¿CÓMO FUNCIONA?

Aplicaciones de las nuevas tecnologías de secuenciación

Rosario Carmona Muñoz
Licenciada en Biología. Plataforma Andaluza de Bionformática, Universidad de Málaga
rosariocarmona@gmail.com

En el número 128 de *Encuentros en la Biología* (1) aparecía publicada una visión general de las tres generaciones de la secuenciación (NGS, *new generation sequencing)*. Estas tecnologías evolucionan a pasos agigantados, y desde aquel artículo hasta hoy cabe reseñar, por ejemplo, el considerable avance de la tecnología de *Illumina*, con su sistema *HiSeq 2500/ 1500*, capaz de generar hasta 600 Gigabases por reacción. Destaca también el desarrollo de un novedoso sistema de secuenciación: *Ion Torrent*, en el que los nucleótidos no se detectan por fluorescencia o emisión de luz, sino por el cambio de pH como resultado de la liberación de un protón tras la incorporación del nucleótido. En general, las distintas tecnologías pretenden aumentar el número de nucleótidos secuenciados y disminuir su coste. Gracias a ello, las nuevas tecnologías, además de servir para secuenciar, tienen otras aplicaciones, algunas de las cuales comento en este artículo (Figura 1).

Estrategias para la secuenciación de genomas *de novo*

Existen dos estrategias principales a la hora de secuenciar un genoma. La elección de una u otra dependerá del tamaño y la complejidad del genoma en cuestión, de manera que cada una supla las carencias de la otra y así recabar la máxima información posible:

- **BAC a BAC**: también se denomina *secuenciación aleatoria jerárquica*. Consiste en digerir el genoma en fragmentos solapantes que se clonarán en BAC (cromosomas bacterianos artificiales) y se secuenciarán por separado para acabar ensamblándolos. Esta estrategia reduce la complejidad del ensamblaje, especialmente en zonas con elementos repetitivos, pero posee la desventaja del alto coste de tiempo y esfuerzo que supone producir y mapear una genoteca de BAC.

- **Método WSG** (*Whole-genome Shotgun Sequencing*): en este método se trocea el genoma aleatoriamente en pequeños fragmentos de tamaño definido, que se secuenciarán (lecturas) y ensamblarán computacionalmente para generar una secuencia consenso. Cuando el genoma contiene muchos elementos repetitivos, el ensamblaje se complica enormemente: al romper en fragmentos una región repetitiva, muchas de las lecturas resultantes son iguales o muy similares, y esto puede ocasionar que las secuencias de la misma repetición se colapsen en una única repetición, por lo que podrían acabar conectándose dos fragmentos realmente distantes en el genoma (quimeras). Además, el carácter aleatorio de la generación de la secuencia implica que algunas partes del genoma estarán cubiertas por varias lecturas, mientras que otras regiones podrían no estar ni siquiera representadas, lo que dejará huecos en el ensamblaje. Por tanto, para tener la certeza de que cada una de las bases del genoma ha sido secuenciada al menos en una lectura, se necesita una cobertura (número medio de veces que se secuencia cada nucleótido) mínima de 20-30X cuando se usan lecturas largas y de 72X cuando se emplean lecturas cortas. Este método se ha vuelto muy popular a raíz de la construcción de lecturas **pareadas** (*paired-end* y *mate pairs*). Mientras que en la estrategia original por WSG se obtiene la lectura de un único extremo de cada fragmento (*single end*), la secuenciación con pareadas proporciona la secuencia de ambos extremos de cada fragmento, siempre separados por una distancia conocida. Los nuevos protocolos tratan de aumentar la distancia entre estos extremos para que puedan resolverse secuencias repetitivas cada vez más largas.

Resecuenciación

La resecuenciación se puede llevar a cabo cuando se dispone de un genoma de referencia, preferiblemente de la misma especie, o, en su defecto, de alguna especie muy cercana. Consiste en alinear (mapear, del inglés *mapping*) las lecturas sobre el genoma de referencia para detectar las diferencias entre ambos. Los algoritmos de mapeo son mucho más rápidos y precisos que los de ensamblaje. Además, como se parte de un genoma de referencia, tiene la ventaja de no necesitar tanta cobertura, y su coste es menor. La resecuenciación permite el estudio de la variación genética entre individuos, al mismo tiempo que aumenta la representatividad de cada especie en las bases de datos, esto es, incrementa el número de individuos secuenciados de una misma especie.

