Proyecto: Entrega 1

AUTORES:

Mariana Aguirre Llano

Sofia Ospina Aristizábal

Mariana Salazar Díaz Granados

Juanita Vélez Uribe

Profesor
Simón Villegas
Bioinformática

Universidad EIA
ENVIGADO
2025 – 02

1. **Resumen QC.**

En los datos crudos (PRE) las lecturas mostraron una calidad aceptable en general, aunque se observó la típica caída en la calidad hacia los extremos de las secuencias, algo muy común en datos de Illumina (Andrews, 2010). El contenido de GC presentó algunas advertencias, probablemente relacionadas con la composición particular del genoma del microorganismo, ya que este tipo de sesgos son normales en genomas pequeños o con regiones repetitivas (Schmieder & Edwards, 2011). También se detectó cierta variabilidad en la longitud de las lecturas y la presencia de secuencias sobre-representadas, lo que puede deberse a contaminantes o a regiones muy expresadas del genoma. A pesar de esto, los niveles de duplicación fueron aceptables y no se detectó un exceso de adaptadores.

Después del proceso de trimming (POST) se notó una mejora clara en la calidad de los reads. Las regiones de baja calidad en los extremos fueron eliminadas, lo que aumentó la confiabilidad de las secuencias (Bolger et al., 2014). Aunque el contenido de GC siguió marcando advertencias, esto no es preocupante porque suele estar relacionado con características biológicas del genoma y no necesariamente con errores técnicos. La longitud de las lecturas pasó a ser más variable, algo esperado tras el recorte, pero se redujo notablemente la presencia de adaptadores y secuencias sobre-representadas. En general, se perdió un pequeño porcentaje de lecturas, pero lo que quedó fue de mejor calidad, lo que resulta más útil para el ensamblaje posterior.

El análisis de ensamblaje con QUAST confirmó que el trimming tuvo un impacto positivo en la preparación de los datos. Los valores de N50 y L50 fueron altos y el número de contigs bajo, lo que indica que se logró una buena continuidad en los ensamblajes. La longitud total de los genomas ensamblados fue muy cercana a la de la referencia, lo cual demuestra que no hubo una pérdida importante de información. Además, el hecho de contar con contigs largos y pocos fragmentos respalda la calidad de los datos procesados (Gurevich et al., 2013).

En resumen, el trimming resultó clave para mejorar los datos: eliminó errores y contaminantes, redujo advertencias en las métricas de calidad y permitió ensamblajes más confiables. Aunque se sacrificó una pequeña fracción de lecturas, los resultados muestran que se mantuvo la integridad del genoma y que las tres líneas (ancestral y evolutivas) pudieron ensamblarse con buena continuidad, en línea con lo que han reportado otros estudios en análisis de datos microbianos (Del Fabbro et al., 2013).

2. **Discutir por qué se están empleando los reads de la línea ancestral, y no la línea evolucionada, para ensamblar el genoma.**

El propósito del presente proyecto es la identificación de un organismo a partir de su información genética. Para lograr esto, es necesario tener, en primer lugar, el genoma del organismo de interés y un genoma de referencia con el cual se pueda realizar una comparación que permita determinar si existen suficientes similitudes entre ambos que permitan caracterizar el organismo de interés.

En este contexto, se emplean los reads de la línea ancestral para ensamblar el genoma porque representan el estado original del microorganismo, sin las modificaciones acumuladas en las líneas evolucionadas. Esto proporciona una base más estable y confiable para la construcción del genoma de referencia, ya que trabajar con un genoma ancestral reduce la interferencia de mutaciones posteriores y mejora la eficiencia del ensamblaje y del mapeo de lecturas (Vieira et al., 2020; Anselmetti et al., 2015). A partir de esta secuencia ensamblada, las lecturas de las líneas evolutivas pueden compararse con el genoma ancestral para identificar similitudes, divergencias o mutaciones, lo que no solo permite correlacionar los organismos, sino también analizar los cambios adaptativos que se han dado a lo largo de la evolución experimental.

3. Resultados de Quast.

A) Definir y discutir las distintas métricas (N50, N90, auN, L50, L90).

- **N50:** Corresponde a la longitud mínima de los contigs más largos que, en conjunto, cubren el 50 % del tamaño total del ensamblaje, en donde, un valor alto indica que gran parte del genoma está representada por contigs largos, lo que sugiere un ensamblaje más continuo y menos fragmentado (Miller et al., 2010).
  En este caso, el ensamblaje de scaffolds_raw presenta un N50 de 57,404 bp, superior al de contigs_raw (48,050 bp) y también mayor al de los ensamblajes con trimming, lo que indica que los contigs en el ensamblaje sin trimming son en promedio más largos, es decir, que hay una mejor continuidad genómica y menor fragmentación.

