ACTA BIOLÓGICA COLOMBIANA
http://www.revistas.unal.edu.co/index.php/actabiol

Facultad de Ciencias
Departamento de Biología
Sede Bogotá

UNIVERSIDAD NACIONAL DE COLOMBIA

ARTÍCULO DE INVESTIGACIÓN / RESEARCH ARTICLE

MOLECULAR

EFECTO DEL FILTRADO DE SECUENCIAS EN EL ENSAMBLADO DEL GENOMA DE *Bacillus altitudinis* 19RS3 AISLADO DE *Ilex paraguariensis*

Effect of sequence filtering on the assembly of the *Bacillus altitudinis* 19RS3 genome isolated from *Ilex paraguariensis*

Iliana Julieta CORTESE[1], María Lorena CASTRILLO[1], Pedro Darío ZAPATA[1], Margarita Ester LACZESKI[1,2]*

[1]Laboratorio de Biotecnología Molecular, Instituto de Biotecnología Misiones "Dra. María Ebe Reca" (InBioMis), CONICET, Facultad de Ciencias Exactas, Químicas y Naturales/FCEQyN, Universidad Nacional de Misiones/UNaM, Ruta 12 km 7 ½, Posadas, Misiones, Argentina.

[2]Cátedra de Bacteriología, Dpto. de Microbiología, Facultad de Ciencias Exactas, Químicas y Naturales/FCEQyN, Universidad Nacional de Misiones/UNaM, Avenida Mariano Moreno 1375, Posadas, Misiones, Argentina.

***For correspondence:** melaczeski@fceqyn.unam.edu.ar

Received: 16th April 2020, **Returned for revision**: 16th July 2020, **Accepted**: 18th August 2020.
Associate Editor: Oscar Ruiz.

Citation/Citar este artículo como: Cortese IJ, Castrillo ML, Zapata PD, Laczeski ME. Efecto del filtrado de secuencias en el ensamblado del genoma de *Bacillus altitudinis* 19RS3 aislado de Ilex paraguariensis. Acta Biol Colomb. 2021;26(2):170-177. Doi: http://dx.doi.org/10.15446/abc.v26n2.86406

RESUMEN

El filtrado de secuencias es un paso esencial sin importar el tipo de tecnología aplicada para la secuenciación de un genoma, en el cual las lecturas de baja calidad o una parte son eliminadas. En un ensamblado la construcción de un genoma se realiza a partir de la unión de lecturas cortas en cóntigos. Algunos ensambladores miden la relación que existe entre secuencias de una longitud fija (k-mer) que puede verse afectada por la presencia de secuencias de baja calidad. Un enfoque común para evaluar los ensamblados se basa en el análisis del número de cóntigos, la longitud del cóntigo más largo y el valor del N50, definido como la longitud del cóntigo que representa el 50 % de la longitud del conjunto. En este contexto, el presente estudio tuvo como objetivo evaluar el efecto del uso de lecturas crudas y filtradas en los valores de los parámetros de calidad obtenidos en el ensamblado del genoma de *Bacillus altitudinis* 19RS3 aislada de *Ilex paraguariensis*. Se realizó el análisis de calidad de ambos archivos de partida con el *software* FastqC y se filtraron las lecturas con el *software* Trimmomatic. Para el ensamblado se utilizó el *software* SPAdes y para su evaluación la herramienta QUAST. El mejor ensamblado para *B. altitudinis* 19RS3 se obtuvo a partir de las lecturas filtradas con el valor de k-mer 79, que generó 16 cóntigos mayores a 500 pb con un N50 de 931 914 pb y el cóntigo más largo de 966 271 pb.

Palabras clave: análisis de secuencias, biología computacional, control de calidad, genomas bacterianos.

ABSTRACT

Sequence filtering is an essential step regardless of the type of technology applied for sequencing a genome, in which low-quality readings or a portion are eliminated. In an assembly, the construction of a genome is carried out from the union of short reads in contigs. Some assemblers measure the relationship between sequences of a fixed length (k-mer) that can be affected by the presence of low-quality sequences. A common approach to evaluating assemblies is based on the analysis of the number of contigs, the length of the longest contig, and the value of N50 defined as the length of the contig representing 50 % of the length of the assembly. In this context, the objective of this study was to evaluate the effect of the use of crude and filtered reads on the values of the quality parameters obtained from the genome assembly of *Bacillus altituidinis* 19RS3 isolated from *Ilex paraguariensis*. The quality analysis of both starting files was performed with the FastqC software and the readings were filtered with the Trimmomatic software. The SPAdes software was used for the assembly and the QUAST tool for its evaluation. The best assembly for *B. altitudinis* 19RS3 was obtained from the filtered readings with the value of k-mer 79, which generated 16 contigs greater than 500 bp with a N50 of 931 914 bp and the longest contig of 966 271 bp.

