Julieta López - Sofía Upegui – Sebastián Reina – Mateo Pulgarín

Proyecto Integrador Bioinformática – Entrega 1

Resumen QC

Al hacer el Control de calidad (QC) de los datos se determinó un error normal de ilumina, que indica una caída de la calidad a medida que aumenta la lectura de la secuencia. Una de las razones es debido a la disminución de la intensidad de la señal fluorescente, con cada ciclo de secuenciación, la intensidad de la señal disminuye porque los fluoróforos se degradan, lo que lleva a una incorporación incompleta de nucleótidos en algunas hebras. De la misma forma una baja calidad al inicio de la secuenciación es debido a que la polimerasa necesita un tiempo mientras se estabiliza. Adicionalmente, se determinó que un sector inicial de los reads mostraba poca homogeneidad en el contenido de secuencia por base, lo que luego se tomó como normal al compararlo con las gráficas ideales proporcionadas por los manuales de QC que explican que este es un comportamiento normal de las lecturas de ilumina. Esto se da por *mis-priming*, un evento en el que secuencias en las hebras de ADN por secuenciar son muy similares a los adaptadores en la celda y erróneamente se unen a ellos, lo que genera un desbalance inicial en la composición por bases. Pero a medida que se siguen secuenciando, las hebras que sí quedaron bien pegadas son más estables y en mayor cantidad por lo que generan señales más fuertes que las erróneas, por lo que las terminan opacando y se muestran contenidos más estables.

Posteriormente, se llevaron a cabo cuatro *trimmings* diferentes que se dividieron entre los integrantes del equipo (20, 22, 26 y 28). Al observar los resultados de estos, se decidió realizar el corte en un *Phred score* de 28. Esta elección se fundamentó en el análisis de los reportes de FastQC y MultiQC de todos los integrantes, donde se observó que este valor ofrecía la mejor calidad. Aunque es común la caída de calidad en los extremos de las lecturas Illumina, el *boxplot* "*Per base sequence quality*" mostró que las medianas se mantienen en 30 o superiores, lo que confirmó que el *trimming* aplicado era adecuado. Además, al comparar con los reportes previos al corte, se evidenció que las medianas rondaban en el valor de 28, lo cual reforzó la decisión tomada.

Posteriormente, se propuso evaluar la opción de recortar los extremos 5' y 3' manteniendo el mismo umbral de 28, para compararla con los resultados obtenidos en el trimming ya realizado. Para sustentar esta evaluación se consultaron dos artículos y un repositorio, que sirvieron de referencia en la toma de decisiones. El análisis se centró en los reportes de Fastp, considerando principalmente los siguientes aspectos:

1. ***Reads* totales después del corte**: Al recortar en 5' y 3' se pierde cerca del 17% de los *reads*, mientras que con trimming en 28 la pérdida es solo del 10%.
2. ***Reads* de baja calidad después del corte**: Los valores se mantienen prácticamente iguales, con una diferencia mínima de 0,6%.
3. ***Reads* demasiado cortos después del corte**: El recorte en los extremos genera un 5% más de *reads* cortos (8% vs 3%), lo que incrementa secuencias repetitivas, *misassemblies*, *gaps* y cobertura anómala.

Luego de esto en los reportes FastQC se tomaron en cuenta las siguientes métricas:

1. **Calidad de secuencia por base:** El recorte en extremos deja medianas cercanas a 34, mientras que con *trimming* en 28 bajan a 30 (igualmente aceptable).
2. **Puntuaciones de calidad por secuencia:** La calidad promedio de las lecturas es similar en ambos casos; los picos son amplios y se concentran en valores ≥30.
3. **Secuencias sobrerrepresentadas:** No se detectaron secuencias sobrerrepresentadas de forma relevante en ninguno de los métodos.
4. **Niveles de duplicación de las secuencias:** Los dos presentan picos adecuados en 1 y 2; no hay diferencias significativas.

Al comparar los reportes de FastQC entre el recorte de extremos (5' y 3') y el *trimming* en 28, no se observan diferencias significativas. Sin embargo, considerando que el *trimming* en 28 conserva mejor la longitud de los reads sin comprometer la calidad se concluye que esta es la opción más adecuada para los procesos posteriores.

