ESTUDIO DEL TRANSCRIPTOMA MEDIANTE RNA-SEQ CON ÉNFASIS EN LAS ESPECIES VEGETALES NO MODELO*

Gustavo Rodríguez-Alonso[1], Svetlana Shishkova[1,2]

[1]Departamento de Biología Molecular de Plantas, Instituto de Biotecnología, Universidad Nacional Autónoma de México. Avenida Universidad 2001, Colonia Chamilpa, Cuernavaca, Morelos 62210, México
[2]Autor de correspondencia correo E: sveta@ibt.unam.mx

RESUMEN

Históricamente el desarrollo de la biología se ha limitado al estudio de unas pocas especies, a las cuales conocemos como especies modelo, y cuyas particularidades representan ventajas prácticas para su mantenimiento y estudio en el laboratorio. Sin embargo, la viabilidad actual para secuenciar el genoma y los transcriptomas de virtualmente cualquier especie permite la inclusión de nuevos organismos como modelo de estudio. En esta revisión se presenta una descripción general de las principales plataformas de secuenciación de transcriptomas (RNA-seq), así como los pasos básicos para el ensamblaje de transcriptomas cuando no se cuenta con un genoma de referencia. Finalmente, se proveen algunos ejemplos de estudios de transcriptoma aplicados a organismos no modelo.

PALABRAS CLAVE:
Transcriptoma, RNA-seq, Organismos no modelo, Illumina, Secuenciación masiva.

ABSTRACT

The advance of biology has largely relied on the detailed study of a small number of species collectively known as model organisms. These species were chosen as experimental models due to their peculiarities, which represent practical advantages for their study in laboratory conditions. However, the improvement of sequencing technologies allows nowadays the genome and transcriptome sequencing of virtually any organism without the need of a reference genome, and therefore, new species can be used now to study biological processes that were previously inaccessible from the study of classic model organisms. In this review, we present a general description of the main sequencing technologies that can be used for transcriptome sequencing (RNA-seq) and provide an overview of transcriptome assembly and analysis when no reference genome is available. Finally, some examples of transcriptome analysis applied to non-model organisms are provided.

KEY WORDS:
Transcriptome, RNA-seq, Non-model organism, Illumina, Next Generation Sequencing.

INTRODUCCIÓN

El desarrollo de la biología como ciencia experimental se ha valido del uso de distintos organismos, los cuales, debido a diversas particularidades, se utilizan rutinariamente como objeto de investigación. En 1929 August Krogh pronosticó "para un gran número de problemas, habrá un animal de elección, o unos pocos animales, a los cuales convenga estudiar" (1). A los organismos o especies elegidos para su estudio detallado los conocemos como organismos modelo. Los organismos modelo, que pueden ser procarióticos o eucarióticos, se estudian de manera extensa para generar inferencias y generalidades sobre diversos procesos biológicos. Krogh sugería que la adopción de una especie eucariótica como modelo de estudio debía obedecer a la idoneidad de ésta para el problema a estudiar. Más adelante, en 1975, Hans Krebs expandió la idea de Krogh al añadir que los organismos modelo deben tener un tamaño adecuado para su trabajo en el laboratorio, así como un arreglo anatómico tal que facilite el trabajo experimental (2). Actualmente, la selección de organismos modelo obedece a razones prácticas para su manejo o estudio (3); por ejemplo, las especies eucarióticas que se eligen

*Recibido: 7 de mayo de 2018 Aceptado: 11 de septiembre de 2018

como modelo de estudio se pueden mantener y propagar fácilmente en el laboratorio, tienen ciclos generacionales cortos y alta fecundidad, son especies diploides, lo cual facilita la obtención y el análisis de mutantes; en muchos casos tienen genoma nuclear pequeño, y existen protocolos establecidos para su manipulación genética.

En el caso particular de las plantas, *Arabidopsis thaliana* fue utilizada como modelo de estudio por Friedrich Laibach desde inicios de 1900 (4), sin embargo, su adopción masiva como organismo modelo sucedió hasta principios de 1980, cuando se generalizó su uso al reconocer las ventajas de esta especie para los estudios genéticos (4-6). El estudio de otras especies vegetales, tales como el arroz (*Oryza sativa*), el maíz (*Zea mays*), o la soya (*Glycine max*) ha recibido atención y financiamiento debido a la relevancia económica de su cultivo, pese a que las características genómicas de estas especies distan de las características genómicas ideales de un organismo modelo; por ejemplo, las especies mencionadas son recalcitrantes a la transformación genética; es decir, la obtención de plantas transgénicas de estas especies resulta complicada.

Aunque la adopción de organismos modelo para la investigación ha proporcionado información valiosa acerca de procesos celulares y moleculares fundamentales, las preguntas que pueden ser abordadas con organismos modelo se encuentran inherentemente limitadas por la información que puede recabarse a partir de ellos (7, 8). Las especies modelo representan un porcentaje insignificante entre las 11 millones de especies, aproximadamente, que se estima que existen en el planeta (9). Adicionalmente, no es posible estudiar todos los procesos biológicos a partir del análisis de las especies modelo; por ejemplo, *A. thaliana* no establece interacciones simbióticas con hongos micorrízicos (10) ni con bacterias fijadoras de nitrógeno (11), las cuales son de relevancia agroeconómica y biogeoquímica (12, 13). Debido a esta limitante es necesaria la inclusión de especies no modelo en la actividad científica, es decir, de especies poco estudiadas y que no necesariamente cumplen con todas las características de los organismos modelo, pero que representan sistemas adecuados para el estudio de procesos biológicos inaccesibles mediante el estudio de los organismos modelo actuales (14,15). La adopción de organismos no modelo es posible, en gran medida, gracias a que los costos de secuenciación han disminuido considerablemente, por lo cual es viable secuenciar el genoma y los transcriptomas de virtualmente cualquier organismo pluricelular o de microorganismos cultivables *in vitro*. Adicionalmente, el desarrollo tecnológico y

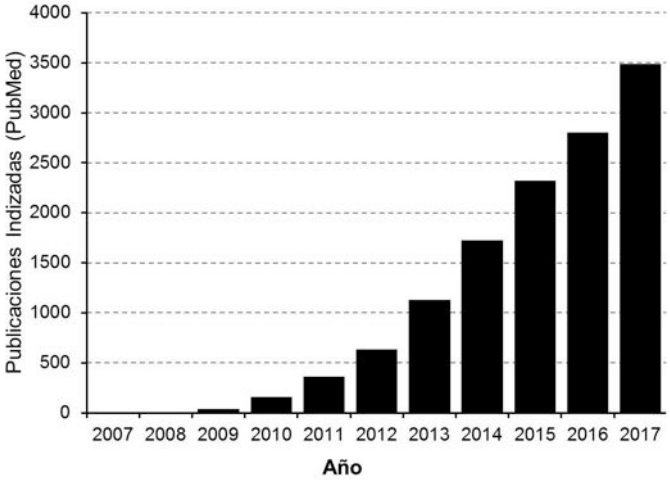

Figura 1. *Aumento en el número de publicaciones que utilizan RNA-seq y que se encuentran indizadas por PubMed. La búsqueda se realizó al restringir a las publicaciones a estudios de expresión génica que incluyen "RNA-seq" en el título o el resumen.*

el aumento de poder de cómputo hacen factible el análisis de transcriptomas en ordenadores personales. Una muestra clara de la adopción y relevancia de la secuenciación como herramienta de estudio es el número creciente de genomas secuenciados depositados en el NCBI (16), cifra que hasta febrero de 2018 asciende a más de 5,000 genomas de eucariontes y aproximadamente 127,000 genomas de procariontes; así como el número de publicaciones indizadas por PubMed sobre secuenciación masiva de cDNA sintetizado a partir de RNA (RNA-seq; Fig. 1). El RNA-seq es una herramienta que permite evaluar el estado transcripcional de un organismo, órgano o tejido, típicamente para comparar distintos tratamientos, condiciones, o estadios de desarrollo (17, 18). A diferencia de otros métodos de transcriptómica que se basan en la hibridación de un conjunto de moléculas de RNA marcadas, tales como los microarreglos, el RNA-seq es un método cuantitativo que no requiere del conocimiento *ab initio* de las secuencias de los RNA mensajeros. Adicionalmente, el RNA-seq permite detectar la transcripción de genes que se expresan a niveles bajos, además de polimorfismos de una sola base e isoformas, es decir, variantes de transcritos que se obtienen mediante el procesamiento diferencial de pre mensajeros provenientes del mismo gen (19).