Keywords: bacterial genome, computational biology, quality control, sequence analysis.

INTRODUCCIÓN

Los avances exponenciales en las tecnologías para generar y procesar grandes conjuntos de datos biológicos, datos -ómicos, promueven un cambio de paradigma en la forma de abordar los problemas biológicos. Las tecnologías -ómicas están dirigidas principalmente a la detección universal de genes (genómica), ARNm (transcriptómica), proteínas (proteómica) y metabolitos (metabolómica) en una muestra biológica específica (Manzoni *et al.*, 2016). Para procesar y analizar el enorme volumen de datos biológicos acumulados, es necesario el empleo de herramientas bioinformáticas que permitan manejar eficientemente la información obtenida por tecnologías de secuenciación de próxima generación (del inglés *Next Generation Sequencing*, NGS) (Aguilar-Bultet y Falquet, 2015).

En un proyecto de secuenciación de genoma, el ADN del organismo objetivo se divide en millones de piezas pequeñas y se lee en un equipo de secuenciación. Estas lecturas varían de 20 a 1000 pares de bases de nucleótidos (pb) en longitud, según el método de secuenciación utilizado. La secuenciación puede ser de lectura única (*single-end*) ó de lectura pareada (*paired-end*). Estas secuencias nucleotídicas se almacenan comúnmente junto con sus valores de calidad (*Phred Quality Score* ó Q) en un formato de texto plano conocido como fastq (Gladman, 2019). Se consideran secuencias de buena calidad aquellas que presentan un valor mayor o igual a 30, por lo que este parámetro se define como Q30.

Sin importar el tipo de tecnología aplicada para la generación de datos genómicos NGS, antes de utilizar las secuencias crudas se recomienda llevar a cabo una evaluación de calidad de cada grupo de datos. Es sumamente importante asegurarse que las lecturas se encuentran libres de adaptadores, secuencias contaminantes, artefactos de baja calidad o secuencias duplicadas, ya que de otra manera podrían interferir en los futuros análisis y procesos (Góngora-Castillo y Buell, 2013). Por lo general, la calidad de las lecturas se evalúa inicialmente mediante la observación de la distribución de bases a lo largo de la lectura, la distribución de calidad Phred, frecuencias de nucleótidos y complejidad de las lecturas (Bishop, 2014; Aguilar-Bultet y Falquet, 2015). Los resultados de esta evaluación pueden utilizarse a la hora de elegir las opciones de filtrado de las secuencias, o bien, verificar si por el contrario cumplen con los estándares de calidad (Andrews, 2010).

El filtrado de secuencias tiene como objetivo eliminar sólo las regiones de baja calidad en un procedimiento conocido como recorte. Este paso no es un proceso trivial y se aborda de diferentes maneras por diversos algoritmos, parámetros y herramientas (Cox *et al.*, 2010; Martin, 2011; Bolger *et al.*, 2014). El principio básico del recorte de lecturas es hacer una estimación de las tasas de error de lectura con el objeto de mantener la información de mayor calidad contenida en las lecturas, y así descartar aquellas de baja calidad o parte de las mismas (Del Fabbro, 2013; Bishop, 2014; Aguilar-Bultet y Falquet, 2015).

El filtrado de secuencias se ha adoptado ampliamente en los estudios más recientes de NGS (Schmieder y Edwards, 2011a). Sin embargo, a pesar de su popularidad, la información que se tiene al respecto sobre su aplicación, su efecto sobre los parámetros de calidad y de ensamblado, así como para la selección del *software* apropiado, continúa siendo escasa. Varios métodos se han descrito individualmente en la literatura (Cox *et al.*, 2010; Smeds y Künstner, 2011; Schmieder y Edwards 2011b; Bolger *et al.*, 2014; Chen *et al.*, 2014) pero su utilidad se ha demostrado solo en casos particulares de ensamblados de genomas (Del Fabbro *et al.*, 2013). La selección de un *software* de filtrado adecuado dependerá del tipo de información obtenida de la secuenciación, es decir, si la misma fue única o pareada. Existe una gran cantidad de *software* disponibles para el filtrado de lecturas, tales como Trimmomatic (Bolger *et al.*, 2014), NGS QC Toolkit (Patel y Hain, 2012), PRINSEQ (Schmieder y Edwards 2011a) y CutAdapt (Martin, 2011).

