Rev. Salud Anim. Vol. 37 No. 2 (Mayo.-ago. 2015): 125-132
ISSN: 2224-4700

ARTÍCULO RESEÑA

Secuenciación y ensamblaje *de novo* de genomas bacterianos: una alternativa para el estudio de nuevos patógenos

Lisandra Aguilar-Bultet[I], Laurent Falquet[II]

[I]Departamento de Microbiología, Dirección de Salud Animal, Centro Nacional de Sanidad Agropecuaria (CENSA), Apartado 10, Mayabeque, Cuba. Correo electrónico: labultet@censa.edu.cu, labultet@gmail.com. [II]Departamento de Biología, División de Bioquímica, Universidad de Fribourg, Instituto Suizo de Bioinformática, Suiza.

RESUMEN: Las tecnologías de secuenciación de nueva generación han revolucionado el campo de la genómica, permitiendo la secuenciación de un gran número de genomas en muy poco tiempo. El genoma brinda, en principio, el catálogo completo de genes que un organismo puede expresar, pero la interpretación de esta información, a todos los niveles (desde el enorme volumen de datos crudos originados por el secuenciador hasta los miles de genes identificados en paralelo asociados a diversas funciones) constituye un gran reto desde el punto de vista computacional. El ensamblaje *de novo* es uno de los procedimientos empleados para reconstruir genomas, principalmente bacterianos. En este artículo se revisan las principales tendencias en la secuenciación y, específicamente, en el ensamblaje *de novo*.

Palabras clave: secuenciación de genomas, tecnologías de secuenciación de nueva generación, ensamblaje *de novo*.

Sequencing and *de novo* assembly of bacterial genomes: an approach to study new pathogens

ABSTRACT: Next Generation Sequencing technologies have revolutionized the field of genomics allowing sequencing of a large number of genomes in a short period of time. The genome provides, in principle, the complete catalogue of genes which can be expressed by a microorganism. But the interpretation of this information, at each level (from the huge volume of raw data originated by the sequencer to the thousands of genes identified in parallel associated with various functions), is a major challenge from the computational point of view. *De novo* assembly is one of the approaches used to reconstruct genomes, particularly bacterial genomes. In this paper the main trends regarding sequencing, and specifically the *de novo* assembly, are reviewed.

Key words: genome sequencing, next generation sequencing, *de novo* assembly.

INTRODUCCIÓN

La caracterización completa de un microorganismo en el laboratorio constituye un proceso muy costoso y laborioso que consume mucho tiempo. Con el incremento de las capacidades de secuenciación a partir del surgimiento de las tecnologías de secuenciación de nueva generación en el año 2005, se ha abierto un nuevo camino en este campo. Entre los principales aportes de estas nuevas tecnologías cabe citar que han facilitado la secuenciación del ADN genómico de un alto número de bacterias, generándose un gran volumen de datos en corto tiempo. La obtención de un genoma completo permite contar, en principio, con el catálogo completo de genes que un organismo puede expresar en cualquier momento de su ciclo de vida. De ahí la importancia de las tecnologías de nueva generación (NGS, por sus siglas en inglés, *Next Generation Sequencing*), que al permitir el procesamiento masivo y en paralelo de las muestras, reducen notablemente los costos y el tiempo para obtener la secuencia genómica, en comparación con la secuenciación automática de Sanger (1).

Para procesar y analizar el enorme volumen de datos biológicos acumulados, como resultado del uso de estas tecnologías, ha sido necesario el empleo de he-

rramientas bioinformáticas que permitan manipular eficientemente esta creciente cantidad de información, herramientas que también se han venido modificando y perfeccionando junto al propio desarrollo de las tecnologías NGS.

La presente revisión tiene como objetivo describir las tendencias actuales para la secuenciación y el ensamblaje *de novo* de genomas.

DESARROLLO

Secuenciación de genomas

La secuenciación de genomas completos es un método poderoso para la rápida identificación de genes en un organismo, y sirve como herramienta básica para posteriores análisis funcionales de los nuevos genes descubiertos. La secuencia genómica provee de un conjunto de virtualmente todas las proteínas que el organismo puede expresar.