Generalidades y conceptos básicos del RNA-seq

En general, los estudios que utilizan al RNA-seq siguen una serie de pasos comunes (Fig. 2), los cuales inician con la extracción de RNA total a

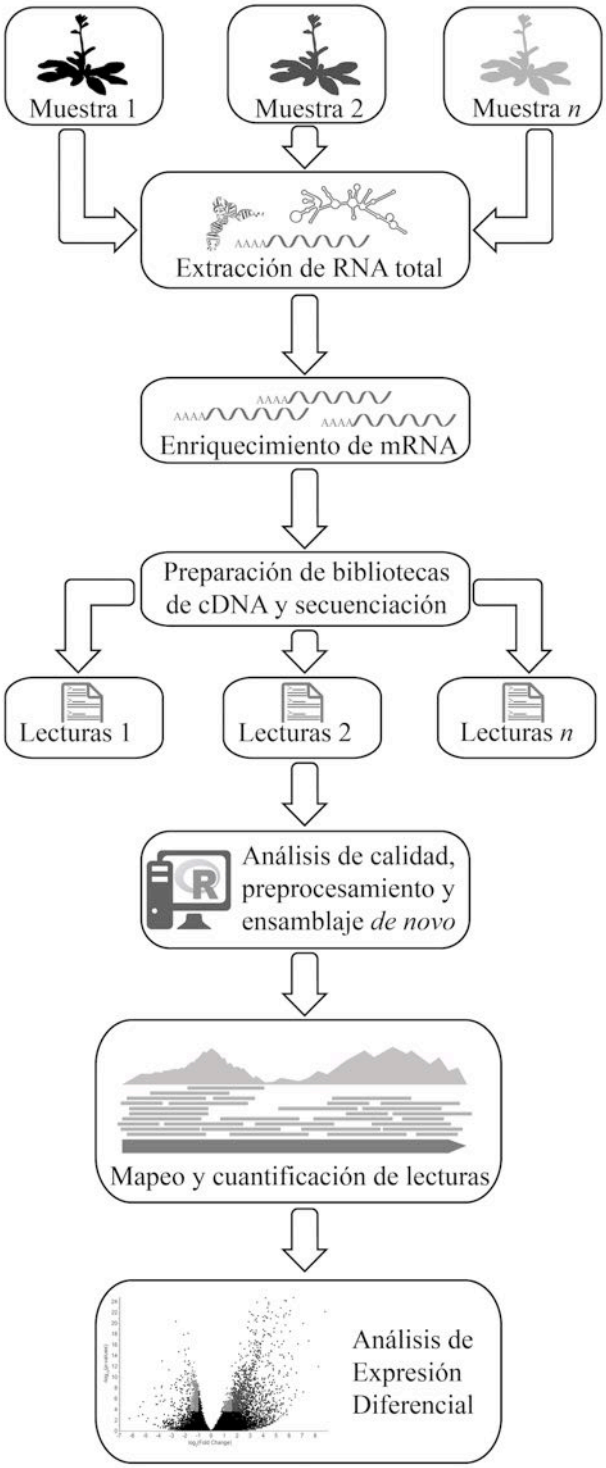

Figura 1. *Diagrama de flujo de un experimento que emplea la secuenciación del transcriptoma mediante RNA-seq. Las muestras se procesan para extraer el mRNA, a partir del cual se prepararán y secuenciarán las bibliotecas de cDNA en alguna de las plataformas de secuenciación disponibles. Las lecturas que se obtienen se procesan, y en el caso de las especies sin genoma de referencia, se ensamblan de novo para reconstruir la secuencia completa de los transcritos y estimar sus niveles de abundancia en cada condición. A partir de los niveles de abundancia es posible realizar análisis de expresión diferencial y proceder hacia los objetivos particulares de cada estudio.*

partir de las muestras biológicas de interés. La extracción se realiza mediante el método de conveniencia para cada especie. En el caso de las plantas, por ejemplo, el consorcio OneKP pone a disposición una serie de protocolos estandarizados para una gran variedad de familias y géneros (20). Dado que en cualquier transcriptoma el RNA ribosomal constituye a la mayor parte del RNA total, existen dos métodos principales de enriquecimiento de RNA mensajero (mRNA), estos son la selección positiva de los mRNA mediante la captura de RNA poliadenilado ("polyA-capture") y el enriquecimiento de mRNA mediante la eliminación de RNA ribosomal ("rRNA-depletion") (21). En general, la técnica más utilizada es la selección de RNAs poliadenilados pues los mRNAs eucarióticos, con muy pocas excepciones, son de este tipo. Sin embargo, el enriquecimiento de mRNA mediante la eliminación de RNA ribosomal es útil en procariontes, cuyos mRNAs no se poliadenilan; y en sistemas eucarióticos en casos muy particulares.

Una vez que el mRNA se enriqueció en la muestra, se procede a la preparación de bibliotecas ("*libraries*") para secuenciación. El mRNA se utiliza para la síntesis de ambas cadenas de cDNA, cuya secuencia permite deducir la secuencia de nucleótidos del mRNA original. Según la plataforma de secuenciación, las cuales se discutirán más adelante, se puede secuenciar a las moléculas completas de cDNA, o bien, se pueden fragmentar mediante métodos mecánicos y realizar una selección de fragmentos por tamaño, usualmente de entre de 200-1,000 nucleótidos. La secuenciación de los fragmentos cortos es mucho más barata, y debido a ello es la técnica más utilizada. A las secuencias cortas que se obtienen a partir de estos fragmentos de cDNA se les conoce como lecturas ("*reads*"), las cuales se pueden generar en los formatos de lecturas simples ("*single-end*") cuando la secuenciación se realiza sólo en alguno de los extremos de cada fragmento, o bien como lecturas apareadas ("*paired-ends*") cuando a partir del mismo fragmento se secuencian los dos extremos. Ambos formatos de lectura permiten cuantificar los niveles de abundancia de transcritos. Sin embargo, cuando no se cuenta con un genoma de referencia, las lecturas apareadas y más largas conducen a un ensamblaje *de novo* de mejor calidad, esto es, al traslape de lecturas en un mayor número de transcritos probables ("*contigs*") y de mayor tamaño. Cuando existe un genoma de referencia, las lecturas apareadas permiten

además detectar isoformas o eventos de empalme ("*splicing*") alternativo de RNAs mensajeros. Una vez que se ensamblaron los contigs, las lecturas se mapean sobre ellos y se cuantifican; al normalizar el número de lecturas mapeadas sobre cada transcrito con base en la longitud de cada contig y el tamaño de la biblioteca de secuenciación, se puede calcular la abundancia relativa de cada transcrito en las muestras biológicas que se secuenciaron. Los análisis posteriores al ensamblaje del transcriptoma estarán determinados por los objetivos particulares de cada estudio, aunque el uso más común del RNA-seq es la detección de expresión diferencial entre muestras biológicas.