Particularmente, en un ensamblado en el que la construcción de un genoma se realiza a partir de la unión de lecturas cortas en fragmentos más largos, conocidos como cóntigos, la utilización de lecturas filtradas adquiere suma importancia para la obtención de mejores resultados. El filtrado eliminará aquellas lecturas propensas a errores y proporcionará una mejor guía para configurar los parámetros de entrada apropiados para el ensamblador a utilizar (Bolger *et al.*, 2014). En los últimos años, el número de *software* desarrollados para el ensamblado de secuencias cortas ha aumentado considerablemente. SPAdes (Bankevich *et al.*, 2012) es un ensamblador de acceso libre basado en la construcción de grafos de Brujin (Medvedev *et al.,* 2011). Este mide la relación que existe entre secuencias de nucleótidos de una longitud fija (k-mer) creada, y genera un grafo donde los nodos son los k-mers y las conexiones del grafo indican que los k-mers son adyacentes y se solapan (k-1 nucleótidos). El k-mer también puede verse afectado por la presencia de secuencias de baja calidad, lo que aumenta la probabilidad de errores en los ensamblados.

Un enfoque habitual para evaluar los ensamblados obtenidos se basa en el análisis del número de cóntigos mayores a 500 pb, la longitud del cóntigo más largo y el valor de N50. El valor N50 se define como la longitud del cóntigo más grande de todos los cóntigos, clasificados de menor a mayor, que representa el 50 % de la longitud del conjunto. Cuanto menor es el número de cóntigos y mayor es el valor del N50, se puede decir que el ensamblado es mejor.

En estudios previos, se aisló una cepa de *Bacillus altitudinis* codificada como 19RS3, de plantines de *Ilex paraguariensis* St. Hilaire 1822 (Aquifoliaceae), con propiedades de promoción del crecimiento vegetal (Laczeski *et al.*, 2020). Con el fin de adquirir nuevos conocimientos que aporten a la comprensión de los mecanismos biológicos utilizados por esta bacteria, se secuenció su genoma completo.

En este contexto, se puso a prueba la hipótesis de que el filtrado de lecturas mejora los parámetros de calidad de los

ensamblados generados para un genoma procariota, por lo que el objetivo de esta investigación fue evaluar el efecto del uso de lecturas pareadas crudas y lecturas pareadas filtradas, como archivos de partida, en el valor de k-mer, el número de cóntigos, la longitud del cóntigo más largo y el valor de N50 obtenidos en el ensamblado del genoma de la cepa de *B. altitudinis* 19RS3.

MATERIALES Y MÉTODOS

Datos genómicos

Los datos genómicos utilizados se obtuvieron como producto de la secuenciación del genoma completo de *B. altitudinis* 19RS3, una bacteria endófita recuperada de *I. paraguariensis* aislada en la provincia de Misiones, Argentina. Esta cepa fue seleccionada por presentar propiedades de promoción del crecimiento vegetal en ensayos realizados *in vitro* y en vivero con plantines orgánicos de *Ilex paraguariensis* (Laczeski *et al.*, 2020). La misma se encuentra depositada bajo el número de acceso LBM250 en el cepario del Laboratorio de Biotecnología Molecular del Instituto de Biotecnología Misiones "Dra. María Ebe Reca" (InBioMis) perteneciente a la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones.

El ADN genómico se extrajo a partir del protocolo de Sambrook modificado (Sambrook, 2001; Cariaga Martinez y Zapata, 2007). La construcción de la librería *TruSeq Nano DNA* (350) y la secuenciación *whole genome de novo paired-end* se llevó a cabo por Macrogen (Seoul, Korea) con la tecnología Illumina HiSeq.

Análisis de calidad y filtrado de lecturas

Se realizó el análisis de calidad de las lecturas pareadas crudas obtenidas de la secuenciación y de las lecturas pareadas filtradas. Para ello se utilizó el *software* FastqC (Andrews, 2010) versión 0.11.9. Se evaluaron los principales parámetros de interés: cantidad de lecturas, longitud promedio de las lecturas, rango de longitud de las lecturas y porcentaje de bases nucleotídicas de guanina y citocina (GC). La calidad de cada grupo de lecturas se determinó a partir del análisis de los distintos módulos de FastqC (Del Fabbro *et al.*, 2013; Rana *et al.*, 2016; Rodriguez Hernáez, 2017).