El método de secuenciación automática de Sanger dominó la industria de secuenciación por casi 20 años, llevando a innumerables logros en este campo, como fue la secuenciación del primer genoma bacteriano *Haemophilus influenzae* y la primera secuencia completa del genoma humano. A pesar de las mejoras técnicas durante los últimos años, las limitaciones de la tecnología de Sanger trajo consigo la necesidad de desarrollar nuevas y mejores alternativas para la secuenciación de un gran número de genomas en corto tiempo (1). Es por ello que surgen las tecnologías de secuenciación de nueva generación.

La tecnología 454, conocida como pirosecuenciación, fue la primera NGS en salir al mercado entre los años 2004 y 2005 (2). A continuación surgieron *Illumina* en 2006 (3), basada en secuenciación por síntesis, *SOLiD* en 2007, basada en secuenciación por ligación, y *Ion Torrent* en el año 2010 (4), basada en detección de pH, las cuales necesitan de la amplificación del ADN previamente a su secuenciación. Además, se han desarrollado tecnologías que no necesitan del paso inicial de amplificación, sino que secuencian directamente una sola molécula de ADN, entre las que se encuentran *Helicos*, salida al mercado en 2008 (5) y *SMRT Pacific Biosciences* (*PacBio*) en 2010 (6).

Características generales de las tecnologías NGS

La mayor ventaja ofrecida por las NGS es la capacidad para producir un inmenso volumen de datos de forma económica, pudiendo llegar a millones o billones de *reads*[1] en solamente una corrida del equipo para un único genoma, en comparación con la secuenciación automática de Sanger, que puede llegar solo hasta cientos de *reads* (7), pero con una longitud de hasta 1000 pb aproximadamente. Por tanto, con las tecnologías NGS se incrementa considerablemente la cobertura del genoma, que no es más que la cantidad promedio de veces que un nucleótido es representado en un conjunto de secuencias crudas al azar (8). Sin embargo, los datos de NGS generalmente son secuencias más cortas (a excepción de los producidos por *PacBio*), que representan un reto desde el punto de vista computacional para su ensamblaje, debido a la longitud y la enorme cantidad de secuencias.

Aplicaciones de las tecnologías NGS en el estudio de patógenos

Antes de enfrentarse a un proyecto de secuenciación, los investigadores deben realizarse varias interrogantes: ¿Cuál es el microorganismo en cuestión? ¿Existe un genoma que se pueda utilizar como referencia?¿Qué características tiene el genoma? A partir de todo lo anterior se decide la tecnología más adecuada para la secuenciación y, por consiguiente, el método a emplear para el análisis de los datos, incluido el ensamblaje del genoma.

Las aplicaciones de las tecnologías NGS se extienden a muchos campos de la biología, como la genómica, la transcriptómica, la epigenética, la genómica de poblaciones, la metagenómica, entre otros. Inicialmente, la aplicación más obvia fue la secuenciación de genomas completos, incluyendo resecuenciación o secuenciación *de novo*. Proyectos de resecuenciación requieren un genoma de referencia en el que se alinean los *reads* para detectar variaciones. Por el contrario, los proyectos *de novo* no tienen genoma de referencia disponible, y los *reads* se utilizarán únicamente para la reconstrucción de todo el genoma. En la actualidad, las NGS se emplean, además, para la secuenciación de ARN con el objetivo de cuantificar la expresión génica. Paralelamente, han surgido técnicas más específicas como la secuenciación acoplada a inmunoprecipitación de la

[1] *reads*: secuencias obtenidas como resultado del proceso de secuenciación.

cromatina (*ChIP-seq*) para el estudio de las interacciones ADN-proteínas, la secuenciación de ADN con bisulfito (*Methyl-seq*) para el análisis de la metilación del ADN, la secuenciación de ADN asociada a sitios de restricción (*RAD-seq*), entre otras.

Las NGS se han empleado en el estudio de microorganismos patógenos. Un ejemplo palpable lo constituyen los estudios realizados en *Avibacterium paragallinarum*, agente causal de la coriza infecciosa en pollos, que afecta considerablemente a la industria avícola, provocando grandes pérdidas por concepto de retardo en el crecimiento y disminución en la producción de huevos. La obtención de varios borradores de diferentes cepas virulentas de este genoma ha permitido la identificación de los factores de virulencia más importantes de este microorganismo (9, 10, 11, 12), con especial énfasis en la toxinas RTX (por sus siglas en inglés, *Repeats in ToXins*) y CDT (por sus siglas en inglés, *Cytolethal Distending Toxins*) (11); trabajos que han sido validados experimentalmente en el laboratorio (13, 14), y que han demostrado que las técnicas bioinformáticas brindan una adecuada aproximación a la realidad biológica.