Una vez realizado el análisis de los datos usando las herramientas de control de calidad y *trimming* se continuo con el ensamblaje del genoma. En este proceso, los *reads* se agrupan primero en *contigs* (secuencias continuas sin huecos) y luego se organizan mediante *scaffolds*, que permiten ordenar y orientar esos *contigs* aun cuando existan regiones sin secuenciar. Teniendo en cuenta que el punto de partida del experimento es con una población ancestral de bacterias *E. Coli* las cuales se sometieron a condiciones de cultivo durante varias generaciones. Bajo estas condiciones y con el tiempo surgen mutaciones que pueden otorgar ventajas adaptativas. Por lo tanto, para poder comparar e identificar los cambios genómicos responsables de esas adaptaciones es necesario tener un genoma de referencia el cual se obtiene al ensamblar los *reads* de la línea ancestral con los cuales se tiene una aproximación del genoma al inicio del experimento antes de la aparición de mutaciones adaptativas. De esta forma la comparación posterior al mapear los reads de las poblaciones evolucionadas permite identificar con mayor precisión los cambios genómicos responsables de las adaptaciones observadas.

Resultados de Quast.

Quast permite determinar la calidad de los ensamblajes que se hicieron de los datos crudos previos al *trimming* y los datos posteriores, así entonces es posible comprender cómo se unieron y las características de los *contigs* en los genomas construidos. Para esto Quast cuenta con métricas como la longitud total del ensamblaje, que se busca que sea lo más larga posible; *contigs* más largo y número de *contigs*, que permiten ver la composición del ensamblaje.

Métricas:

N50 y N90 muestran la longitud del *contig* más corto en el grupo de *contigs* que forman el 50% y 90%. En otras palabras, la longitud del *contig* más corto de los que sumados llegan al 50% y 90% del total del ensamblaje. Por otro lado, L50 y L90 muestran el número de *contigs* necesarios para formar el 50% y 90% del genoma.

También está la métrica auN la cual representa el área bajo la curva de la gráfica Nx y fue propuesta por Heng Li al encontrar problemas con N50. N50 representa solo un punto en esa gráfica y si se conectan 2 *contigs* más largos que N50 o 2 *contigs* más cortos que N50 su valor no cambiara y solo mejoraría si se conecta un *contig* más corto que N50 y uno más largo que N50. Por lo tanto, si el análisis se centra únicamente en N50 hay posibilidad de una mala interpretación. De esta forma auN otorga una sensibilidad a mejoras pequeñas y se incrementa cada vez que 2 *contigs* se conectan, además es menos propenso a saltos abruptos ya que integra toda la distribución de tamaños y no dependen sólo de un punto específico.

Hay métricas que se podrían considerar esperables para un buen ensamblaje como alto N50, bajo L50 y bajos *misassemblies*. Sin embargo, La calidad de un ensamblaje no se define únicamente por una métrica de longitud (como N50, que solo mide contigüidad), sino por la síntesis de al menos dos criterios técnicos: Contigüidad y Precisión Estructural. Un ensamblaje superior es aquel que es largo y, al mismo tiempo, fiel a la arquitectura genómica real. Ignorar la precisión estructural cuantificada por QUAST (es decir, tolerar un alto número de *misassemblies* y errores base a base) lleva a un ensamblaje que, aunque sea largo, es estructuralmente erróneo (Falsa Contigüidad). Por lo tanto, un bajo número de *misassemblies* de QUAST es esencial para validar la calidad reportada por la N50. Por último, la decisión de mejores métricas dependerá de los investigadores y de los objetivos que persiguen.