De acuerdo con lo evaluado y con la finalidad de obtener lecturas de mejor calidad, se realizó el filtrado de éstas con el *software* Trimmomatic (Bolger *et al.*, 2014) versión 0.39. Se seleccionaron las siguientes opciones: eliminación de adaptadores (ILLUMINACLIP:TrueSeq3-PE.fa:2:30:10), eliminación de las principales bases de baja calidad o N bases (LEADING:3), eliminación de las bases de baja calidad o N finales (TRAILING:3), escaneo de lecturas desde el extremo 5´ con el método de "ventana deslizante" (SLIDINGWINDOW:4:15) y eliminación de lecturas con menos de 36 bases de largo (MINLEN: 36) (Del Fabbro *et al.*, 2013; Rana *et al.*, 2016; Rodriguez Hernáez, 2017).

Ensamblado del genoma

Se realizó el ensamblado del genoma de *B. altitudinis* 19RS3 con el *software* SPAdes (Bankevich *et al.*, 2012) versión 3.12.0. Se evaluó la aplicación de valores de k-mer impares entre 55 a 87. Este *software* establece como requisito utilizar valores de k-mer impares y menores a 128. Como archivos de partida se utilizaron las lecturas pareadas crudas, y las lecturas pareadas obtenidas luego del filtrado con el *software* Trimmomatic.

Para la evaluación de los ensamblados se utilizó la herramienta *Quality Assessment Tool for Genome Assemblies*-QUAST (Gurevich *et al.*, 2013) versión 4.0.

RESULTADOS

En la secuenciación del genoma de *B. altitudinis* 19RS3 se generó un total de 9 938 250 lecturas pareadas de 101 pb con una cobertura promedio de 266. El contenido GC fue 41,01 % y el valor de calidad Q30 fue 91,57 %.

Las lecturas crudas obtenidas presentaron valores altos de calidad según el análisis realizado con el *software* FastqC. La Fig. 1 presenta en la parte superior los parámetros básicos de calidad y en la parte inferior los gráficos de calidad por base para las secuencias sentido y anti-sentido. Estos gráficos representan cada una de las bases de una secuencia *versus* sus valores de calidad Q30. Dichos valores dividen el gráfico horizontalmente en una franja verde con valores de calidad altos (Q30 mayor o igual a 30), una franja amarilla con valores de calidad intermedios (Q30 menor a 30 y mayor a 20) y una franja roja con valores de calidad bajos (Q30 menor a 20). Las bases de las lecturas sentido-crudas mostraron una distribución completa sobre la franja verde. Sin embargo, las lecturas anti-sentido mostraron una distribución menos homogénea a medida que se aumenta la posición en pb de la secuencia.

Luego del filtrado, y a pesar de que el número total de lecturas disminuyó, el 93,87 % del genoma secuenciado continuó representado. Debido al recorte se obtuvieron secuencias con un tamaño entre 36 y 101 pb. Las lecturas sentido-filtradas mantuvieron la misma calidad que las lecturas sentido-crudas en todas sus bases. Sin embargo, las lecturas anti-sentido filtradas mostraron una gran mejoría en la calidad de sus últimas bases con respecto a lo observado en las lecturas anti-sentido crudas. El porcentaje de GC no se vio afectado por el filtrado (Fig. 2).

A partir de la evaluación realizada con el *software* QUAST se obtuvieron dos tablas correspondientes a los ensamblados realizados a partir de las lecturas crudas (Tabla 1) y para los ensamblados realizados a partir de las lecturas filtradas (Tabla 2). En ambas tablas se resaltan en letra negrita los valores principales que se tuvieron en cuenta para realizar la comparación. Al utilizar valores crecientes de k-mer de 55 a 87, se logró disminuir el número de cóntigos

a

Parámetro	Valor
Archivo	19R_1.fastq.gz
Tipo de archivo	Bases convencionales
Total de secuencias	4969125
Secuencias de baja calidad	0
Longitud de la secuencia	101
% GC	41

b

Parámetro	Valor
Archivo	19R_2.fastq.gz
Tipo de archivo	Bases convencionales
Total de secuencias	4969125
Secuencias de baja calidad	0
Longitud de la secuencia	101
% GC	40