La secuenciación de genomas también ha sido aplicada recientemente al estudio de los mecanismos de resistencia a antibióticos, específicamente a la linezolida, de especies de *Staphylococcus* como *S. epidermidis*, el cual se caracteriza por ser un contaminante frecuente de muestras de laboratorio y estar presente en la piel y mucosas de humanos y diversos animales (bovinos, caninos, equinos), pero también se ha descrito que puede causar mastitis, ocasionalmente, en ganado bovino. Como resultado de la secuenciación por NGS de 28 aislados, se identificaron mutaciones en el ARNr 23S y en las riboproteínas L3 y L4, que están relacionadas con la resistencia a la linezolida (15). También se ha estudiado la resistencia a antibióticos en *Salmonella enterica* serovar Heidelberg, bacteria de trasmisión alimentaria que puede afectar a humanos, cerdos, pavos, entre otros. Las NGS han permitido el estudio de algunas cepas muy patogénicas de este serovar que por métodos convencionales no se pueden diferenciar de otras menos patogénicas (16).

Se ha aplicado también esta metodología para la identificación y tipificación de cepas de *Streptococcus suis*, patógeno de gran incidencia para la industria porcina mundial, que provoca meningitis y en menor medida septicemia, artritis, neumonía, entre otros; constituye, además, un peligro para los humanos con riesgo para la vida. Se secuenciaron y compararon un total de 85 aislados de *S. suis*, lo cual permitió no solo la identificación taxonómica de los mismos, sino que se demostró su potencial patogénico y epidémico. A partir del análisis del genoma núcleo en todos los aislados, se lograron diferenciar claramente aquellos que provocan síntomas severos, de aquellos que solo causan meningitis esporádicas de menor impacto (17).

Para los micoplasmas también se han empleado las NGS, como es el caso de *Mycoplasma mycoides mycoides* «Small Colony» (MmmSC), el cual es el agente responsable de la pleuroneumonía contagiosa bovina, enfermedad de notificación obligatoria de la Organización Mundial de Sanidad Animal (OIE). Se emplearon 20 cepas representativas de lugares geográficos diferentes. Como resultado de la secuenciación y ensamblaje de los genomas, se pudo obtener un robusto árbol filogenético que identificó un linaje específico a las cepas europeas, mientras las cepas africanas se ubican dispersas en varias ramas. Se pudo determinar, además, el ancestro común más reciente y los posibles orígenes de las cepas africanas y de los brotes recientes en Europa (18).

Por otro lado, la secuenciación del genoma de la cepa Illinois de *Mycoplasma suis*, agente causal de la anemia infecciosa porcina, permitió la identificación de mecanismos metabólicos responsables probablemente de su adaptación, crecimiento y desarrollo en los eritrocitos (19).

Ensamblaje de genomas

Al proceso de descifrar la secuencia genómica a partir de pequeños fragmentos de ADN, en conjunto con alguna información adicional disponible, se le denomina ensamblaje de genomas. Las estrategias para el ensamblaje de genomas se pueden dividir en dos categorías: ensamblaje por comparación, en el que se utiliza un genoma como referencia; y ensamblaje *de novo*, en el cual se utiliza solo la información obtenida de la secuenciación para reconstruir el genoma en cuestión, sin conocimiento *a priori* de la organización del mismo (20). Sin embargo, en esta última estrategia algunas informaciones previas son útiles, como la talla esperada del genoma, el contenido de GC y el contenido de regiones repetitivas, ya que ayudan a elegir la mejor estrategia a seguir. Estos datos pueden ser inferidos a partir de secuencias de organismos relacionados. El ensamblaje *de novo* con datos de NGS se limita generalmente a proyectos de genomas microbianos debido a su pequeña talla (21, 22). Esta metodología es la más desafiante, y a la que nos referiremos en lo adelante.

De manera general, los pasos para el ensamblaje *de novo* se resumen en la Figura.

