Discusión de Resultados - Grupo 4

Resumen del Quality Control

Para realizar este análisis se simula un flujo de trabajo; el primer paso es el control de calidad (QC), donde se evalúa la calidad de las lecturas crudas obtenidas de la secuenciación. Esta etapa permite identificar errores y regiones de baja confianza que pueden afectar el ensamblaje y análisis posteriores.

Al observar los reportes iniciales de *fastq* y *MultiQC* se observa que los parámetros básicos de calidad son adecuados. En las gráficas de Per Base Sequence Quality el valor Phred promedio supera 26, lo que indica un nivel alto. Sin embargo, en la línea celular evolucionada, las primeras posiciones muestran una dispersión amplia, con algunos reads por debajo de 20, lo que también sucede a lo largo de toda la secuencia y se acentúa en los extremos. Esto se podría explicar por la diferencia de 0.7 millones de reads frente a la línea ancestral, pero no representa un problema crítico ya que el promedio se mantiene en rangos aceptables.

Otros parámetros señalan una falla en el contenido relativo de ATGC, con valores disparados consistentes entre lecturas, lo que sugiere un artefacto técnico confirmado por la alerta de secuencias sobre representadas. También se reportaron alertas en el contenido de GC, aunque se conserva una distribución simétrica y baja desviación; y en la longitud de reads, probablemente por la tagmentación, ya que se observan fragmentos más cortos, pero no es una cantidad significativa, la alerta solo significa que hay al menos un read de tamaño diferente a los demás.

En general, las lecturas presentan buena calidad y sería posible un ensamblaje de novo. Sin embargo, dado que el objetivo es el análisis de variantes, se requiere mayor exigencia para garantizar resultados confiables. La literatura recomienda valores Phred >30 para evitar falsos positivos (Illumina, 2011; O'Rawe et al., 2015), pero se estableció un umbral de trimming en 28 para evitar pérdida significativa de reads, complementado con un sliding window de 28 y una depuración por longitud mínima de 80 pb, teniendo en cuenta que la mayoría de las lecturas presentaban longitudes superiores a 140 pb.

En la gráfica Per Base Sequence Content se evidenció una variación inusual en las primeras 10 pb del extremo 5´ y las últimas 5 pb del extremo 3´, consistente en todas las lecturas. Aunque inicialmente se sospechó de adaptadores, se descartó tras comparar con secuencias de Illumina. Al intentar recortar estas bases, la calidad del mapeo disminuyó, especialmente en profundidad, que bajó de 54x a 34x. Si bien 30x es suficiente para llamado de variantes, la literatura indica que algunas librerías presentan un sesgo en el extremo 5' (Illumina Inc., 2020), lo que produce advertencias en composición de bases, pero no puede corregirse por recorte y en la mayoría de los casos no impacta de manera adversa los análisis posteriores (*Per Base Sequence Content*, n.d.). Considerando que el recorte reducía el N50 y la profundidad, se decidió no realizar esta modificación.

Por qué se están empleando los reads de la línea ancestral, y no la línea evolucionada, para ensamblar el genoma

Al construir un genoma de referencia con la cepa ancestral, se establece una base, que representa el estado inicial del microorganismo, y los reads de las poblaciones evolucionadas se alinean contra esta base. Así, las diferencias que se detecten por INDEL´s o SNP´s pueden atribuirse directamente al proceso de evolución experimental.

Esto permite responder la pregunta biológica planteada pues las mutaciones observadas pueden asociarse con mejoras en el crecimiento, resistencia a estrés, eficiencia metabólica, etc. Lo que conecta directamente las variaciones genómicas con los fenómenos fenotípicos adaptativos.

Resultados de Quast y selección del ensamblaje

La herramienta **Quast** permite evaluar la calidad de ensamblajes genómicos mediante distintas métricas de continuidad y precisión, en este caso se realizó una comparación entre el genoma de referencia ancestral ensamblado a partir de los datos crudos y de los datos depurados con trimming.

Para determinar cuál genoma de referencia ensamblado es el más adecuado, se evaluaron las siguientes métricas (Gurevich et al., 2013) (*How-to Guides*, s. f.):