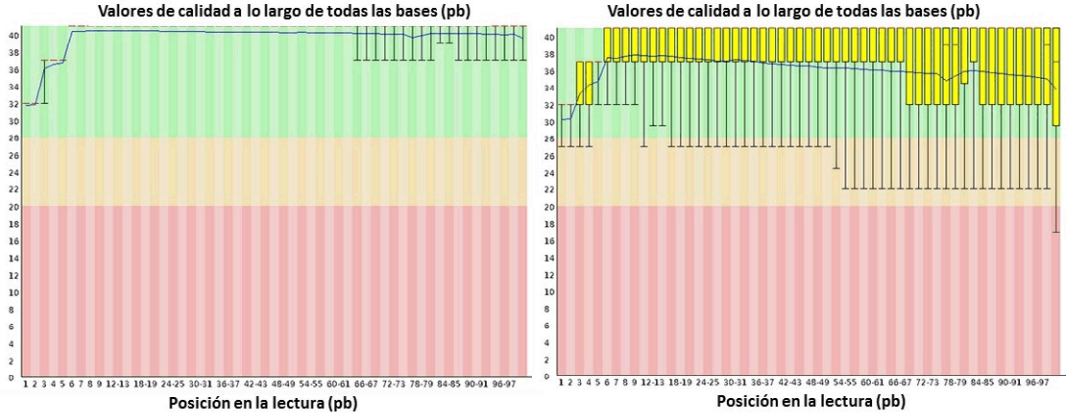

Figura 1. Análisis de calidad realizado por el *software* FastQC para las lecturas crudas del genoma de *Bacillus altitudinis* 19RS3 (Bacteria promotora del crecimiento vegetal aislada de *Ilex paraguariensis* St. Hil.). En la parte superior se presentan los parámetros básicos de calidad. En la parte inferior se observan los gráficos de calidad por base para las secuencias sentido (a) y anti-sentido (b). El eje X corresponde a la posición de cada base en una secuencia y el eje Y al valor de calidad Q30. Los colores representan la calidad según los valores de Q30, alta en verde (Q30 mayor o igual a 30), intermedia en amarillo (Q30 menor a 30 mayor a 20) y baja en rojo (Q30 menor a 20).

y aumentar paralelamente el valor del N50. Si se observa el k-mer 79 en ambas tablas, valor con el cual se obtuvo el mejor ensamblado con las lecturas filtradas, se puede notar que existe una diferencia de cinco cóntigos, 157 763 pb en el valor de N50 y de 1464 pb en la longitud del cóntigo más largo, entre ambos ensamblados.

Si bien las diferencias entre resultados fueron notorias en todos los ensamblados, cabe destacar que el mejor ensamblado para *B. altitudinis* 19RS3 se obtuvo a partir de las lecturas filtradas con el valor de k-mer 79, con el que se generaron 16 cóntigos mayores a 500 pb con un N50 de 931 914 pb y el cóntigo más largo con una longitud de 1464 pb mayor al obtenido en el ensamblado realizado con las lecturas crudas (Tabla 2). Aunque se observaron valores mayores de N50 al utilizar un k-mer 83 en el ensamblado con lecturas filtradas, la diferencia entre ambos valores fue mínima. Es por esto que se decidió seleccionar como mejor ensamblado aquel que presentó un menor número de cóntigos.

DISCUSIÓN

Si bien el filtrado de secuencias es un paso esencial en el procesamiento de lecturas crudas, generalmente se omite en la bibliografía a la hora de realizar el ensamblado de un genoma. Esta falta de información y la consecuente ausencia de este proceso puede derivar en interpretaciones biológicas erróneas de los resultados obtenidos. La evaluación del efecto del filtrado de secuencias en los parámetros de calidad de los ensamblados obtenidos para *B. altitudinis* 19RS3 indicó que a medida que el valor de k-mer aumentó los ensamblados mostraron grandes mejoras en los demás parámetros evaluados, como el número de cóntigos y el valor de N50. Al utilizar las lecturas filtradas se obtuvo un mejor ensamblado con el valor de k-mer 79; sin embargo, al utilizar las lecturas crudas se observó un mejor ensamblado con el valor de k-mer 81. Este valor mayor para las lecturas crudas podría deberse a que como indicaron Del Fabbro *et al.* (2013), la inclusión de lecturas de baja calidad conduce a la generación de falsos k-mer (Zerbino y Birney, 2008), lo cual afecta la conectividad entre los nodos y tiene un impacto significativo en la longitud y el número final de cóntigos (Smeds y Künstner, 2011; Del Fabbro *et al.*, 2013). A diferencia de los ensamblados que obtuvimos a partir de las lecturas filtradas, cuyo número de cóntigos disminuyó notablemente, Del Fabbro *et al.* (2013) generaron ensamblados más fragmentados con un mayor número de cóntigos a partir de las lecturas filtradas. Cabe destacar que su trabajo se realizó con genomas de mayor complejidad que el nuestro, factor que podría estar vinculado con una menor eficiencia del filtrado de las lecturas.