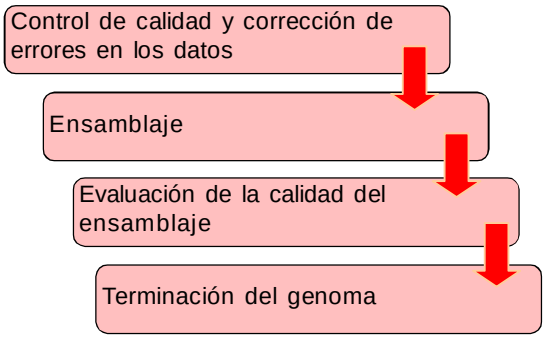

FIGURA. Representación del flujo de trabajo a seguir de manera general para la obtención de la secuencia genómica de un organismo./ *Graphical representation of the workflow followed to obtain the genome sequence of an organism.*

Control de calidad y corrección de los datos

El control de calidad de los datos crudos sirve como un chequeo rápido para identificar y excluir datos con problemas serios de calidad, lo cual permite ahorrar gran cantidad de tiempo en los análisis posteriores (23). Las herramientas empleadas chequean la calidad de las bases (probabilidad de que la base asignada sea la correcta), la distribución de los nucleótidos, la distribución del contenido de GC, secuencias repetidas, entre otros parámetros. En dependencia de la calidad inicial de los datos, serán los procesos de corrección a seguir. En los casos de secuencias cortas, producidas por tecnologías NGS de segunda generación (*Illumina*, *IonTorrent*, SOLiD), la tendencia es a filtrar los *reads* que tengan poca calidad, o cortarlos a partir de la posición en la cual la calidad comienza a decaer. Se estima que el porcentaje de error de los datos crudos se encuentra entre 0.1-1%.

En plataformas de tercera generación, como *PacBio RS*, se construyen dos tipos de bibliotecas: CLR (*Continuous Long Reads*, por sus siglas en inglés, *reads* largos continuos) y CCS (*Circular Consensus Sequences*, por sus siglas en inglés, secuencias consenso circulares). Durante la corrección de los errores de la biblioteca CLR, se usan *reads* cortos de alta calidad, como los CCS o los producidos por *Illumina* o *454*. Se estima que el porcentaje de error de los datos CLR crudos es de aproximadamente del 15% (24); por ello se recomienda corregirlos, ya sea con datos de segunda generación o con secuencias CCS. Pero la plataforma *PacBio RS* está obsoleta y casi en desuso. En la versión actual, *PacBio RS II*, solo se construye la biblioteca CLR, pero con una mayor cobertura de *reads*, lo que permite agruparlos de acuerdo a la longitud y posteriormente utilizar los más cortos para la corrección de los más largos, y de esta manera eliminar errores. Con el desarrollo del paquete de programas HGAP se ha facilitado la manipulación de este tipo de datos, desde su corrección hasta el ensamblaje (25). De este modo, según los fabricantes, se reducen los errores por debajo del 0.1%.

Metodologías empleadas en el ensamblaje *de novo*

Todos los métodos de ensamblaje se basan en la simple suposición de que fragmentos de ADN altamente similares se originan de la misma posición dentro del genoma. De esta manera, la similitud entre secuencias de ADN se usa para conectar fragmentos individuales en secuencias contiguas más largas, denominadas *contigs* (secuencias consenso obtenidas a partir del ensamblaje de los *reads*).

Las tecnologías NGS han reformado la biología en la actualidad, incluyendo el ensamblaje de genomas. En comparación con el método tradicional de Sanger, el rendimiento de los datos obtenidos por estas nuevas tecnologías es significativamente mayor y los costos mucho menores (1). Sin embargo, representan un nuevo reto desde el punto de vista computacional para el ensamblaje *de novo*, debido a la corta longitud de los fragmentos secuenciados.

Los *reads* de corta longitud constituyen un problema cuando hay repeticiones en el genoma (7). Las repeticiones o regiones repetitivas son segmentos de ADN que aparecen más de una vez a lo largo del genoma. Cuando un *read* proviene de una región repetitiva, y es más corto que esta, no se sabe con certeza de cuál copia de la repetición se obtuvo. Es por ello que durante el ensamblaje, se pueden crear falsas uniones en el genoma en las regiones de repeticiones. Además, debido a las regiones repetitivas, muchas veces resulta difícil ensamblar todos los fragmentos para que se logre, en un solo evento, reconstruir la secuencia del genoma completo, incluso en genomas microbianos pequeños.