Tecnologías de secuenciación más comunes

Actualmente las tres tecnologías principales para la secuenciación de transcriptomas son: la secuenciación por síntesis, la secuenciación por semiconducción y la secuenciación en tiempo real de molécula única. De éstas, la secuenciación por síntesis, desarrollada por la compañía Solexa y comercializada por la compañía Illumina, es la más utilizada (22). Las bibliotecas para la secuenciación por síntesis se preparan a partir de fragmentos de cDNA, cada uno de los cuales se liga a secuencias adaptadoras (Fig. 3). Los adaptadores fijan a los fragmentos en un arreglo espaciado y ordenado

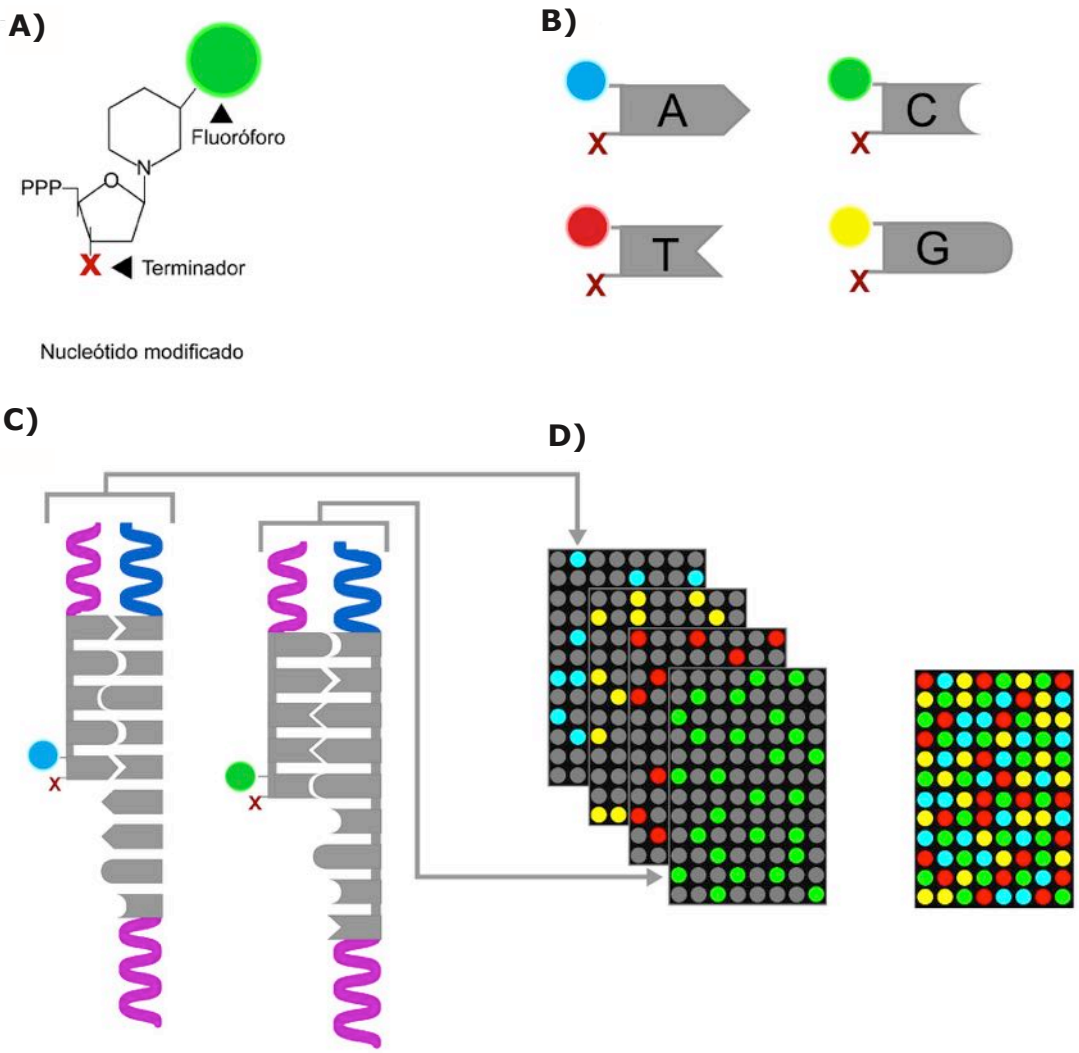

Figura 3. *Secuenciación por síntesis. La tecnología de secuenciación por síntesis, desarrollada por la compañía Solexa, que actualmente pertenece a Illumina, utiliza nucleótidos modificados (A) que contienen una marca fluorescente distintiva para cada base (B) y un terminador en la posición 3' del nucleótido. El terminador limita cada ciclo de síntesis a la adición de un solo nucleótido por fragmento a secuenciar (C). Después de que el nucleótido se agregó a la cadena naciente, se capturan cuatro imágenes (D) para identificar al nucleótido que se agregó en cada cluster. Una vez que se resuelven los nucleótidos adicionados en un ciclo de secuenciación en cada cluster, el fluoróforo, así como el bloqueo del sitio 3' se escinden químicamente y se procede a un nuevo ciclo de secuenciación.*

sobre una superficie de secuenciación, o chip. Cada fragmento está separado de otros espacialmente y se amplifica mediante una reacción de PCR de modo que se generan grupos discretos ("*clusters*") de copias de la misma molécula, lo cual permite detectar más fácilmente a las señales provenientes de cada cluster. La secuenciación se realiza mediante la adición consecutiva de 3'-O-azidometil-dNTPs, los cuales son nucleótidos modificados que incluyen un fluoróforo específico para cada base nitrogenada y una modificación química en la posición 3', de modo que la reacción de síntesis se detiene con la adición de cada nucleótido (23). En cada paso de adición de nucleótidos los fluoróforos se excitan mediante un láser y se capturan cuatro imágenes, cada una en la longitud de emisión de un fluoróforo particular; así en cada ciclo de secuenciación se detecta la señal de cada cluster en sólo una de las cuatro imágenes. De este modo se puede identificar, o resolver, al nucleótido que se incorpora en cada cluster durante cada ciclo de secuenciación. Una vez que las imágenes se capturaron, tanto el fluoróforo como el bloqueo del nucleótido modificado se remueven y se puede realizar un nuevo ciclo de síntesis. En teoría, el número de ciclos de secuenciación puede ser ilimitado, sin embargo, la adición de nucleótidos, la remoción de fluoróforos o del bloqueo no siempre es homogénea dentro de un cluster. Como resultado, la certeza de lectura de bases decrece de manera inversamente proporcional al número de ciclos de secuenciación, por lo cual los valores de calidad de las lecturas decaen después de varios ciclos de secuenciación. Actualmente se pueden obtener datos confiables hasta de 300 ciclos de síntesis, es decir, en el formato de lecturas apareadas se pueden secuenciar hasta 600 nucleótidos del mismo fragmento.

