en el que el sesgo inicial fue corregido y las secuencias se estandarizaron en un rango uniforme de ~127 pb. Además, el porcentaje de lecturas únicas aumentó ligeramente hasta 86.7% y no se detectaron secuencias sobre-representadas, confirmando una mejora en la homogeneidad y confiabilidad de los datos para análisis posteriores.

1.3.Evol2

Los datos crudos de Evol2 presentaron una calidad promedio de Q34, con un sesgo marcado en la composición de bases durante los primeros nucleótidos. La longitud de las lecturas se distribuyó entre 35 y 150 pb, y el porcentaje de lecturas únicas alcanzó el 86%, lo que refleja un bajo nivel de duplicación. Se aplicaron los mismos parámetros de depuración utilizados en Evol1 (recorte de 20 bases iniciales, sliding window con $Q \geq 30$ y longitud mínima de 80 pb). Tras el filtrado, la calidad global se mantuvo en Q34, el sesgo de composición al inicio de las lecturas se corrigió y las secuencias se estandarizaron en un

rango de ~127 pb. Asimismo, el porcentaje de lecturas únicas aumentó ligeramente a 87% y no se detectaron secuencias sobre-representadas, lo que mejora la uniformidad y la confiabilidad de los datos para análisis posteriores.

2. **¿Por qué se emplean los reads de la línea ancestral, y no la línea evolucionada para ensamblar el genoma?**

Se utiliza la línea ancestral para ensamblar el genoma porque representa el punto de partida y referencia basal del experimento. Las líneas evolucionadas pueden acumular mutaciones, inserciones, deleciones y reordenamientos que introducen heterogeneidad en las lecturas, lo que aumenta la fragmentación y reduce la precisión del ensamblaje de novo. Por ello, el flujo estándar en evolución experimental es ensamblar un genoma de referencia a partir de la línea ancestral y luego mapear las lecturas de las líneas evolucionadas para identificar las variantes adquiridas con mayor confiabilidad.

3. **Resultados de Quast.**

Las métricas generadas por el informe del quast asociadas para el caso de los scaffolds en el ancestro antes de realizar la limpieza las cuales se tomarán como referencia para comparar con el resultado del informe generado en el ancestro luego de la limpieza fueron las siguientes: Métricas para ancestral crudo: N50: 51960, N90: 13692, L50: 28, L90: 92; métricas para el ancestro limpio: N50: 23694, N90: 6386, L50: 58, L90: 206.

Con base en estos valores y teniendo en cuenta lo que nos indica cada una de estas métricas podríamos analizar para el caso de N50 y N90:
Nos indica que, el largo del scaffold más corto del conjunto que se usa para poder reconstruir el 50% del genoma es menor en el N50 y el N90 en el reporte del ancestral limpio que en el ancestral crudo es mayor en el ancestral crudo, a su vez esto comunica que los scaffolds quedaron más fragmentados y tendrán que usarse más para lograr el objetivo de reconstruir el genoma, en complemento podemos observar la métrica N'per 100 Kb la cual nos habla de la cantidad de bases desconocidas en el reporte del ensamble, para el caso del limpio fue más bajo lo cual nos indica que a pesar de su poca continuidad con respecto al crudo se tiene mayor confiabilidad en la reconstrucción del genoma.

Para el caso de las métricas L50 y L90 tenemos que:
Esta métrica nos indica la cantidad de scaffolds usados para reconstruir el 50 y 90 por ciento del genoma respectivamente, en concordancia entonces con el análisis anterior al tener scaffolds más continuos en el caso del ancestral crudo, se necesitaron menos cantidad de scaffolds a diferencia del ancestral limpio para el cual se necesitaron más de hecho el doble, pasando de 28 a 58 scaffolds para el 50% y de 92 a 206 para el 90%, esto igualmente se ve afectado por la métrica que N'per 100 kb la cual nos indica mejor confiabilidad en los datos limpios a pesar de la fragmentación.