Para abordar estos problemas debidos a los fragmentos cortos en las tecnologías NGS de segunda generación, se han desarrollado nuevos secuenciadores que incrementan la longitud de los mismos mientras mantienen el rendimiento (cantidad de ADN que puede ser procesado por unidad de tiempo) (26). Por ejemplo, el secuenciador SMRT de *Pacific Biosciences* produce *reads* de varios kb de longitud (2-10 kb) (6, 27), incluso en las últimas versiones pueden alcanzar los 20 kb (http://www.pacificbiosciences.com/products/smrt-technology/smrt-sequencing-advantage/). Sin embargo, esta tecnología todavía no es estable en térmi-

nos de calidad de los *reads* (1). No obstante, se han obtenido interesantes y buenos resultados a partir del ensamblaje de datos de *PacBio*, previamente corregidos con fragmentos de alta calidad y corta longitud, obteniéndose borradores de genomas con una calidad mejorada en relación con los anteriormente obtenidos, en el caso de *Avibacterium paragallinarum* (11). También se obtuvo el genoma de la cepa JF4335 de *Clostridium chauvoei* combinando datos de *Illumina* y datos de *PacBio*, con excelentes resultados (28). Actualmente la nueva plataforma *PacBio RS II,* y el subsiguiente procesamiento de los *reads* con el paquete HGAP, han permitido la obtención de un solo *contig* como resultado del proceso de ensamblaje, lo cual supera cualquier programa previamente empleado (25, 29).

Computacionalmente es posible alargar los *reads* de NGS usando la técnica de extremos pareados (PE, por sus siglas en inglés, *paired-ends*). En este caso, se conoce la secuencia de ambos extremos (que serían los *reads*) de un fragmento de ADN y la distancia aproximada entre ellos. Cuando la longitud del fragmento de ADN es menor que dos veces la longitud del *read*, los dos extremos se solapan, lo cual permite unirlos para formar una secuencia más larga.

Las estrategias empleadas por los programas ensambladores de secuencias pueden agruparse en tres paradigmas principales: Greedy, Overlap-Layout-Consensus y gráficos de Bruijn (30).

Greedy: Es el algoritmo más sencillo e intuitivo. El ensamblador siempre conecta los *reads* que mejor se solapan, de manera iterativa, mientras no contradigan el ensamblaje ya construido. Sin embargo, esta metodología no es ampliamente empleada, ya que es inherentemente un proceso de ensamblaje local, no emplea información global, y no resuelve de manera eficiente largas regiones repetitivas en los genomas. Greedy se usa generalmente para el ensamblaje de datos originados por secuenciación Sanger.

OLC (por sus siglas en inglés, Overlap-Layout-Consensus): Este método primero identifica todos los pares de *reads* que se solapan lo suficientemente bien y organiza esta información en un gráfico en el cual hay un nodo por cada uno de ellos y un conector (*edge*) por cada solapamiento entre los mismos. Esta estructura del gráfico permite el desarrollo de complejos algoritmos de ensamblaje que tienen en cuenta la relación global entre los *reads*. De esta manera se definen caminos, que corresponden con los segmentos del genoma que están siendo ensamblados. Finalmente, se reconstruye el genoma mediante la búsqueda de un único camino que atraviese todos los nodos solo una vez. Este paradigma dominó el mundo del ensamblaje hasta la emergencia de las tecnologías NGS.

Gráficos De Bruijn: Los ensambladores basados en gráficos De Bruijn modelan la relación entre subcadenas exactas de longitud *k* dentro de los *reads*. De manera similar al método OLC, los nodos en el gráfico representan *k-mers*, y los conectores indican que *k-mers* adyacentes se solapan por *k*-1 letras, por lo que la longitud del *k-mer* correlaciona con la longitud del solapamiento que el ensamblador es capaz de detectar. En esta metodología no se modelan directamente los *reads*, sino que están implícitamente representados por los conectores en el gráfico de Bruijn. La mayoría de los ensambladores usan la información global de los *reads* para refinar la estructura del gráfico, resolver repeticiones y eliminar patrones no consistentes. Además, incorporan métodos de corrección de errores para mejorar la calidad del ensamblaje.