a

Parámetro	Valor
Archivo	19R_1_trim_PE.fastq.gz
Tipo de archivo	Bases convencionales
Total de secuencias	4664919
Secuencias de baja calidad	0
Longitud de la secuencia	36-101
% GC	41

b

Parámetro	Valor
Archivo	19R_2_trim_PE.fastq.gz
Tipo de archivo	Bases convencionales
Total de secuencias	4664919
Secuencias de baja calidad	0
Longitud de la secuencia	36-101
% GC	41

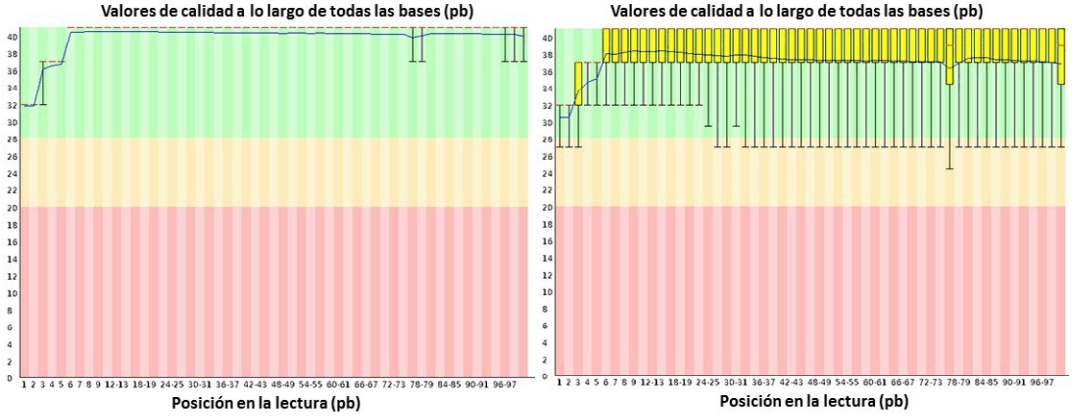

Figura 2. Análisis de calidad realizado por el *software* FastQC para las lecturas filtradas del genoma de *Bacillus altitudinis* 19RS3 (Bacteria promotora del crecimiento vegetal aislada de *Ilex paraguariensis* St. Hil.). En la parte superior se presentan los parámetros básicos de calidad. En la parte inferior se observan los gráficos de calidad por base para las secuencias sentido (a) y anti-sentido (b). El eje X corresponde a la posición de cada base en una secuencia y el eje Y al valor de calidad Q30. Los colores representan la calidad según los valores de Q30, alta en verde (Q30 mayor o igual a 30), intermedia en amarillo (Q30 menor a 30 mayor a 20) y baja en rojo (Q30 menor a 20).

De manera similar al presente trabajo, los creadores del *software* Trimmomatic (Bolger *et al.*, 2014) realizaron la evaluación de los ensamblados obtenidos a partir de lecturas crudas y filtradas. Como resultado, obtuvieron mejoras hasta de un 77 % y un 55 % en los valores de N50 y en la longitud del cóntigo más largo, respectivamente, a partir de las lecturas filtradas. Además, notaron que el ensamblado obtenido a partir de las lecturas crudas contenía una coincidencia perfecta de 34 pb con una secuencia de adaptador. Estos datos dan un mayor soporte a los resultados obtenidos por nuestro grupo y destacan la necesidad de trabajar con lecturas procesadas con el fin de asegurar la eliminación de secuencias de adaptadores que podrían incorporarse erróneamente en el ensamblado final.

Chen *et al.* (2014) también compararon los efectos del uso de lecturas crudas y filtradas para la generación de ensamblados. Al igual que en nuestro trabajo, sus resultados mostraron que en todos los casos el filtrado mejoró las puntuaciones de calidad de las lecturas. Coincidimos con Chen *et al.* (2014) al decir que Trimmomatic presenta una amplia variedad de opciones para la realización del filtrado de las lecturas y además recomendamos su uso por su fácil manejo y accesibilidad.

En concordancia con los resultados generados por Chen *et al.* (2014) y Bolger *et al.* (2014), los ensamblados obtenidos en nuestro trabajo para el genoma de *B. altitudinis* 19RS3 sustentan la necesidad del filtrado de las secuencias como paso previo a realizar el ensamblado de un genoma en estudio. Además, como lo mencionan Del Fabbro *et al.* (2013), consideramos que existe poca literatura disponible que trate sobre el manejo y procesamiento de lecturas cortas producidas durante un experimento NGS, y que la interpretación biológica de los datos puede verse influenciada por la falta de su aplicación.