3.1. Comparación entre scaffolds crudos vs. Scaffolds limpios

Como se mencionó anteriormente a pesar de mostrar mejores métricas el ensamble del genoma crudo, en cuanto a la calidad y exactitud se nota mejor confiabilidad en el ensamble realizado en el genoma limpio, por lo cual entonces no necesariamente unas mejores métricas de ensamble en cuanto a longitud de scaffolds y cantidad necesaria de estos indica que a priori los resultados sean los mejores ni confiables, siempre pueden

observarse otras métricas que nos darán mayor información, para este caso podría decirse que es preferible usar los datos limpios a pesar de estar más fragmentados ya que tiene menos bases desconocidas y para los siguientes pasos de identificación de variantes puede ser preferible, en cuestión que de paso con la mayor fragmentación en los datos limpios pueden explicarse por los parámetros usados en algunos de los pasos anteriores particularmente parámetros de limpieza muy agresivos que pudieron ocasionar perdida de información o incluso por la calidad inicial de las muestras.

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

Indexar el genoma es un paso previo al alineamiento de los reads contra un genoma de referencia. Lo que se hace es preparar el archivo .fasta del genoma para que los programas de alineamiento, como BWA o Bowtie2, puedan trabajar de manera más rápida y organizada. Cuando uno indexa, el programa genera unos archivos adicionales que funcionan como un mapa o guía para que no tenga que leer todo el genoma de principio a fin cada vez que entra una lectura. En cambio, con esos archivos ya puede ir directamente a la región que corresponde, lo que hace que el proceso sea mucho más eficiente. Una comparación sencilla es pensar en un libro de miles de páginas. Si no tiene índice, uno se demora un montón buscando el tema que necesita. Pero si sí tiene índice, uno revisa el capítulo y va directo a la página que corresponde. Con el genoma pasa lo mismo: el índice evita que el programa pierda tiempo buscando en todo el archivo. En este proceso se crean varios archivos auxiliares, como .bwt, .pac, .amb, .ann y .sa. Cada uno tiene información que ayuda al programa a encontrar rápido las coincidencias. Por ejemplo, el .bwt está relacionado con la transformación de Burrows-Wheeler, que es la base de cómo trabaja BWA. Este paso se hace con el comando bwa index. Aunque en el archivo .fasta no se ve que ocurra nada, esta paso es un requisito fundamental. Si no se hace, el alineador ni siquiera reconoce el genoma y el mapeo no corre.

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

Cuando se termina el alineamiento de los reads, se tiene un archivo .bam. Este archivo es la versión comprimida y en binario del .sam y guarda toda la información sobre las lecturas y cómo se alinearon contra el genoma. El problema es que para abrir ese archivo en IGV no basta con tener el .bam. Se necesita también un archivo .bai, que es el índice del .bam. Ese archivo se crea con el comando samtools index. El .bai cumple la función de guía dentro del archivo .bam: le dice a IGV en qué parte del archivo están los datos de una región específica del genoma. Gracias a eso, IGV puede cargar rápidamente solo la parte que uno quiere ver en lugar de tener que cargar todo el archivo completo en memoria, que puede ser muy grande y pesado. Antes de indexar, el archivo .bam debe estar ordenado por coordenadas, lo cual se hace con samtools sort. Si no está ordenado, el índice no funciona y el programa no permite navegar por el genoma. Si uno no hace ese paso, IGV no carga bien el archivo, puede mostrar errores o simplemente quedarse en blanco. En cambio, con el índice ya se puede explorar el genoma de forma interactiva, revisar la cobertura de las lecturas, analizar variantes y comprobar cómo quedaron alineados los reads. En clase vimos que IGV es muy útil porque permite ver de manera gráfica cómo se distribuyen las lecturas y detectar cosas como regiones con poca cobertura, errores de alineamiento o mutaciones puntuales. Todo eso solo es posible si el archivo está indexado. Incluso recuerdo que me pasó una vez que IGV no me

abría nada y resultó ser porque me faltaba crear el .bai. Desde ahí entendí que indexar el mapeo es tan obligatorio como haber hecho el mapeo en sí.