- **N90:** Es la longitud mínima de los contigs que cubren el 90 % del ensamblaje, lo que permite evaluar la continuidad de la mayor parte del genoma, identificando si las secuencias que componen el ensamblaje en su tramo final son cortas o largas (Bradnam et al., 2013).
  Para scaffolds_raw se obtiene un N90 de 13,786 bp, que también es superior al de contigs_raw (12,320 bp). Esto sugiere que incluso los contigs más pequeños que componen la mayor parte del genoma son relativamente largos, favoreciendo un ensamblaje más representativo y con menos gaps.

- **auN:** Representa el área bajo la curva Nx y resume la continuidad global del ensamblaje, además de que es menos sensible a valores atípicos (Pevzner et al., 2001). En nuestro reporte se observa que el valor obtenido para scaffolds_raw (59,411.2) es mayor que en los otros ensamblajes (por ejemplo, 54,258.9 en contigs_raw). Por lo que, se puede confirmar que el ensamblaje de scaffolds es más uniforme y no está dominado por contigs extremadamente cortos o largos, reflejando una buena calidad general.

- **L50:** Indica el número mínimo de contigs que cubren el 50 % del ensamblaje. Pues bien, en cuanto menor sea este valor, más concentrada está la información genómica en pocos contigs, lo que indica un ensamblaje más robusto (Miller et al., 2010).
  En nuestros resultados se ve que en scaffolds_raw, el valor es 28, menor que en contigs_raw (30), lo que implica que se necesita un menor número de contigs para cubrir la mitad del genoma, favoreciendo un ensamblaje menos fragmentado y, por lo tanto, más robusto.
- **L90:** Es el número mínimo de contigs que cubren el 90 % del ensamblaje, en donde se refleja con mayor precisión el nivel de fragmentación del genoma, mostrando cuántos contigs son necesarios para abarcar casi toda la secuencia (Bradnam et al., 2013).
  En nuestro caso, scaffolds_raw tiene un L90 de 92, mejor que los 100 contigs requeridos en contigs_raw. Por lo tanto, esto refuerza la idea de que la secuencia está concentrada en menos fragmentos, lo que es deseable para un ensamblaje de calidad.

B) Discutir si mejores métricas significan mejor ensamblaje.

Las métricas como N50, N90 y auN con valores altos, junto con valores bajos de métricas como L50 y L90, suelen asociarse con ensamblajes más continuos y menos fragmentados. No obstante, estos indicadores no siempre reflejan una mayor exactitud biológica, ya que un valor de N50 elevado puede deberse a misassemblies, es decir, uniones incorrectas de secuencias que generan errores en el ensamblaje (Bradnam et al., 2013).

Por ello, estas métricas deben ser interpretadas junto con otras evaluaciones, como la tasa de genes recuperados y la detección de errores de ensamblaje, usando herramientas como QUAST (Miller et al., 2010). Por lo tanto, se puede concluir que mejores métricas sugieren un ensamblaje más continuo, pero no garantizan que sea el más correcto biológicamente**.**

C) Comparar el ensamblaje de los datos crudos vs. los datos depurados por calidad. ¿Cuál ensamblaje prefieres y por qué?

Después de realizar el reporte y analizar los resultados de QUAST, se observa que el ensamblaje con datos crudos presenta mejores métricas de continuidad, como un N50 más alto (57,404 vs. 50,016 en el depurado) y menor fragmentación (L90 = 92 vs. 95). Sin embargo, esto no necesariamente implica que sea el mejor ensamblaje, ya que, un factor muy importante para tener en cuenta es el número de N's por cada 100 kb, debido a que estos representan regiones donde el ensamblador no pudo determinar la secuencia real y dejó gaps (Bradnam et al., 2013). Teniendo en cuenta lo anterior, se puede decir que el ensamblaje depurado tiene una secuencia más confiable debido a que presenta una menor cantidad de N's en comparación con el ensamblaje de datos crudos (22.29 vs. 22.69).

Este punto es clave, porque tener menos N's significa que el ensamblaje está mejor soportado por las lecturas y reduce el riesgo de que existan regiones con información incierta, lo que puede introducir errores en el análisis biológico (Miller et al., 2010), lo que resulta fundamental para análisis posteriores, por ejemplo, el mapeo de lecturas, porque si la referencia tiene

demasiados gaps, las lecturas pueden quedar sin alinear o alinearse en posiciones incorrectas, afectando la interpretación de los resultados (Li & Durbin, 2009).

Por lo tanto, se prefiere el ensamblaje generado con datos depurados, ya que, aunque presenta una ligera fragmentación adicional, es biológicamente más confiable y ofrece una representación más precisa del genoma. Esto asegura que, al realizar el mapeo, la tasa de alineación sea mayor y que los resultados observados en IGV sean más representativos y útiles para la detección de variantes (Robinson et al., 2011).

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

Indexar el genoma hace referencia al proceso de generar archivos auxiliares que permiten acceder de forma rápida a cualquier posición de la secuencia de referencia. Pues bien, este proceso no modifica el genoma, sino que crea una especie de "mapa" que facilita a los alineadores (como BWA) ubicar eficientemente cada lectura en su posición correspondiente.