- **# de contigs:** número de fragmentos en los que se divide el genoma ensamblado. Un menor número indica un ensamblaje más continuo y con menos fragmentación.
- **Largest contig:** longitud del fragmento más grande ensamblado. Un valor mayor refleja un ensamblaje de mayor calidad, ya que implica que el ensamblador logró reconstruir regiones largas del genoma sin interrupciones.
- **N50:** longitud mínima tal que al menos el 50% del genoma está contenido en *contigs* de igual o mayor tamaño. Cuanto mayor sea el N50, mejor es la continuidad del ensamblaje, ya que indica que la mayoría de los fragmentos son suficientemente largos y consistentes.
- **N90:** similar al N50, pero considerando el 90% del genoma.
- **auN:** similar a N50/N90, ya que integra todos los posibles valores (desde el 1% hasta el 100%) y calcula el área bajo la curva acumulativa de longitudes de *contigs*., valores más altos de auN significan mejor ensamblaje.
- **L50 / L90:** número de contigs necesarios para cubrir el 50% o 90% del genoma; valores menores son preferibles, pues indican que una mayor proporción del genoma está representada en menos fragmentos.
- **# de N's por 100 kbp:** cantidad de bases ambiguas (N) por cada 100,000 bases; Un valor menor señala mayor precisión y menor incertidumbre en el ensamblaje.
- **# total de N's:** número total de bases ambiguas en el ensamblaje; Valores menores indican un ensamblaje más confiable, con menos huecos cubiertos por símbolos de ambigüedad.

Tabla 1:Resultados de quast

Métricas	filter_scaffold_anc_raw (R)	filter_scaffold_anc_trimmed (T)	/
Número de contigs	191	205	**(R)**
Largest contig	157883	128117	**(R)**
N50	60681	47571	**(R)**
N90	13213	11951	**(R)**
auN	60853	52313	
L50	26	31	**(R)**
L90	91	102	**(R)**

Número de N’s por 100 kbp	16.11	18.11	**(R)**
Número total de N’s	730	820	**(R)**

De acuerdo con los datos proporcionados por Quast, el genoma ensamblado con los datos crudos presenta mejores valores en la mayoría de las métricas en comparación con el ensamblaje generado a partir de los datos depurados usando trimming.

Sin embargo, se decidió seleccionar como genoma de referencia el ensamblado con datos depurados. Esta decisión se justifica en que, aunque las métricas del ensamblaje crudo sean ligeramente mejores, el ensamblaje con los datos depurados se encuentra condicionado por los cortes realizados en el proceso de QC, lo que garantiza una mayor calidad en las lecturas utilizadas para el ensamblaje. Esto aporta confiabilidad en el llamado de variantes, donde la calidad de los datos iniciales es determinante para obtener mejores resultados en el análisis posterior de los genomas **Evol1** y **Evol2**.

Para estimar la profundidad y la cobertura del ensamblaje se emplearon los comandos ´samtools depth´ y ´samtools coverage´. Estos análisis proporcionaron información sobre la cantidad promedio de veces que cada posición fue leída y el porcentaje total del ensamblaje cubierto por lecturas. Los resultados mostraron una profundidad promedio de 47**x** y una cobertura de **%**, lo que confirma que tras la depuración de los datos el ensamblaje tiene una representación adecuada para su análisis posterior.

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

La indexación de genomas es un paso fundamental en el análisis bioinformático moderno, que permite la búsqueda eficiente y el mapeo de secuencias de ADN o ARN, dadas las crecientes bases de datos genómicas (Alser et al., 2022). La indexación del genoma se refiere al preprocesamiento de la secuencia de referencia para generar unas estructuras llamadas índices, estos índices facilitan la localización rápida de subsecuencias genómicas dentro de un genoma de referencia, lo cual es crucial para el alineamiento de lecturas y otras tareas computacionales (Alser et al., 2021). Este proceso es indispensable debido al volumen masivo de datos generados por las tecnologías de secuenciación de alto rendimiento, que han reducido drásticamente los costos y tiempos de secuenciación (Jalili et al., 2018).

Es necesario indexar el genoma para facilitar búsquedas exhaustivas y análisis complejos (Alser et al., 2022), lo que permite ahorrar tiempo y memoria de procesamiento, permitiendo un manejo eficiente de los recursos computacionales y una aceleración significativa en las distintas etapas del análisis genómico, desde la identificación de variantes hasta el ensamblaje de novo (Alser et al., 2022) (Alser et al., 2020). Esta capacidad de procesamiento es vital, ya que permite a los investigadores explorar patrones genéticos complejos y realizar descubrimientos significativos (Montecuollo & Schmid, 2020). La indexación genómica es particularmente importante para el ensamblaje de novo, donde se reconstruye un genoma completo a partir de fragmentos cortos, y para la identificación de variaciones genéticas, como polimorfismos de un solo nucleótido e inserciones/deleciones (indels) (Jalili et al., 2018). Además, es necesario debido a que la mayoría de los programas no pueden recibir un fasta sin indexar, los archivos indexados son precisamente los que les permiten navegar por el genoma.

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

El mapeo o alineamiento, es el proceso bioinformático mediante el cual los reads de ADN generadas por el secuenciador se alinean o "mapean" contra un genoma de referencia conocido.