Las bibliotecas de secuenciación por semiconducción también se preparan a partir de cDNA, ligado mediante adaptadores a una esfera sobre la cual se realizará la amplificación clonal para generar un cluster (Fig. 4). Sin embargo, esta tecnología de secuenciación tiene como base un principio químico distinto al de la secuenciación por síntesis. Durante la síntesis de ácidos nucleicos, la reacción de ataque nucleofílico para la adición de un nucleótido a la cadena naciente libera a un protón como producto. Este principio se aprovecha en las plataformas de secuenciación por semiconducción (24), desarrolladas por IonTorrent, las cuales consisten en un transistor de efecto de campo sensible a iones que está acoplado a un circuito integrado (25). El chip de secuenciación es un arreglo de micropozos que contienen a las esferas con los clusters de cDNA; el transistor asocia un cambio de voltaje (ΔV) al cambio de pH (ΔpH) generado por la liberación de protones debido a la adición de nucleótidos. Dado que la liberación de protones es inherente a la reacción de adición de los nucleótidos, éstos no están marcados con fluorescencia ni incluyen un terminador de la reacción de síntesis. La secuencia de los fragmentos se resuelve mediante la inyección de cada tipo de nucleótido de forma secuencial y al observar la presencia o ausencia de cambios de voltaje en cada cluster. En regiones en las cuales existen repeticiones de una sola base se puede estimar el número de nucleótidos repetidos a partir de la magnitud del cambio de voltaje, pues éste es proporcional al número de protones liberados. Actualmente el intervalo dinámico para la detección de repeticiones de una sola base es una limitante técnica de este tipo de secuenciación, sin embargo, al no requerir nucleótidos modificados, los costos de secuenciación son más bajos que en la plataforma Illumina, además, se pueden obtener lecturas de hasta 700 nucleótidos (26). Una ventaja adicional de la secuenciación por semiconducción es que dado que el sistema forma parte de un circuito integrado, las dimensiones de los secuenciadores pueden reducirse hasta dispositivos portátiles, proporcionando así la posibilidad de realizar secuenciación *in situ*.

Por último, la secuenciación en tiempo real de molécula única o SMRT-seq, por sus siglas en inglés (Single Molecule Real Time sequencing; Fig.5), desarrollada por Pacific Biosciences (PacBio), es la tecnología que permite obtener secuencias más largas, con un intervalo de tamaño entre 40 y 60,000 nucleótidos (27). A diferencia de la secuenciación por síntesis o por semiconducción, la secuenciación de molécula única en tiempo real no requiere generar clusters de secuenciación. El fragmento a secuenciar se liga a un adaptador con estructura secundaria de tallo-asa en cada extremo del cDNA (Fig. 5 A) con lo cual, al desnaturalizarse, el fragmento a secuenciar queda circularizado convirtiéndose en una molécula de cadena sencilla. La molécula circular es reconocida por una DNA polimerasa que la utilizará en rondas de síntesis consecutivas e ininterrumpidas. La DNA polimerasa se encuentra fija en el fondo de un nanopozo, formalmente conocido como guía de onda de modo cero, de 70 nm de diámetro y 100 nm de altura. La solución de reacción incluye a los nucleótidos ligados a un fluoróforo en el extremo trifosfatado (28), por lo cual el fluoróforo se escinde tras la formación del enlace fosfodiester durante la síntesis del DNA. Los nucleótidos marcados, así como los fluoróforos liberados, difunden libremente en la guía de onda. La secuencia se obtiene al excitar a los fluoróforos desde el fondo del pozo con un láser y registrar la emisión de fluorescencia del nucleótido que se está

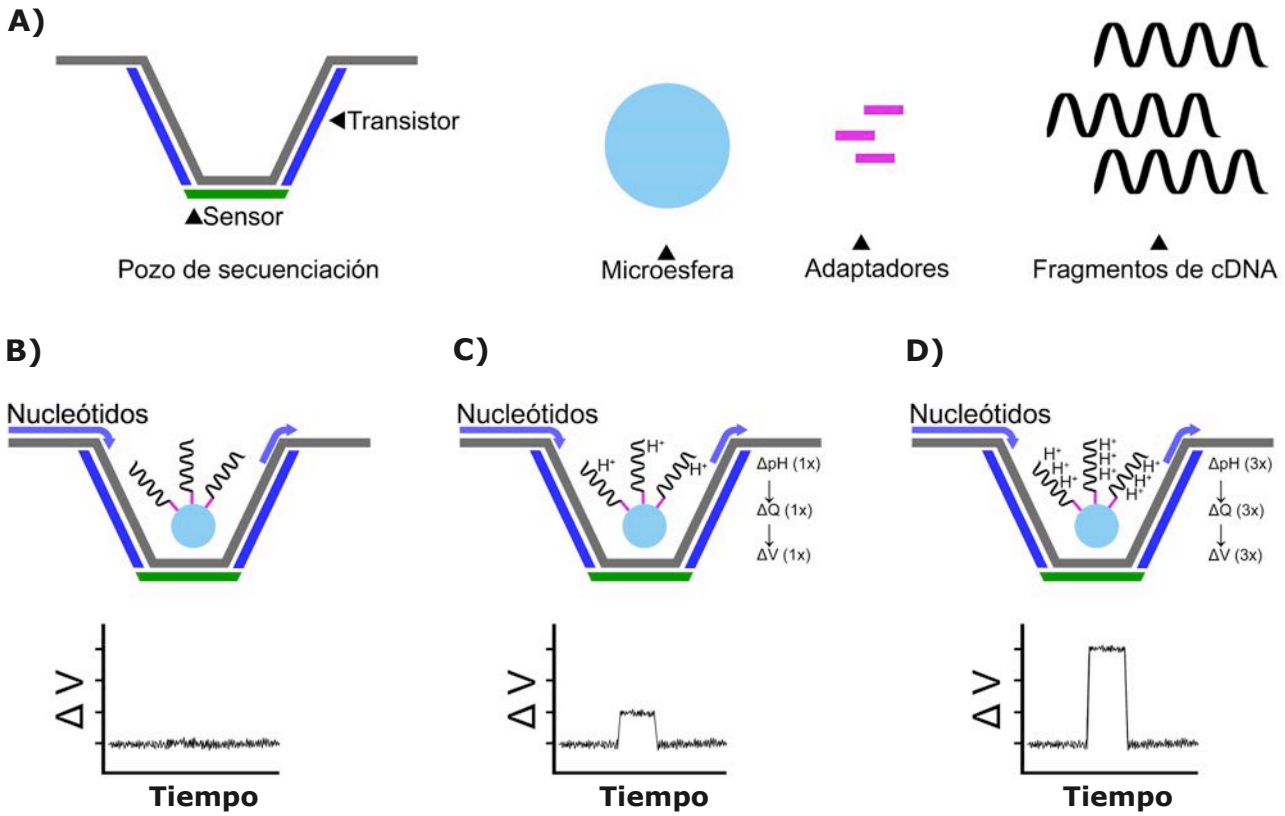

Figura 4. *Secuenciación por semiconducción. Esta secuenciación de la compañía IonTorrent, utiliza transistores de campo sensible a iones en los nanopozos de secuenciación (A). La secuenciación ocurre mediante la inyección controlada de cada nucleótido de forma secuencial a los micropozos y al registrar el cambio de voltaje (ΔV) asociado al cambio de pH (ΔpH) debido a la liberación de protones. Si en un pozo durante un ciclo no se agregan nucleótidos a la cadena naciente, no se registra un ΔV (B), a diferencia de aquellos pozos en los cuales un nucleótido es adicionado (C). Cuando en un micropozo a la cadena naciente del DNA se agrega más de un nucleótido debido a repeticiones continuas de la misma base en la secuencia, el ΔV será proporcional al número de nucleótidos adicionados (D). Después de haber registrado la presencia o ausencia del ΔV en un ciclo, se lava a las moléculas del nucleótido que no fueron adicionados y el ciclo puede repetirse mediante la inyección de un nucleótido diferente.*

adicionando. Debido a las dimensiones de la guía de onda, el volumen observacional está restringido a 20 zeptolitros ($20x10^{-21}$ L), por lo cual la fluorescencia registrada corresponde al nucleótido contenido en el sitio activo de la polimerasa (29). En general, los resultados que se obtienen mediante SMRT-seq son de buena calidad debido a que cada fragmento se secuencia de manera continua múltiples veces al estar circularizado; lo cual permite deducir la secuencia del mRNA original a partir del consenso que se obtiene al eliminar la secuencia de los adaptadores y alinear a las sublecturas resultantes (Fig. 5D).