La indexación es necesaria porque el alineamiento de millones de lecturas sería computacionalmente muy costoso si el software tuviera que recorrer toda la secuencia cada vez que compara una lectura. No obstante, gracias al índice, el alineador puede buscar posiciones potenciales en tiempo mucho menor, haciendo que el mapeo sea más rápido y reproducible (Li & Durbin, 2009).

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

Ver el resultado del mapeo en IGV significa explorar gráficamente cómo se alinearon las lecturas de secuenciación contra el genoma de referencia, permitiendo identificar regiones con buena cobertura, variantes, o posibles errores.

Para esto es necesario indexar el mapeo, lo que consiste en generar un archivo auxiliar (.bai para archivos BAM) que funciona como un índice, permitiendo que IGV cargue de forma rápida únicamente la región del genoma que se desea visualizar, sin leer todo el archivo completo. Sin esta indexación, la carga de datos sería extremadamente lenta o incluso impráctica (Robinson et al., 2011)

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

En la línea evolutiva 1, se obtuvieron cerca de dos millones de lecturas, de las cuales el 98.84% se alinearon al genoma de referencia derivado de la línea ancestral. Este porcentaje de mapeo es considerado muy alto en análisis de secuenciación, ya que indica una fuerte correspondencia entre las lecturas y la referencia, lo cual respalda la calidad de la preparación de la librería y del proceso de secuenciación (Li & Durbin, 2009). Además, el 97.74% de las lecturas estuvieron correctamente emparejadas, lo que sugiere que la gran mayoría de las secuencias se encuentran en la orientación y distancia esperada, lo cual es un indicador de integridad de los datos (Li et al., 2009). El bajo porcentaje de singletons (0.32%) confirma que casi todas las lecturas pudieron ser alineadas junto a su par, reduciendo la probabilidad de errores sistemáticos en el mapeo.

En la línea evolutiva 2, los resultados fueron consistentes con los de la línea 1, con porcentajes de mapeo cercanos al 99%. Esta alta tasa de alineamiento muestra que, a pesar de los procesos evolutivos experimentales, el genoma de la línea 2 mantiene una alta similitud con el genoma ancestral. La cobertura observada fue uniforme, lo cual es importante ya que una distribución homogénea de la profundidad de lectura a lo largo del genoma minimiza sesgos y asegura que la mayoría de las regiones están representadas de manera confiable (Okou et al., 2015). Asimismo, la proporción elevada de lecturas correctamente emparejadas refleja que no hubo grandes reordenamientos estructurales que impidieran el mapeo.

En conjunto, las estadísticas de mapeo de Qualimap muestran que ambas líneas evolutivas presentan tasas de alineamiento superiores al 98%, con una gran proporción de lecturas correctamente emparejadas y niveles bajos de singletons. Estos resultados indican que los genomas evolucionados conservan una alta similitud con el genoma ancestral, y que las diferencias que se identifiquen en análisis posteriores —como polimorfismos de un solo nucleótido (SNPs) o indeles— serán cambios puntuales más que reestructuraciones genómicas amplias. Esto concuerda con lo observado en estudios de evolución experimental, donde los genomas de las poblaciones evolucionadas suelen mantener una alta conservación global respecto al ancestro, mientras acumulan mutaciones específicas que reflejan adaptaciones al ambiente (Barrick & Lenski, 2013).

7. Bibliografía

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics. Disponible en: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Anselmetti, Y., Berry, V., Chauve, C., Ponty, Y., Scornavacca, C., & Tannier, E. (2015). Ancestral gene synteny reconstruction improves extant species scaffolding. *BMC Genomics, 16*(S10), S11. https://doi.org/10.1186/1471-2164-16-S10-S11

Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics, 14*(12), 827–839. https://doi.org/10.1038/nrg3564

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience, 2*(1). https://doi.org/10.1186/2047-217X-2-10

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS ONE, 8*(12), e85024. https://doi.org/10.1371/journal.pone.0085024

FastQC. (2015). *FastQC documentation*. Babraham Institute. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

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

Illa, E., Sargent, D. J., Girona, E. L., et al. (2011). Comparative analysis of rosaceous genomes and the reconstruction of a putative ancestral genome for the family. *BMC Evolutionary Biology, 11*, 9. https://doi.org/10.1186/1471-2148-11-9

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., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., & Zwick, M. E. (2015). Microarray-based genomic selection for high-throughput resequencing. *Nature Methods, 4*(11), 907–909. https://doi.org/10.1038/nmeth1110

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences, 98*(17), 9748–9753. https://doi.org/10.1073/pnas.171285098

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology, 29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics, 27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

Vieira, G., Lassalle, F., Cornet, L., Ginevra, C., & Lechat, P. (2020). Using in silico predicted ancestral genomes to improve the efficiency of paleogenome reconstruction. *Ecology and Evolution, 10*(15), 8205–8216. https://doi.org/10.1002/ece3.6925