Muchos de los programas ensambladores incluyen además, el proceso de *scaffolding*, mediante el cual intentan conectar los *contigs* obtenidos empleando la información que brindan las bibliotecas PE, cuando varios *reads* en el extremo de un *contig* «apuntan» todos hacia otro *contig*. A pesar de no conocer la secuencia entre ellos, se pueden conectar, dejando una distancia aproximada, determinada por la longitud del inserto. Por tanto, se puede inferir cuándo dos *contigs* son adyacentes, si cada *read* PE se ubica en cada *contig* a la distancia y orientación esperadas.

Otra herramienta muy útil en el proceso de *scaffolding* es la construcción de bibliotecas *mate-pairs*, en la cual, de manera similar a la biblioteca PE, se conocen las secuencias de dos segmentos separados por una distancia determinada que también se conoce, pero en este caso la distancia puede ser de hasta 150 kb (31). Esta información puede ser utilizada para concatenar un conjunto de *contigs* y formar *scaffolds*. Se puede inferir que dos *contigs* son adyacentes en el genoma si un extremo de un *mate-pair* está unido al primer *contig* y el otro extremo está fusionado al segundo. Esta estrategia la utilizan programas como Euler (32), Bambus (33), Celera (34), Velvet (35) y Arachne (36).

La compañía *Pacific Bioscences*, con el lanzamiento de su máquina secuenciadora, también lanzó un *pipeline* para el análisis y ensamblaje de los datos, denominado *SMRT Pipe*®. Este *pipeline* incluye múltiples herramientas para la corrección de errores, ensamblaje *de novo*, ensamblaje híbrido (cuando se combinan datos de más de una tecnología), detección de variantes estructurales, detección de modificaciones postranscripcionales, entre otras.

Evaluación de la calidad del ensamblaje

Aunque utópico, el objetivo final del ensamblaje *de novo* es obtener un número de fragmentos igual al número total de cromosomas, en el caso de las bacterias un cromosoma, y también pueden haber plásmidos que se obtienen como fragmentos separados. Por lo tanto, el investigador siempre busca obtener el menor número de *contigs* posible.

Existen algunos indicadores métricos que permiten evaluar la calidad del ensamblaje cuantitativamente. Se calcula generalmente la talla mínima, máxima y media de los *contigs*, así como la talla total del ensamblaje, la cual debe coincidir con la talla esperada del genoma. Pero el principal valor estadístico es el valor N50, el cual corresponde con el menor de los mayores *contigs* que cubren la mitad del genoma. Aunque el valor N50 constituye un indicador acerca de la contigüidad del genoma (i.e. cuán acertado fue el programa ensamblador en unir las secuencias contiguas en una única secuencia más larga), no es una señal de precisión y calidad del genoma ensamblado. Tampoco brinda una correcta estimación de si la talla del ensamblaje difiere o no de la talla esperada (30).

Para una mejor valoración de la calidad, desde el punto de vista cualitativo, se puede realizar el alineamiento de los *reads* con los *contigs* obtenidos, proceso denominado remapeo (30). La visualización de estos alineamientos permite analizar la consistencia del genoma y si los *contigs* son confiables; además, la información PE y las variaciones inesperadas de cobertura permiten identificar potenciales regiones mal ensambladas (37). Otra estrategia es comparar la secuencia obtenida con otras secuencias genómicas, ya sea una secuencia de referencia, o genomas de organismos relacionados (38).

Proceso de terminación del genoma

El proceso de terminación del genoma es el paso que más tiempo consume en la secuenciación. Muchas veces se asocia también al proceso de cierre de *gaps*, pero esencialmente está dedicado a corregir los errores que quedaron en el ensamblaje, y errores de secuenciación. Los métodos *in silico* empleados durante la fase de terminación del genoma incrementan considerablemente la posibilidad de cerrar el genoma, pero aun así, cuando quedan *gaps* entre los *contigs*, o las conexiones entre ellos son ambiguas, solo las técnicas de laboratorio pueden ayudar. Entre las técnicas más empleadas se encuentran «*chromosome walking*» (39), PCR múltiple optimizado (40) asociado a secuenciación Sanger, y la técnica mapa óptico (41, 42). Pero estas técnicas son muy laboriosas y costosas.