Consideraciones para el diseño experimental de estudios que involucran al RNA-seq

Aunque las aplicaciones del RNA-seq son diversas, existen consideraciones básicas que deben tomarse en cuenta durante la planeación de experimentos de RNA-seq para optimizar la robustez y fiabilidad de los resultados que se derivan a partir de ellos. El proyecto *ENC*yclopedia *O*f *D*NA *E*lements (ENCODE, 30) recomienda incluir al menos dos réplicas biológicas de cada condición, aunque algunos estudios sugieren que la robustez estadística de los resultados se alcanza con al menos seis réplicas biológicas (31). Por su parte, las réplicas técnicas de secuenciación no se requieren pues la variabilidad asociada a las plataformas de secuenciación es mínima (32).

A diferencia del genoma, el transcriptoma es dinámico y difiere entre una condición y otra, o entre distintos órganos. Por ello, y debido a que la calidad del ensamblaje depende de la cantidad y calidad de lecturas obtenidas, se recomienda estimar la profundidad de secuenciación según los objetivos de cada experimento (33). Obtener

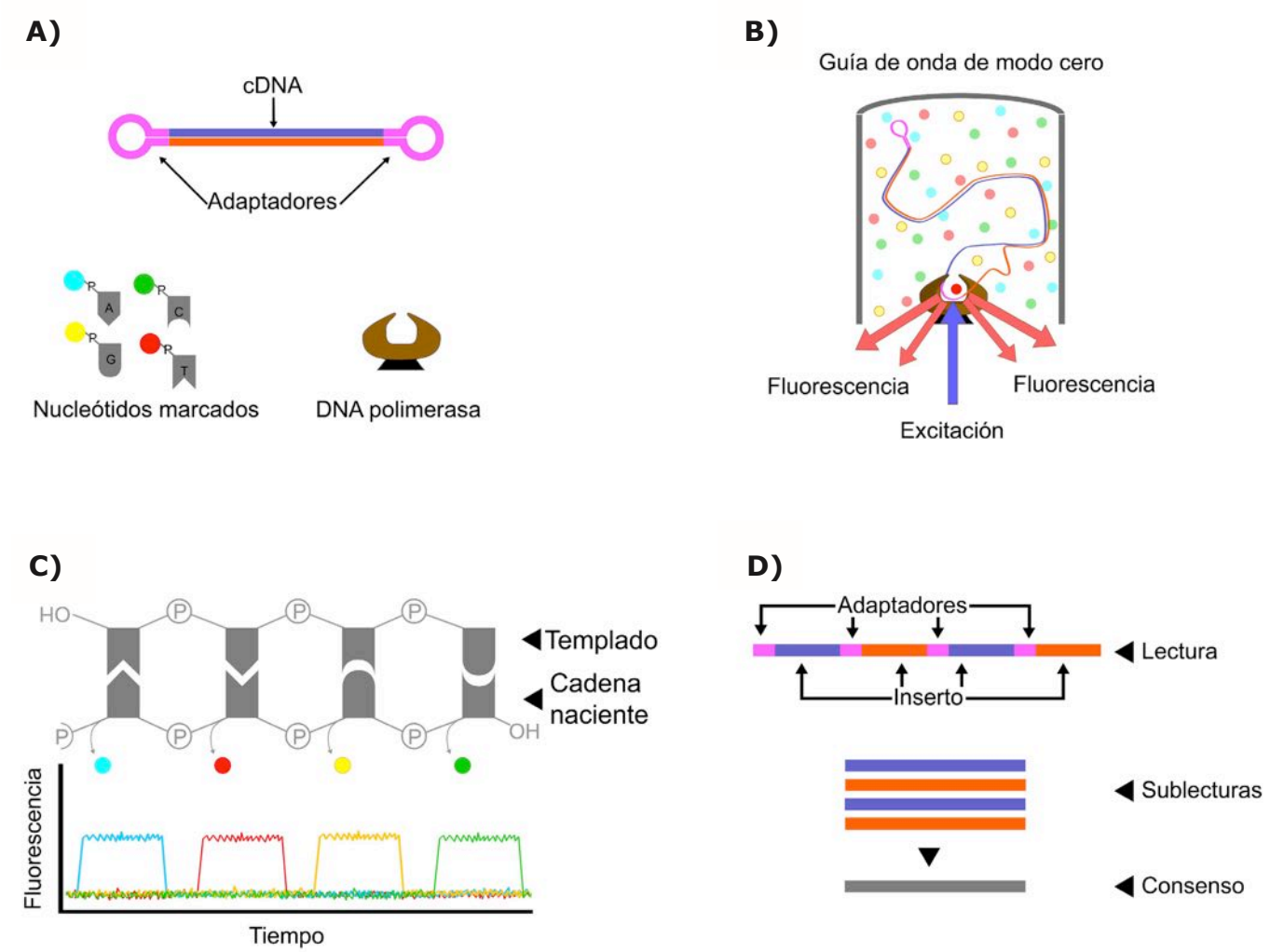

Figura 5. *Secuenciación de molécula única en tiempo real (SMRT sequencing). La tecnología de secuenciación SMRT utiliza nucleótidos con marcas fluorescentes en el extremo trifosfatado y adaptadores que circularizan al fragmento de cDNA a secuenciar (A). La secuenciación ocurre en un nanopozo, o guía de onda de modo cero, en el fondo del cual está inmovilizada una DNA polimerasa. La secuencia se obtiene al excitar con láser, desde el fondo del pozo, a los fluoróforos que marcan a cada nucleótido. Debido a que la DNA polimerasa inmoviliza al nucleótido modificado mientras éste se incorpora a la cadena naciente, el fluoróforo de este nucleótido emitirá fluorescencia durante más tiempo que los nucleótidos que difunden libremente en la solución (B y C). Al ser una molécula circular, el mismo fragmento de cDNA se secuencia varias veces de forma ininterrumpida dentro del nanopozo; dado que las secuencias de los adaptadores se conocen, son fácilmente identificables y pueden ser removidas. Las sublecturas generadas después de la remoción de adaptadores se pueden alinear para obtener una secuencia consenso, la cual corresponderá a la secuencia de cDNA derivado del mRNA original (D).*

al menos 200 millones de lecturas mapeables permite detectar transcritos poco abundantes o eventos raros de empalme alternativo (34).

Finalmente, una práctica recomendable pero no del todo extendida es la inclusión de estándares cuantitativos ("*spike-in*") en las muestras de RNA. Estos estándares son moléculas de RNA exógeno, de secuencia conocida, que se agregan a las muestras en una concentración determinada durante cualquier paso experimental. El ERCC (External RNA Control Consortium) es un consorcio internacional que provee de 92 RNAs sintéticos con un intervalo de tamaño de 250 a 2500 nucleótidos y del 5 al 51% de contenido de GC (35). El uso de estándares cuantitativos provee de una manera directa para estimar las tasas de error de lectura de bases y los sesgos de cobertura (36, 37). Adicionalmente, es posible contar con controles positivos y negativos en los análisis de expresión diferencial al agregar el *spike-in* en concentraciones significativamente distintas o iguales, respectivamente.