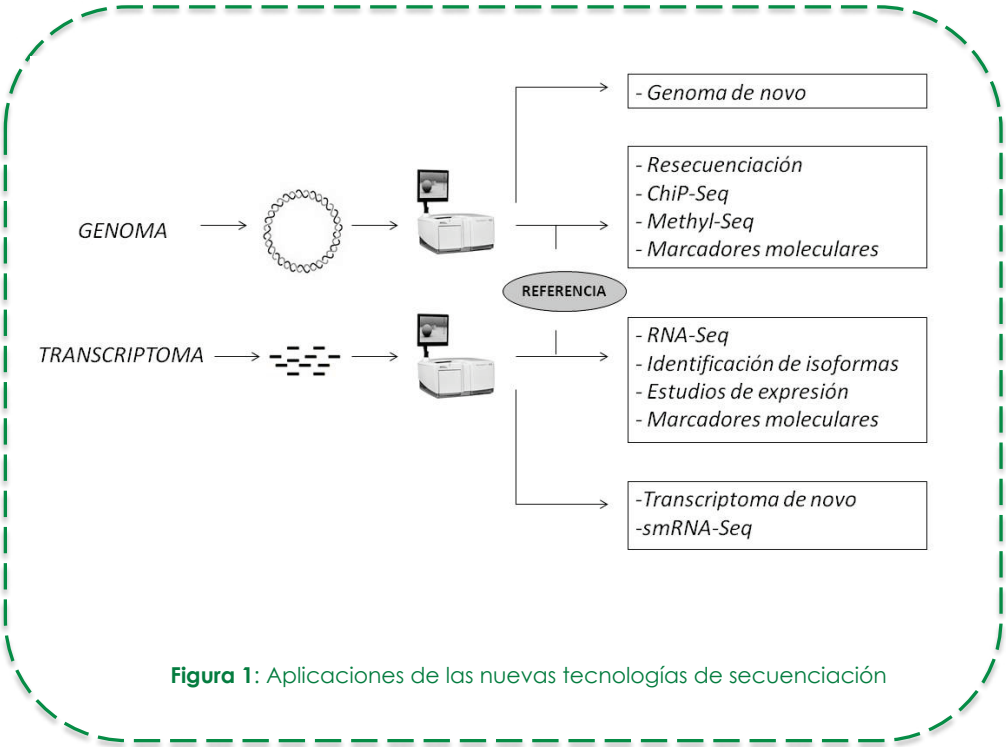

Figura 1: Aplicaciones de las nuevas tecnologías de secuenciación

Interacciones DNA-proteína: *ChiP-Seq*

Las interacciones DNA-proteína desempeñan una función clave en los distintos procesos celulares. Uno de los primeros métodos utilizado para su estudio en masa es *ChiP-on-chip*, que combina la inmunoprecipitación de cromatina con las micromatrices. Al acoplar la NGS a los fragmentos de DNA purificados que se obtienen del *ChiP-on-chip*, se generó el médodo de *ChiP-Seq*, que proporciona mapas de interacción DNA-proteína de mayor resolución y con menos ruido de fondo. La *ChiP-Seq* se ha usado, por ejemplo, para mapear sitios de unión de factores de transcripción en los genes de diferenciación celular y de proliferación (2).

Metilación: *Methyl-Seq*

Se sabe que la metilación de las citosinas en los eucariotas es clave para la regulación de la replicación y de la transcripción. Las citosinas darán lugar a uracilo cuando el DNA se trata con bisulfito de sodio, y se secuenciarán como timinas, mientras que las metilcitosinas se seguirán secuenciando como citosinas. Este hecho, junto con la aplicación de NGS, permite que los investigadores generen el metiloma de un genoma completo (3).

RNA-Seq en estudios de expresión génica

Las mejoras en la eficacia y calidad, así como el abaratamiento de los costes de la secuenciación de genomas completos está llevando a los investigadores a sustituir las clásicas micromatrices por aquellas basadas en NGS. La *RNA-Seq* consiste en la secuenciación profunda de cDNA de diferentes tipos celulares, mutantes, condiciones ambientales o estados de desarrollo, y la cuantificación de las lecturas correspondientes a cada transcrito como medida de su nivel de expresión en valores absolutos. Esta técnica es mucho más eficaz para distinguir entre genes parálogos y para detectar transcritos poco abundantes, y permite cuantificaciones reproducibles. Además resulta útil para la identificación de polimorfismos y de nuevas isoformas de ayuste. Al contrario que las micromatrices, la *RNA-Seq* no requiere necesariamente un genoma de referencia, pues hay programas capaces de cuantificar la expresión aún cuando no existan anotaciones disponibles, si bien es cierto que se obtienen mejores resultados cuando hay se dispone de un genoma de referencia para determinar la identidad de los genes.