Anotación de genomas

Después de un exitoso ensamblaje del genoma, el próximo reto consiste en interpretar la información que contiene. Para ello es necesaria la identificación de las principales características del genoma, proceso conocido como anotación. La anotación de genomas comprende dos etapas fundamentales: la anotación estructural (predicción de regiones codificantes) y la anotación funcional (asignación de información biológica a los genes previamente predichos).

Los métodos para la anotación estructural en un genoma se dividen en dos categorías: método *ab initio* o *de novo*, y método por comparación (43). El método *ab initio* utiliza algoritmos estadísticos o de reconocimiento de patrones para determinar si la secuencia de interés es codificante o no, mediante la detección de patrones o motivos específicos en la secuencia. Por otro lado, el método por comparación identifica zonas de alta similitud en organismos relacionados o en bases de datos de proteínas para reconocer las regiones codificantes. Sin embargo, este método es menos exitoso en la identificación de nuevos genes y en nuevos organismos, ya que las bases de datos están sesgadas hacia los genes altamente expresados en los organismos más estudiados.

Para la anotación funcional, también se emplean distintos métodos. La función de un gen se puede inferir mediante la búsqueda de secuencias homólogas en bases de datos, empleando algoritmos de alineamiento local, como *BLAST* (44). Esta asignación se realiza tomando como base la premisa de que genes con secuencias compartidas, también comparten su función. Otra metodología empleada es la búsqueda de motivos y dominios funcionales. Aunque un dominio funcional no permite asignar un nombre directamente al gen, sí puede dar una idea de la familia génica a la que pertenece el gen, o indicar el grupo de procesos en los que pueda estar involucrado.

CONCLUSIONES

Las NGS han tenido un gran impacto en muchos campos de la biología actual, permitiendo el estudio simultáneo de varios microorganismos de interés. Sin embargo, supone un gran reto desde el punto de vista computacional por el volumen de datos que ofrece. No podemos referirnos a una tecnología «mejor» para la secuenciación de genomas, ni tampoco a la «mejor» metodología para el ensamblaje de secuencias. En dependencia de las características del microorganismo en cuestión, y del objetivo final de la investigación, se escoge la tecnología para la secuenciación y la subsiguiente metodología para el ensamblaje. Se han

descrito múltiples aplicaciones de la secuenciación de genomas para el estudio de microorganismos patógenos. Entre ellos se destaca la resecuenciación, mediante la cual se corrigen errores o se completan secuencias genómicas previamente obtenidas por una tecnología diferente.

Una vez obtenida la secuencia genómica y anotados los genes que contiene, la misma constituye una herramienta poderosa para la caracterización del microorganismo, a través de la identificación de factores de virulencia, análisis comparativos y filogenéticos con otros microorganismos, estudio de redes metabólicas, entre otros, así como elementos que puedan contribuir al diagnóstico, la prevención y el control de la enfermedad que produce el agente.