CONCLUSIONES

Más allá del ensamblador que se seleccione, o de la estrategia k-mer que se utilice, en el presente trabajo remarcamos la importancia de realizar el procesamiento de datos genómicos a partir del filtrado y recorte de lecturas, con el fin de optimizar las etapas de análisis posteriores. Destacamos que este procedimiento, además de ser rápido y fácil de realizar, permite la obtención de ensamblados de mejor calidad para genomas bacterianos, con un mayor valor de k-mer, un menor número de cóntigos, una mayor longitud del cóntigo más largo y un mayor valor del N50, lo que asegura la generación de resultados fehacientes y representativos. Resaltamos, además, la utilidad del *software* Trimmomatic a la hora de realizar el procesamiento de

Tabla 1. Ensamblado del genoma de *Bacillus altitudinis* 19RS3 (Bacteria promotora del crecimiento vegetal aislada de *Ilex paraguariensis* St. Hil.). Resultados obtenidos de la evaluación por QUAST de los ensamblados realizados con SPAdes a partir de lecturas crudas.

	# cóntigos	Longitud del cóntigo más largo	Longitud total	GC (%)	N50
k-mer 55	28	778 854	3 783 072	41,16	553 021
k-mer 61	26	830 193	3 783 478	41,16	774 385
k-mer 65	23	932 555	3 781 549	41,16	774 285
k-mer 67	20	932 563	3 781 839	41,16	774 289
k-mer 69	20	932 571	3 781 909	41,16	774 293
k-mer 71	20	932 579	3 782 087	41,16	774 405
k-mer 73	21	932 719	3 783 779	41,16	774 409
k-mer 75	22	932 805	3 784 068	41,17	774 143
k-mer 77	22	932 986	3 785 125	41,17	774 147
k-mer 79	21	933 937	3 785 553	41,17	774 151
k-mer 81	**18**	**964 807**	3 786 254	41,17	**928 480**
k-mer 83	20	964 870	3 788 948	41,17	894 982
k-mer 85	20	964 876	3 789 114	41,18	894 986
k-mer 87	20	964 882	3 788 415	41,17	894 990

QUAST: *software* utilizado para la evaluación de los ensamblados.
SPAdes: *software* utilizado para realizar los ensamblados.
K-mer: valor k utilizado para cada ensamblado realizado con el *software* SPAdes.
cóntigos: número total de cóntigos de longitud ≥ 500 pb.
Longitud total: número total de bases en los cóntigos de longitud ≥ 500 pb.
GC (%): porcentaje de guanina y citosina.
N50: longitud del cóntigo más grande de todos los cóntigos clasificados de menor a mayor que representa el 50 % de la longitud del conjunto.
Las cifras destacadas en letra en negrita corresponden a los valores obtenidos para el mejor ensamblado generado con las lecturas crudas.

secuencias pareadas y asegurar la eliminación de secuencias de baja calidad y adaptadores.

AGRADECIMIENTOS

Este trabajo fue financiado por el proyecto del Instituto Nacional de la Yerba Mate (INYM, Argentina) "Biofertilizantes: validación a campo y estudios de trazabilidad de la utilización de *Bacillus* sp. como fertilizante para yerba mate" Res. N° 274/17 (INYM - PRASY).

REFERENCIAS

Aguilar-Bultet L, Falquet L. Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. Rev Salud Anim. 2015; 37(2):125-132.

Andrews S. FastQC a quality control tool for high throughput sequence data. Babraham bioinformatics [monografía en Internet] 2010. Disponible en: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Citado: 30 abr 2020.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77. Doi: https://doi.org/10.1089/cmb.2012.0021

Bishop OT. Bioinformatics and data analysis in microbiology. Grahamstown, Sudafrica: Caister Academic Press; 2014. 264 p.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120. Doi: https://doi.org/10.1093/bioinformatics/btu170

Cariaga Martinez AE, Zapata PD. Protocolos de extracción de ADN. El laboratorio de biología molecular. Edición ampliada. Buenos Aires, Argentina: Editorial universitaria; 2007. p. 23-39.

Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next generation sequencing short read sequences. Source Code Biol Med. 2014;9:8. Doi: https://doi.org/10.1186/1751-0473-9-8

Cox MP, Peterson DA, Biggs PJ. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics. 2010;11:485. Doi: https://doi.org/10.1186/1471-2105-11-485

Tabla 2. Ensamblado del genoma de *Bacillus altitudinis* 19RS3 (bacteria promotora del crecimiento vegetal aislada de *Ilex paraguariensis* St. Hil.). Resultados obtenidos de la evaluación por QUAST de los ensamblados realizados con el *software* SPAdes a partir de lecturas filtradas con Trimmomatic.