Statistics without reference	PRE	POST
# contigs	192	206
# contigs (>= 0 bp)	264	332
# contigs (>= 1000 bp)	173	188
# contigs (>= 5000 bp)	129	138
# contigs (>= 10000 bp)	104	108
# contigs (>= 25000 bp)	64	66
# contigs (>= 50000 bp)	30	29
Largest contig	167 767	136 463
Total length	4 530 485	4 529 272
Total length (>= 0 bp)	4 546 456	4 551 231
Total length (>= 1000 bp)	4 516 959	4 516 449
Total length (>= 5000 bp)	4 410 954	4 398 833
Total length (>= 10000 bp)	4 224 688	4 178 909
Total length (>= 25000 bp)	3 570 348	3 500 304
Total length (>= 50000 bp)	2 411 175	2 276 309
N50	57 404	50 016
N90	13 799	12 670
auN	59 020	55 201
L50	28	29
L90	92	99
GC (%)	50.74	50.74
Per base quality		
# N's per 100 kbp	22.73	22.74
# N's	1030	1030

Figura 1. Reporte de Quast de los datos crudos vs *trimmed data*

Una vez observado el reporte de Quast se observan los resultados de las métricas anteriormente mencionadas. En el caso de N50 y L50 se observa que L50 es mayor para los datos a los que se les realizo el *trimming* lo que significa que para cubrir ese mismo porcentaje se necesitaron más *contigs* indicando que el ensamblaje se fragmento un poco más. Se puede llegar a la misma conclusión con las métricas N90 y L90. Para el caso de auN también disminuye para *trimmed data* lo que indica que se pierde un poco de continuidad. Por otro lado, el número de huecos (N's) es la misma en las dos lo que no ayuda a saber con seguridad si el ensamblaje es más limpio al eliminar registros dudosos. Por último, el contenido de CG (%) es prácticamente idéntico lo cual sugiere que el *trimming* no introdujo sesgo fuerte en la composición de bases.

Se concluye que el ensamblaje *trimmed* es con el cual se continuara trabajando ya que, aunque tiene una ligera perdida de continuidad y el número de huecos no cambio, se tiene la seguridad de que es realizado a partir de lecturas con una calidad *Phred* de 28 y como la perdida de continuidad no es tan se valoró más la calidad en este caso.

Por último, se realizó un alineamiento de los reads evolucionados frente al genoma de referencia que en este caso es el ensamblaje en el paso anterior. Las lecturas evolucionadas se alinean con la finalidad de verificar que el ensamblaje sea bueno, medir su calidad y, sobre todo, para detectar variantes y cambios evolutivos en las poblaciones que se comparan. Para esta alineación se usó BWA donde se requiere que el genoma ensamblado de referencia sea indexado. Después de la alineación, se convierten los archivos en formato SAM a BAM y con ayuda de *samtools* se aplica un ordenamiento, eliminación de duplicados y en general se limpian los BAM finales para análisis posteriores. Para terminar, se usa QualiMap para observar la calidad del ensamblaje proporcionando información sobre el porcentaje de lecturas alineadas y la cobertura genómica. Los resultados consolidados se visualizaron nuevamente con MultiQC, permitiendo interpretar globalmente la calidad de los ensamblajes y las alineaciones.

Observando el reporte de MultiQC se observó que en el evolucionado 1 se mapeo el 99.92% y el evolucionado 2 se mapeo el 97.8%. Son porcentajes bastante buenos con respecto al mapeo y la sección que no se mapeo puede deberse a errores de secuenciación, contaminación o mutaciones importantes. Además, se analizó el reporte individual de cada evolucionado alineado con el ensamblaje de referencia, donde el *secondary aligments* dio 0 lo que indica que no hubo lecturas que se mapearan en más de una posición.

Mismatches and indels	
General error rate	0.84%
Mismatches	1,797,272
Insertions	1,657
Mapped reads with at least one insertion	0.1%
Deletions	5,315
Mapped reads with at least one deletion	0.34%
Homopolymer indels	38.87%

Mismatches and indels	
General error rate	0.89%
Mismatches	1,792,631
Insertions	1,594
Mapped reads with at least one insertion	0.11%
Deletions	4,929
Mapped reads with at least one deletion	0.34%
Homopolymer indels	39.75%