El resultado del mapeo se suele almacenar en un archivo BAM, que es un archivo grande el cual contiene millones de lecturas alineadas y está ordenado por coordenadas genómicas. Para visualizarlo de manera eficiente se debe indexar, debido a que la indexación crea un archivo auxiliar, que contiene la extensión ".bai", este archivo actúa como un índice de contenidos para el archivo BAM (Li et al., 2009) que se debe ingresar al programa. Este archivo ".bai" le indica a IGV dónde buscar en el archivo ".bam" para encontrar las lecturas que se alinean a una región genómica específica, permitiendo el acceso aleatorio rápido a los datos (Thorvaldsdottir et al., 2012) (Li et al., 2009).

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

Tras el ensamblaje del genoma de referencia Ancestral, se realizaron los mapeos con los genomas Evol1 y Evol2, obteniéndose resultados satisfactorios que confirman la calidad del ensamblaje y del proceso de alineamiento:

1. Mapeo con Evol1:

De un total de 1,801,148 lecturas, el 100 % se alineó con el genoma de referencia, lo que respalda la pertinencia del ensamblaje. Sin embargo, se observó una proporción considerable de lecturas duplicadas, aspecto que debe tenerse en cuenta en análisis posteriores.

La profundidad media de cobertura fue de 55.34X, valor considerado adecuado, ya que para lecturas de Illumina se acepta como mínimo 50X.

La calidad de mapeo alcanzó un valor de 58.54, lo que refleja una alta confiabilidad, pues valores cercanos a 60 representan mayor precisión en el alineamiento.

En cuanto a la tasa de error general, se obtuvo 1.44 %, lo que indica buena fidelidad en las lecturas. No obstante, se identificó un 49 % de INDEL's en regiones homopoliméricas, lo cual corresponde a sesgos conocidos en plataformas Illumina.

2. Mapeo con Evol2:

En este caso, de un total de 1,672,482 lecturas, el 100 % también se alineó correctamente con el genoma de referencia, confirmando nuevamente la solidez del ensamblaje. Al igual que en Evol1, debe considerarse la presencia de lecturas duplicadas.

La profundidad media de cobertura fue de 52.47X, valor que cumple con los parámetros de calidad para Illumina, aunque ligeramente inferior al obtenido en Evol1.

La calidad de mapeo alcanzó 58.42, lo que representa una excelente confiabilidad en el alineamiento.

La tasa de error general fue de 1.51 %, muy similar a la observada en Evol1, indicando buena fidelidad de lectura. Asimismo, se presentó un porcentaje comparable de INDEL´s en regiones homopoliméricas, consistente con los sesgos técnicos esperados de Illumina.

En conclusión, el mapeo de los genomas Evol1 y Evol2 frente al genoma de referencia ensamblado fue exitoso, mostrando una cobertura adecuada, alta calidad de alineamiento y baja tasa de error. Estos resultados validan la confiabilidad del ensamblaje y permiten avanzar con el análisis de llamado de variantes.

Referencias

Alser, M., Rotman, J., Deshpande, D., Taraszka, K., Shi, H., Baykal, P. I., Yang, H. T., Xue, V., Knyazev, S., Singer, B. D., Balliu, B., Koslicki, D., Skums, P., Zelikovsky, A., Alkan, C., Mutlu, O., & Mangul, S. (2021). Technology dictates algorithms: recent developments in read alignment. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02443-7

Alser, M., Eudine, J., & Multu, O. (2022). Taming Large-Scale genomic analyses via sparsified genomics. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-2277358/v1

Accelerating Genome Analysis: a primer on an ongoing journey. (n.d.). IEEE Journals & Magazine | IEEE Xplore. https://doi.org/10.1109/MM.2020.3013728

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

How-to guides. (s. f.). How-to Guides. https://australianbiocommons.github.io/how-to-guides/genome_assembly/assembly_qc#:~:text=Informe%20QUAST,-https://bio&text=QUAST%20devuelve%20diversas%20m%C3%A9tricas%20que,menor%20sea%20el%20L50%2C%20mejor.

Illumina. (2011). *Quality Scores for Next-Generation Sequencing*. http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/

Illumina Inc. (2020). *Preparación de ADN de Illumina sin PCR, tagmentación*. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/illumina-dna-pcr-free-data-sheet-770-2020-003-translations/illumina-dna-pcr-free-data-sheet-770-2020-003-esp.pdf

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

Montecuollo, F., & Schmid, G. (2020). ER-index: A referential index for encrypted genomic databases. *Information Systems*, *96*, 101668. https://doi.org/10.1016/j.is.2020.101668

Next generation indexing for genomic intervals. (n.d.). IEEE Journals & Magazine | IEEE Xplore. https://doi.org/10.1109/TKDE.2018.2871031

O'Rawe, J. A., Ferson, S., & Lyon, G. J. (2015). Accounting for uncertainty in DNA sequencing data. *Trends in Genetics*, *31*(2), 61–66. https://doi.org/10.1016/J.TIG.2014.12.002

Per Base Sequence Content. (n.d.). Retrieved September 29, 2025, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2012). 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