	# cóntigos	Longitud del cóntigo más largo	Longitud total	GC (%)	N50
k-mer 55	28	778 854	3 782 731	41,16	553 021
k-mer 61	24	830 193	3 781 196	41,16	774 421
k-mer 65	21	964 992	3 781 853	41,16	927 810
k-mer 67	18	964 996	3 783 031	41,16	927 814
k-mer 69	19	964 386	3 783 999	41,16	927 818
k-mer 71	18	964 394	3 782 835	41,16	927 786
k-mer 73	18	964 402	3 785 025	41,16	928 507
k-mer 75	19	964 492	3 784 944	41,16	928 507
k-mer 77	17	966 288	3 787 201	41,17	929 616
k-mer 79	**16**	**966 271**	3 788 682	41,18	**931 914**
k-mer 81	17	964 864	3 789 487	41,18	932 026
k-mer 83	17	964 870	3 789 762	41,18	932 030
k-mer 85	19	964 876	3 790 302	41,18	930 818
k-mer 87	20	964 882	3 791 338	41,18	894 990

QUAST: *software* utilizado para la evaluación de los ensamblados.
SPAdes: *software* utilizado para realizar los ensamblados.
Trimmomatic: *software* utilizado para el filtrado de las lecturas.
K-mer: valor k utilizado para cada ensamblado realizado con el *software* SPAdes.
cóntigos: número total de cóntigos de longitud ≥ 500 pb.
Longitud total: número total de bases en los cóntigos de longitud ≥ 500 pb.
GC (%): porcentaje de guanina y citosina.
N50: longitud del cóntigo más grande de todos los cóntigos clasificados de menor a mayor que representa el 50 % de la longitud del conjunto.
Las cifras destacadas en letra en negrita corresponden a los valores obtenidos para el mejor ensamblado generado con las lecturas filtradas.

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PloS one. 2013;8(12):e85024. Doi: https://doi.org/10.1371/journal.pone.0085024

Gladman S. De novo Genome Assembly for Illumina Data. Melbourne Bioinformatics [monografía en Internet]. 2019. Disponible en: https://www.melbournebioinformatics.org.au/tutorials/tutorials/assembly/assembly-protocol/#protocol. Citado: 30 abr 2020.

Góngora-Castillo E, Buell CR. Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Nat Prod Rep. 2013;30(4):490-500. Doi: https://doi.org/10.1039/c3np20099j

Gurevich A, Vladislav S, Nikolay V, Glenn T. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-1075. Doi: https://doi.org/10.1093/bioinformatics/btt086

Laczeski ME, Onetto AL, Cortese IJ, Mallozzi GY, Castrillo ML, Bich GA, *et al*. Isolation and selection of endophytic spore-forming bacteria with plant growth promoting properties isolated from *Ilex paraguariensis* St. Hil. (Yerba mate). An Acad Bras Cienc. 2020; 92. Doi: https://doi.org/10.1590/0001-3765202020181381

Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, *et al*. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2016;19(2):286-302. Doi: https://doi.org/10.1093/bib/bbw114

Martin M. Cutadapt removes adapter sequences from high throughput sequencing reads. EMBnet J. 2011;17(19:10-12. Doi: https://doi.org/10.14806/ej.17.1.200

Medvedev P, Pham S, Chaisson M, Tesler G, Pevzner P. Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. J Comput Biol. 2011;18(11): 1625-34. Doi: https://doi.org/10.1089/cmb.2011.0151

Patel RK, Jain M. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012;7:e30619. Doi: https://doi.org/10.1371/journal.pone.0030619

Rana SB, Zadlock FJ, Zhang Z, Murphy WR, Bentivegna CS. Comparison of De Novo Transcriptome Assemblers and k-mer strategies using the killifish, *Fundulus heteroclitus*.

PLoS One. 2016;11(4):e0153104. Doi: https://doi.org/10.1371/journal.pone.0153104

Rodríguez Hernáez JI. Ensamblaje y caracterización genómica de una bacteria celulolítica aislada del rumen bovino (Tesis de Licenciatura en Biotecnología). Buenos Aires: Facultad de Ingeniería y Ciencias Exactas, Universidad Argentina de la Empresa; 2017. 138 p.

Sambrook J, Rusell DW. Molecular cloning: a laboratory manual. Nueva York, USA: Cold spring harbor laboratory press; 2001. 1546 p.

Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011a;6:e17288. Doi: https://doi.org/10.1371/journal.pone.0017288

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

Smeds L, Künstner A. ConDeTri-A Content dependent read trimmer for Illumina data. PLoS One. 2011;6:e26314. Doi: https://doi.org/10.1371/journal.pone.0026314

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821-829. Doi: https://doi.org/10.1101/gr.074492.107