Ensamblaje *de novo* de transcriptomas de organismos sin genoma de referencia

Las lecturas que se obtienen mediante las plataformas de secuenciación deben ser preprocesadas para identificar y remover a los sitios no informativos. Estos incluyen a las secuencias adaptadoras de la plataforma de secuenciación, bases ambiguas y bases no resueltas. La remoción de sitios no informativos resultará en lecturas de diversos tamaños, a partir de las cuales es necesario descartar también a las lecturas muy cortas, pues éstas son más propensas a ser mapeadas en más de un sitio. Algunas herramientas bioinformáticas útiles y de acceso abierto para el preprocesamiento de las lecturas son FastQC, Trimmomatic y FastX (38-40). En el caso de las lecturas apareadas, es necesario fusionar a las lecturas que provienen del mismo fragmento cuando éstas se superponen en sus extremos (41).

Una vez que las lecturas han sido preprocesadas, se procede al ensamblaje del transcriptoma. Los algoritmos más comunes para el ensamblaje *de novo* se basan en gráficas de de *Bruijn*. Entre los algoritmos disponibles destaca Trinity (42) como el paquete más utilizado debido a su flexibilidad y el manejo eficiente de los recursos computacionales. De manera general, los ensambladores descomponen virtualmente a las lecturas de secuenciación en fragmentos pequeños, conocidos como palabras o k-meros, los cuales se utilizan en rondas de alineamiento para encontrar a los k-meros que se superponen en los extremos 5' o 3' (Fig. 6). De esta manera, a partir de un k-mero inicial ("*seed*") elegido aleatoriamente, se empieza el ensamblaje de los contigs y se continúa hasta que la secuencia ya no puede ser extendida. Cuando para alguno de los extremos se obtienen dos posibles k-meros que pueden ser agregados, se abre una burbuja ("*bubble*") y se continúa la extensión del contig en las dos posibles variantes, hasta que esta burbuja se cierra, o colapsa, debido a la existencia de un solo k-mero para ambas posibilidades. Cada burbuja duplica el número de variantes del contig, las cuales pueden representar isoformas, variantes alélicas o errores de secuenciación que resultan en artefactos del ensamblaje. Los tamaños de k-mero y burbuja pueden ser ajustados a conveniencia. Idealmente, deben realizarse varios ensamblajes al variar éstos dos parámetros hasta encontrar al "mejor" ensamblado del transcriptoma, el cual estará definido por la generación de contigs de mayor tamaño, con la mayor proporción de lecturas mapeadas en pares, y con un número bajo de lecturas que mapean a más de un sitio.

Anotación del transcriptoma

Una vez que las lecturas se han fusionado en contigs, es necesario asignar a cada uno de éstos una identidad y función putativa. A este proceso se le conoce como anotación y se basa principalmente en la búsqueda por homología de secuencias caracterizadas y depositadas en bases de datos tales como RefSeq (43). En general, la anotación requiere de la identificación de los marcos abiertos de lectura presentes en cada contig, una tarea que puede realizarse con programas tales como TransDecoder (44). Posteriormente, la secuencia de aminoácidos traducida a partir del contig se utiliza como sonda para interrogar a las bases de datos e identificar a sus secuencias homólogas. Paqueterías como Blast2GO (45) permiten asignar un conjunto de categorías jerárquicas, u ontologías génicas, a partir de la homología entre la proteína predicha para cada contig y las secuencias depositadas en las bases de datos. Otras herramientas disponibles para la anotación de secuencias son InterProScan (46), el cual identifica firmas proteicas; Annocript (47), el cual permite anotar probables RNAs largos no codificantes; y KAAS (48), el cual mapea a las secuencias proteicas en rutas metabólicas y de señalización.

Actualmente, la anotación de las secuencias de organismos no modelo depende de las secuencias depositadas en bases de datos de acceso público, las cuales a su vez fueron anotadas a partir de organismos modelo. Esto representa una limitante, pues en los organismos no modelo sólo se podrá anotar a aquellas secuencias que contengan a un ortólogo putativo anotado en el genoma de alguna especie modelo. Adicionalmente, las secuencias de RNA que carecen de marco abierto de lectura, tales como los RNAs largos no codificantes o los precursores de microRNAs, serán difíciles de anotar pues presentan tasas de evolución más altas y por lo tanto más divergencia entre especies (49-50). De esta forma, el porcentaje de secuencias no anotadas, así como el número de secuencias con anotaciones de baja calidad, será proporcional a la distancia evolutiva entre el organismo en estudio y el organismo modelo más cercano.

Mapeo de lecturas y evaluación de la expresión diferencial

Uno de los objetivos más comunes de los experimentos de RNA-seq es identificar a los transcritos que se regulan de manera diferencial entre dos condiciones. Existen distintas herramientas gratuitas que permiten mapear lecturas sobre los contigs cuando no se cuenta con un genoma de referen-