En este apartado se observan más resultados con respecto al evolucionado 1 al lado izquierdo y el evolucionado 2 al lado derecho en los dos se observa que el *general error rate* es menor que el 1% lo que significa que menos de ese porcentaje de las bases presentan errores en comparación con la referencia. El *mismatches* se observa el número total donde la secuencia difiere de la referencia. Las inserciones y deleciones muestran cuantas bases adicionales y cuantas bases faltan en las lecturas alineadas y el valor no es muy alto teniendo en cuenta el tamaño de referencia que es de 4.551.231. En general se concluye que es un buen alineamiento según los resultados de la calidad al alienar casi la totalidad de las lecturas con el ensamblaje de referencia y que esos porcentajes no mapeados seguirán siendo estudiados para comprobar posibles cambios evolutivos en esas lecturas.

¿Qué se ve en IGV?: Al cargar en IGV el archivo de referencia (*scaffolds* en formato FASTA) junto con los archivos BAM y sus índices (.bai), lo que se visualiza es la alineación de las lecturas contra la secuencia de referencia. En la parte superior aparece el gráfico de cobertura, que muestra cuántas lecturas apoyan cada posición del genoma, y en la parte inferior se despliegan las lecturas individuales alineadas, donde se pueden distinguir coincidencias con la referencia y también diferencias marcadas con colores que representan nucleótidos alternativos. Esta visualización permite explorar de manera interactiva la calidad del mapeo, la profundidad de secuenciación y posibles variantes a lo largo del genoma ensamblado.

¿Porque se debe indexar?: Indexar un archivo BAM consiste en generar un archivo adicional (.bai) que funciona como una especie de tabla de contenido, de manera que programas como IGV puedan ubicar rápidamente las regiones del genoma que uno quiere visualizar sin tener que leer todo el archivo de principio a fin. Para que esto sea posible, el BAM debe estar ordenado por coordenadas genómicas, lo cual permite que el índice relacione esas coordenadas con bloques de datos contiguos dentro del archivo. Cuando se abre el BAM en IGV, el programa busca automáticamente el archivo de índice (.bai) y lo usa para cargar solo la parte de la secuencia que el usuario solicita, en lugar de procesar todo el archivo, lo que hace que la exploración sea mucho más rápida y práctica, incluso si se trata de archivos grandes o almacenados de forma remota. Sin el índice, IGV tendría que revisar el archivo completo cada vez, lo que sería muy poco eficiente; por eso, la indexación asegura una visualización ágil, eficiente y compatible.

Referencias

Assess Quality of Assemblies with QUAST - v4.4 | KBase App. (2016). Kbase.us. https://kbase.us/applist/apps/kb_quast/run_QUAST_app/release

BioinfQuests. (2021, June 1). *Different Assembly statistics (N50, L50, NG50, LG50, NA50, NGA50 and Misassemblies)*. YouTube. https://www.youtube.com/watch?v=ViXzKrQo25k

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Hesketh, S. (2024, marzo 4). A guide to reviewing BAM files. *EPI2ME Blog*. https://epi2me.nanoporetech.com/reviewing-bam/

Jia, H., Tan, S., & Zhang, Y. E. (2024). Chasing Sequencing Perfection: Marching Toward Higher Accuracy and Lower Costs. *Genomics, Proteomics & Bioinformatics/Genomics, Proteomics and Bioinformatics*. https://doi.org/10.1093/gpbjnl/qzae024

Laboratorio Evolucion Molecular. (2021, April 10). *Práctica 2 Métricas de Ensamble*. YouTube. https://www.youtube.com/watch?v=Vw5n4kj1LVQ

Narzisi, G., & Mishra, B. (2011). Comparing De Novo Genome Assembly: The Long and Short of It. *PLoS ONE*, *6*(4), e19175. https://doi.org/10.1371/journal.pone.0019175

Nute, M. (2022, July 12). *quality_control_tutorial*. https://gitlab.com/treangenlab/quality_control_tutorial

Per Base Sequence Content. (2021). Babraham.ac.uk. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html

QC Fail Sequencing» Positional sequence bias in random primed libraries. (2017, February 6). QC Fail. https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/

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 (PMCID: PMC3603213)