6. Interpretación de los resultados de las estadísticas de mapeo (Qualimap).

Evol1: En la muestra Evol1 se analizaron un total de 1,412,170 lecturas, de las cuales un 99.9% lograron alinearsecorrectamente contra el genoma de referencia ancestral. Este porcentaje tan alto refleja que el ensamblaje del ancestro representa bien la información genética de la línea evolucionada y que el preprocesamiento de las lecturas fue adecuado. La cobertura promedio alcanzó 38.7X lo cual es más que suficiente para estudios en bacterias, donde usualmente con coberturas mayores a 20X ya se puede tener confianza en la detección de variantes. Aunque la cobertura no es completamente uniforme (desviación estándar de 20), sigue siendo bastante aceptable. En algunos contigs se observan picos de cobertura mucho más altos que en el resto, lo que podría estar relacionado con regiones repetitivas o con sesgos de secuenciación, pero estas zonas no afectan de manera crítica la interpretación global. Para la calidad de mapeo el valor medio fue de MQ = 52.7 que es muy alto, lo que significa que casi todas las lecturas quedaron ubicadas en el lugar correcto dentro del genoma. La tasa de error general fue de solo 0.86%, principalmente debida a mismatches, lo cual es normal considerando que se trata de una línea evolucionada donde pueden existir polimorfismos respecto al ancestro. Por ultimo la tasa de duplicación fue de 31%, lo que suele pasar en experimentos con PCR.

Evol2: En Evol2, se procesaron 1,351,692 lecturas, de las cuales un 97.8% se alinearon al genoma ancestral. Aunque sigue siendo un porcentaje muy alto, es menor al de Evol1, lo que podría deberse a que esta población acumuló más cambios respecto al ancestro o a que haya algunas lecturas contaminantes o de baja calidad que no lograron mapearse. La cobertura promedio fue de 37X que también es adecuada, pero en este caso la desviación estándar fue mayor 23 esto indica que la distribución de la cobertura es menos homogénea que en Evol1. Esto significa que hay regiones del genoma con mucha mayor cobertura y otras con menos, lo que podría reflejar sesgos técnicos de la secuenciación o también regiones repetitivas que dificultan el alineamiento. La calidad de mapeo fue muy similar a Evol1 MQ = 52.6 por lo que las lecturas que sí se mapearon lo hicieron con mucha confianza. La tasa de error general fue de 0.91%, un poco más alta que en Evol1 principalmente por mismatches, lo cual puede interpretarse como que Evol2 tiene un poco más de divergencia con respecto al ancestro. Además, la tasa de duplicación también fue alta 31%, muy similar a Evol1. Un detalle interesante es que en Evol2 el tamaño promedio del inserto fue mayor 210 pb vs 185 pb en Evol1, lo cual indica que la preparación de la librería resultó en fragmentos más largos.

Comparación:

Tabla 1. Comparación Evol1 vs Evol2

Métrica	Evol1	Evol2
Número total de lecturas	1,412,170	1,351,692
% Lecturas mapeadas	99.9%	97.8%

% Lecturas no mapeadas	0.1%	2.2%
Cobertura promedio	38.7X	37X
Desviación estándar de cobertura	20.2	23.1
Calidad media de mapeo (MAPQ)	52.7	52.6
Tasa de error general	0.86%	0.91%
% Lecturas duplicadas	31.9%	31.4%
Tamaño medio del inserto	185 pb	210 pb
GC%	51.1%	51.6%

Al comparar los resultados de Evol1 y Evol2 se ve que ambos mapeos son de buena calidad y adecuados para continuar con el análisis de variantes. Evol1 tuvo un porcentaje de mapeo más alto 99.9% y una cobertura un poco más uniforme, lo que me dice que esta línea evolucionada está más cercana al ancestro. Evol2, en cambio, tuvo un 97.8% de lecturas alineadas y una desviación mayor en la cobertura lo que muestra más variación en ciertas regiones del genoma. Si bien hay diferencias la calidad de mapeo, el nivel de cobertura y las tasas de error en ambas muestras son suficientes para confiar en los alineamientos y usarlos en los siguientes pasos del proyecto.

Referencias

- Harvard Chan Bioinformatics Core. (s. f.). *Quality control using FastQC*. HBC Training. Recuperado de https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/QC_raw_data.html
- Biostars. (2016). *What is the purpose of indexing a genome?* Biostars. Recuperado de https://www.biostars.org/p/212594/
- Li, H. (2009). *Burrow-Wheeler Aligner (BWA)* [Repositorio GitHub]. GitHub. https://github.com/lh3/bwa
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). *Integrative Genomics Viewer (IGV) User Guide*. Broad Institute. Recuperado de https://software.broadinstitute.org/software/igv/UserGuide
- Langmead, B., & Salzberg, S. L. (2012). *Bowtie 2 manual*. Bowtie. Recuperado de http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). *The Sequence Alignment/Map format and SAMtools*. *Bioinformatics, 25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint* arXiv:1303.3997. https://academic.oup.com/bioinformatics/article/29/8/1072/228832