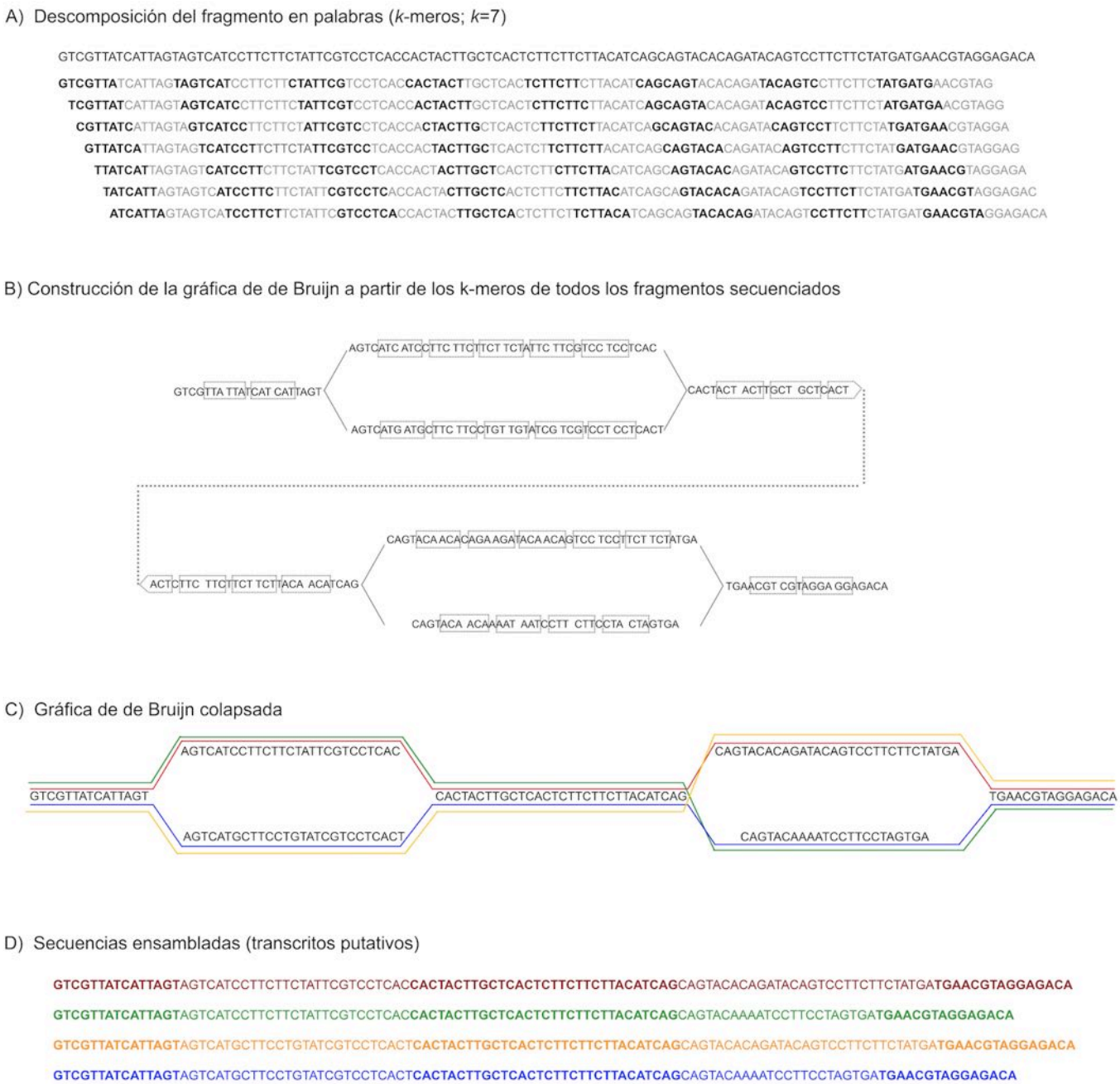

Figura 6. *Ensamblaje de novo del transcriptoma. Las lecturas obtenidas, en este caso a partir de la plataforma Illumina/Solexa, se descomponen virtualmente en palabras, o k-meros, de un tamaño particular definido por el usuario y ejemplificado aquí con un tamaño k=7 (A). Estos k-meros se utilizarán para construir alineamientos que permitan extender a la secuencia hacia los extremos 3' y 5'. Cuando existen dos k-meros que permiten extender el mismo extremo, se abre una burbuja, la cual colapsa mediante la adición de un k-mero común para ambas variantes (B). La gráfica se colapsa al fusionar las regiones sobrelapantes de los k-meros (C) y se leen las posibles variantes ensambladas del transcrito (D).*

cia, entre ellas destaca RSEM (51). Los algoritmos para mapear lecturas sobre los contigs funcionan, básicamente, a partir de alineamientos globales entre las lecturas procesadas y las secuencias de referencia. Cada algoritmo presenta sus propias ventajas y desventajas respecto a los demás, particularmente en la calidad de los alineamientos y la versatilidad para evaluar los costos de las diferencias de un solo nucleótido, o *mismatches*, y la apertura y extensión de *gaps*. Un análisis comparativo de los distintos algoritmos para mapeo de lecturas cortas de RNA-seq (52), provee de la información

necesaria para decidir entre los diferentes algoritmos que existen según los criterios que resulten más relevantes en cada experimento. El mapeo de las lecturas sobre las secuencias de referencia, en este caso los transcritos ensamblados *de novo*, es el primer paso para estimar el nivel de expresión de cada uno de los contigs. Esto se logra al mapear y cuantificar, por separado, a las lecturas que se obtuvieron para cada condición y evaluar si existen diferencias significativas entre el número de lecturas que mapean sobre cada contig en las distintas condiciones. Existen diferentes métodos estadísticos para la evaluación de la expresión diferencial, los cuales pertenecen a los algoritmos paramétricos, cuando éste asume algún tipo de distribución estadística de los datos de expresión; o a los algoritmos no paramétricos, los cuales no ajustan ni asumen ningún tipo de distribución de los datos *a priori*. Los algoritmos paramétricos por lo general asumen una distribución binomial negativa de los datos, lo cual implica que al comparar dos muestras biológicas distintas se espera que la mayor parte de los transcritos no muestren diferencias significativas en su nivel de expresión. Entre los algoritmos paramétricos más utilizados se encuentran edgeR y DESEQ2 (53-54). Por su parte, NOISEQ es el algoritmo no paramétrico más ampliamente utilizado, el cual además permite simular réplicas cuando sólo se tiene una muestra biológica por condición (55). Los umbrales en el nivel de cambio ("*fold change*") de expresión y de valor p ("*p*-value") para definir a un transcrito como diferencialmente expresado se pueden ajustar de acuerdo con el rigor de los análisis; en general se recomienda ser más estricto cuando se dispone de pocas réplicas biológicas y se puede ser más laxo a medida que el número de réplicas biológicas por condición crece. Cabe mencionar que aunque el RNA-seq ha mostrado ser una herramienta muy útil y precisa, cuando se disponen de pocas muestras con pocas réplicas, es aconsejable realizar mediciones del nivel de abundancia de los transcritos por un método independiente, siendo los ensayos de PCR en tiempo real a partir de cDNA (RT-qPCR) el método de elección.

Aplicaciones del RNA-seq

La rápida adopción del RNA-seq como herramienta para el estudio de procesos biológicos permea actualmente en una gran diversidad de disciplinas. En el área de la biología evolutiva, por ejemplo, el análisis del transcriptoma del tomate domesticado y de algunas especies silvestres se utilizó para comparar los patrones de expresión de distintos genes para identificar a aquellos que fueron seleccionados artificialmente durante la domesticación de ésta solanácea, así como los tipos de presión evolutiva que actuaron sobre esta especie durante distintas etapas de su domesticación y diferentes eventos de hibridación con especies silvestres (56). En este ejemplo, el mapeo de las lecturas de RNA se realizó sobre un congénere con genoma secuenciado, es decir, sobre el genoma de referencia de una especie perteneciente al mismo género taxonómico. El conjunto de lecturas que no mapearon sobre el genoma de referencia se utilizaron para realizar un ensamblaje *de novo* e identificar transcritos putativos específicos de las especies silvestres.
Por otro lado, proyectos colaborativos como el OneKP, en el cual se secuenciaron los transcriptomas de más de 1,300 especies pertenecientes a distintas familias de plantas, han permitido realizar análisis de transcriptómica comparativa para obtener filogenias moleculares con mejor resolución (57, 58); dilucidar el origen evolutivo, por ejemplo, de algunos receptores de hormonas vegetales (59) y de las desacetilasas de histonas (60) mediante la neofuncionalización de genes; analizar la divergencia de regiones regulatorias para la subfuncionalización y neofuncionalización de factores transcripcionales y su importancia en la remodelación de redes de regulación genética (61), entre otros.

Adicionalmente, el RNA-seq se ha empleado como herramienta para la exploración de procesos bioquímicos tales como la síntesis de metabolitos secundarios, por ejemplo, de la planta del té (*Camellia sinensis*; 62) o del peyote (*Lophophora williiamsii*; 63); así como la caracterización del gametofito de los helechos (64), la caracterización de interacciones planta-patógeno (65), o la exploración de programas particulares del desarrollo, tales como el agotamiento del meristemo apical de la raíz en cactáceas (66).

El rápido desarrollo de las tecnologías de secuenciación, y de procesamiento de información masiva, permite actualmente que el RNA-seq ofrezca oportunidades diversas y atractivas en distintas áreas de la ciencia. Por ejemplo, la combinación del RNA-seq y de las técnicas histológicas permite realizar análisis de transcriptómica espacial para generar atlas de expresión con resolución a nivel de tejido (67). Por otro lado, la combinación con la técnica de separación de células asistida por dinámica de fluidos (FACS) permite secuenciar el transcriptoma de un solo tipo celular o de células únicas (68). La adopción de estas tecnologías permea ya en el estudio de diferentes especies modelo, y más lentamente, en especies cuyo genoma aún no está secuenciado, con lo cual se ampliará el estudio de especies no modelo y en consecuencia,

se fortalecerá nuestro conocimiento de las distintas especies, sus procesos metabólicos y moleculares.