REFERENCIAS

1. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31-46.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.
3. Bentley DR. Whole-genome re-sequencing. Curr Opin Genet Dev. 2006;16(6):545-52.
4. Pennisi E. Genomics. Semiconductors inspire new sequencing technologies. Science. 2010; 327(5970):1190.
5. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008;320(5872):106-9.
6. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133-8.
7. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13(1):36-46.
8. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121-32.
9. Xu F, Miao D, Du Y, Chen X, Zhang P, Sun H. Draft Genome Sequence of *Avibacterium paragallinarum* Strain 221. Genome Announc. 2013;1(3).
10. Requena D, Chumbe A, Torres M, Alzamora O, Ramirez M, Valdivia-Olarte H, et al. Genome sequence and comparative analysis of *Avibacterium paragallinarum*. Bioinformation. 2013;9(10):528-36.
11. Aguilar-Bultet L, Calderon-Copete SP, Frey J, Falquet L. Draft Genome Sequence of the Virulent *Avibacterium paragallinarum* Serotype A Strain JF4211 and Identification of Two Toxins. Genome Announc. 2013;1(4).
12. Roodt Y. Towards unravelling the genome of *Avibacterium paragallinarum*. Bloemfontein: University of the Free State; 2009.
13. Chen YC, Tan DH, Shien JH, Hsieh MK, Yen TY, Chang PC. Identification and functional analysis of the cytolethal distending toxin gene from *Avibacterium paragallinarum*. Avian Pathol. 2014;43(1):43-50.
14. Kung E, Frey J. AvxA, a composite serine-protease-RTX toxin of *Avibacterium paragallinarum*. Vet Microbiol. 2013;163(3-4):290-298.
15. Tewhey R, Gu B, Kelesidis T, Charlton C, Bobenchik A, Hindler J, et al. Mechanisms of linezolid resistance among coagulase-negative staphylococci determined by whole-genome sequencing. mBio. 2014;5(3): e00894-14.
16. Evans PS, Luo Y, Muruvanda T, Ayers S, Hiatt B, Hoffman M, et al. Complete Genome Sequences of Salmonella enterica Serovar Heidelberg Strains Associated with a Multistate Food-Borne Illness Investigation. Genome Announc. 2014;2(3).
17. Chen C, Zhang W, Zheng H, Lan R, Wang H, Du P, et al. Minimum core genome sequence typing of bacterial pathogens: a unified approach for clinical and public health microbiology. Journal of clinical microbiology. 2013;51(8):2582-91.
18. Dupuy V, Manso-Silvan L, Barbe V, Thebault P, Dordet-Frisoni E, Citti C, et al. Evolutionary history of contagious bovine pleuropneumonia using next generation sequencing of *Mycoplasma mycoides* subsp. *mycoides* «Small Colony». PLoS One. 2012;7(10):e46821.
19. Guimaraes AM, Santos AP, SanMiguel P, Walter T, Timenetsky J, Messick JB. Complete genome sequence of *Mycoplasma suis* and insights into its biology and adaption to an erythrocyte niche. PLoS One. 2011;6(5):e19574.
20. Wajid B, Serpedin E. Review of general algorithmic features for genome assemblers for next generation

sequencers. Genomics Proteomics Bioinformatics. 2012;10(2):58-73.

21.Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010;6(2):e1000667.

22.Chistoserdova L. Recent progress and new challenges in metagenomics for biotechnology. Biotechnol Lett. 2010;32(10):1351-1359.

23.Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. 2013.

24.Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13:238.

25.Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563-569.

26.Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227-40.

27.English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012;7(11):e47768.

28.Falquet L, Calderon-Copete SP, Frey J. Draft Genome Sequence of the Virulent *Clostridium chauvoei* Reference Strain JF4335. Genome Announc. 2013;1(4).

29.Satou K, Shiroma A, Teruya K, Shimoji M, Nakano K, Juan A, et al. Complete Genome Sequences of Eight *Helicobacter pylori* Strains with Different Virulence Factor Genotypes and Methylation Profiles, Isolated from Patients with Diverse Gastrointestinal Diseases on Okinawa Island, Japan, Determined Using PacBio Single-Molecule Real-Time Technology. Genome Announc. 2014;2(2).

30.Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14(3):157-167.

31.Pop M. Genome assembly reborn: recent computational challenges. Brief Bioinform. 2009;10(4):354-366.

32.Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001;98(17):9748-9753.

33.Pop M, Kosack DS, Salzberg SL. Hierarchical scaffolding with Bambus. Genome Res. 2004;14(1):149-159.

34.Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science. 2000;287(5461):2196-2204.

35.Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821-829.

36.Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, et al. ARACHNE: a whole-genome shotgun assembler. Genome Res. 2002;12(1):177-189.

37.Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 2008;9(3):R55.

38.Meader S, Hillier LW, Locke D, Ponting CP, Lunter G. Genome assembly quality: assessment and improvement using the neutral indel model. Genome Res. 2010;20(5):675-684.

39.Wang Z, Ye S, Li J, Zheng B, Bao M, Ning G. Fusion primer and nested integrated PCR (FPNI-PCR): a new high-efficiency strategy for rapid chromosome walking or flanking sequence cloning. BMC Biotechnol. 2011;11:109.

40.Tettelin H, Radune D, Kasif S, Khouri H, Salzberg SL. Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project. Genomics. 1999;62(3):500-507.

41.Nagarajan N, Read TD, Pop M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. Bioinformatics. 2008;24(10):1229-1235.

42.Samad AH, Cai WW, Hu X, Irvin B, Jing J, Reed J, et al. Mapping the genome one molecule at a time--optical mapping. Nature. 1995;378(6556):516-517.

43.Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, et al. Identifying protein-coding genes in genomic sequences. Genome Biol. 2009;10(1):201.

44.Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-410.

Recibido: 30-9-2014.
Aceptado: 2-3-2015.