Secuenciación de transcriptomas *de novo*

La secuenciación de cDNA es una sólida técnica que posibilita la caracterización del transcriptoma de un organismo de forma rápida y barata. Proporciona información sobre los genes de un organismo a menor coste que la secuenciación genómica, ya que solo se investigan aquellas regiones que se están transcribiendo. Tradicionalmente, los proyectos de transcriptómica se basaban en secuenciación de EST (secuencias etiquetadas por su expresión, del inglés *expressed sequence tag*) por el método de Sanger, pero la reciente aplicación de la NGS está poniendo de manifiesto una sorprendente e inesperada complejidad de los genomas eucariotas (sobre todo respecto a las formas de ayuste alternativo), además de recomponer los transcritos completos, sin tener que clonar antes el cDNA. Si se analiza el transcriptoma de una especie cuyo genoma ya está secuenciado, se pueden identificar nuevos transcritos, comprobar y optimizar supuestos transcritos e identificar nuevas isoformas de ayuste. Cuando se pretende descubrir qué genes se están expresando, la muestra de RNA tiene que estar normalizada, con el fin de aumentar la representación de los genes pocos expresados y disminuir la de los sebreexpresados. Esto está revelando miles de transcritos raros que previamente pasaban inadvertidos.

Se recomienda recurrir a las tecnologías de secuenciación que generan lecturas largas para facilitar el ensamblaje del transcriptoma. No obstante, la gran cantidad de lecturas que genera la NGS de lectura corta resulta extremadamente potente para: la *RNA-Seq*, la corrección de la secuencia consenso generada tras el ensamblaje y la detección de formas de ayuste alternativo.

smRNA-Seq

Las tecnologías de NGS también se están empleando para analizar RNA pequeños (smRNA), ya que estas tecnologías producen lecturas lo suficientemente largas para cubrirlos. Gracias a ello es posible conocer el smRNAoma. Destacan por su interés los miRNA (microRNA) y siRNA (RNA pequeños interferentes), que regulan la transcripción y la traducción de los genes. Se ha visto, por ejemplo, que existe una estrecha correlación entre la localización de algunos smRNA y los sitios de metilación en el DNA. En general, la identificación de nuevos smRNA a partir de la smRNA-Seq está basada en la comparación entre especies, o bien en las características de cómo se obtienen los smRNA maduros a partir de sus precursores. Recientemente se han desarrollado proyectos y paquetes de software para llevar a cabo el análisis a gran escala de sets de datos de s*mRNA-Seq*, con el objetivo de anotar dichos smRNAs en el genoma, construir perfiles de expresión y descubrir nuevos smRNA.

Descubrimiento de marcadores

La identificación de marcadores moleculares sirve para evaluar la variación dentro de una población o especie. El desarrollo de marcadores moleculares no siempre requiere poner en marcha reacciones de NGS, ya que se pueden deducir de las secuencias ya publicadas y disponibles con la ayuda de las herramientas bioinformáticas adecuadas. En las especies con el genoma o transcriptoma completamente secuenciado ya existen lecturas (preferiblemente cortas, por razones económicas) para mapear sobre la referencia y detectar los SNP (polimorfismos mononucleotídicos) mediante el algoritmo más apropiado.

Un área en continua evolución

Resulta evidente, por tanto, la trascendencia de las nuevas tecnologías de secuenciación en multitud de técnicas ya existentes y en otras tantas impensables hasta hace apenas unos años. Las NGS seguirán evolucionando, y con ellas todas sus aplicaciones, no solo en la biología, sino en todas las ciencias de la vida.

Bibliografía citada:

1. Bautista R. Las tres generaciones de la secuenciación. Encuentros en Biología. Vol. 3, 128: 27-28 (2010).
2. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods 4: 651–657 (2007).
3. Zhang Y, Jeltsch A. The aplication of next generation sequencing in DNA methylation analysis. Genes 1: 85-101 (2010).
4. Egan AN, Schlueter J, Spooner DM. Applications of next-generation sequencing in plant biology. American Journal of Botany 99: 175-185 (2012).