Agradecimiento. El trabajo de los autores sobre el análisis de transcriptoma de *Pachycereus pringlei*, una especie vegetal no modelo, está parcialmente financiado por proyectos CONACyT 240055 y PAPIIT-UNAM IN201318. A Gustavo Rodríguez-Alonso se le otorgó una beca doctoral del CONACyT (registro 290654).

REFERENCIAS

1. Krogh A (1929) The progress of physiology. Science 70:200-204.
2. Krebs HA (1975) The August Krogh principle: "for many problems there is an animal on which it can be most conveniently studied". J Exp Zool 194:221-226.
3. Russell JJ, Theriot JA, Sood P, Marshall WF, Landweber LF, Fritz-Laylin L, Polka JK, Oliferenko S, Gerbich T, Gladfelter A, Umen J, Bezanilla M, Lancaster MA, He S, Gibson MC, Goldstein B, Tanaka EM, Hu CK, Brunet A (2017) Non-model model organisms. BMC Biol 15:55.
4. Koornneef M, Meinke D (2010) The development of Arabidopsis as a model plant. Plant J 61:909-921.
5. Rédei GP (1975) Arabidopsis as a genetic tool. Annu Rev Genet 9:111-127.
6. Somervill C, Kornneef M (2002) A fortunate choice: the history of Arabidopsis as a model plant. Nat Rev Genet 3:883-889.
7. Bolker J (2012) There's more to life than rats and flies. Nature 491:31-33.
8. Alfred J, Baldwin IT (2015) New opportunities at the wild frontier. eLife 4:e06956.
9. Mora C, Tittlensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and the ocean? PLoS Biol 9:e1001127.
10. Bonfante P, Genre A (2010) Mechanisms underlying beneficial plant-fungus interactions in mycorrhizal symbiosis. Nat Commun 1:48.
11. Mylona P, Pawlowski K, Bisseling T (1995) Symbiotic Nitrogen Fixation. The Plant Cell 7:869-885.
12. Read DJ (2003) Towards Ecological Relevance — Progress and pitfalls in the path towards an understanding of mycorrhizal functions in nature. En: Mycorrhizal Ecology. Ecological Studies (Analysis and Synthesis). Editor: van der Heijden MGA, Sanders IR, vol. 157. Springer, Berlin, Heidelberg.
13. Vicente EJ, Dean DR (2017) Keeping the nitrogen-fixation dream alive. Proc Natl Acad Sci USA. 114:3009-3011.
14. Goldstein B, King N (2016) The Future of Cell Biology: Emerging model organisms. Trends Cell Biol. 26:818-824.
15. Gladfelter AS (2015) How nontraditional model systems can save us. Mol. Biol. Cell 26:3687-3689.
16. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. (2012) NCBI GEO: archive for functional genomics data sets - update. Nuc Ac Res 41:D991-D995.
17. Van Verk MC, Hickman R, Pieterse CM. J, Van Wees SCM (2013) RNA-seq: revelation of the messengers. Trends Plant Sci 18:175-179.
18. Wang Z, Gersten M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57-63.
19. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. PLoS One 9:e78644.
20. Johnson MT, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN (2012) Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. PLoS One 7:e50226.
21. Cui P, Lin Q, Ding F, Xin C, Gong W (2010) A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics 96:259-265.
22. Ari Ş, Arikan M (2016) Next-Generation Sequencing: Advantages, disadvantages, and future. En: Plant Omics: Trends and Applications. Editor: Hakeem K, Tombuloğlu H, Tombuloğlu G. Springer, Cham, Switzerland.
23. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray

LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.

24. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348-352.
25. Lee CS, Kim SK, Kim M (2009) Ion-sensitive field-effect transistor for biological sensing. Sensors (Basel) 9:7111-71131.
26. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA (2014) Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol 80:7583-7591.
27. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. Nuc Ac Res 38:e159.
28. Korlach J, Bibillo A, Wegener J, Peluso P, Pham TT, Park I, Clark S, Otto GA, Turner SW (2008) Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. Nucleos Nucleot Nucl 27:1072-1083.
29. Foquet M, Samiee KT, Kong X, Chauduri BP, Lundquist PM, Turner SW, Freudenthal J, Roitman DB (2008) Improved fabrication of zero-mode waveguides for single-molecule detection. J Appl Phys 103:034301.
30. ENCODE project consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. Science 306:636-640.
31. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M, Barton GJ (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22:839-851.
32. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509-1517.
33. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP (2013) Calculating sample size estimates for RNA sequencing data. J Comput Biol 20:970-978.
34. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analysis. Nat. Rev. Genet. 15:121-132.
35. Risso D, Ngai J, Speed TP, Dudoit S (2014) Normalization of RNA-seq data using factor

analysis of control genes or samples. Nat Biotechnol 32:896-902.
36. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B (2011) Synthetic spike-in standards for RNA-seq experiments. Genome Res 21:1543-1551.
37. Munro SA, Lund SP, Pine PS, Binder H, Clevert DA (2014) Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat Commun 5:5125.
38. Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
39. Bolger AM, Lohse M, y Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.
40. Hannon Lab. Cold Spring Harbor Laboratory. http://hannonlab.cshl.edu/fastx_toolkit/
41. Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614-620.
42. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29:644-652.
43. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nuc Ac Res 40:D130-D135.
44. TransDecoder https://github.com/TransDecoder
45. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674-3676.
46. Zdobnov EM, y Apweiler R (2001) InterProScan-- an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847-848.
47. Musacchia F, Basu S, Petrosino G, Salvemini M, Sanges R (2015) Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. Bioinformatics 31:2199-2201.
48. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nuc Ac Res 35:W182 - W185.
49. Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H (2013) Birth and expression evolution of mammalian microRNA genes. Genome Res 23:34-45.
50. Johnsson P, Lipovich L, Grandér D, Morris KV (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta 184:1063-1071.
51. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics. 12:323.
52. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, Bertone P; RGASP Consortium. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Meth 10:1185-1191.
53. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.
54. Love MI, Huber W, Anders S (2014) Moderate estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.
55. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 21:2213-2223.
56. Koenig D, Jiménez-Gómez JM, Kimura S, Fulop D, Chitwood DH, Headland LR, Kumar R, Covington MF, Devisetty UK, Tat AV, Tohge T, Bolger A, Schneeberger K, Ossowski S, Lanz C, Xiong G, Taylor-Teeples M, Brady SM, Pauly M, Weigel D, Usadel B, Fernie AR, Peng J, Sinha NR, Maloof JN (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. Proc Nat Acad Sci USA, 110:E2655-E2662.
57. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, Wu X, Zhang Y, Wang J, Zhang Y, Carpenter EJ, Deyholos MK, Kutchan TM, Chanderbali AS, Soltis PS, Stevenson DW, McCombie R, Pires JC, Wong GK, Soltis DE, Depamphilis CW (2012) A genome triplication associated with early diversification of the core eudicots. Genome Biol 13:R3.
58. Yang Y, Moore M. J, Brockington SF, Soltis DE, Wong GK, Carpenter EJ, Zhang Y, Chen L, Yan

Z, Xie Y, Xie Y, Sage RF, Covshoff S, Hibberd JM, Nelson MN, Smith SA (2016) Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Mol Biol Evol 32:2001-2014.
59. Bythell-Douglas R, Rothfels CJ, Stevenson DWD, Graham SW, Wong GK, Nelson DC, Bennett T (2017) Evolution of strigolactone receptors by gradual neo-functionalization of KAI2 paralogues. BMC Biol. 15:52.
60. Bourque S, Jeandroz S, Grandperret V, Lehotai N, Aimé S, Soltis DE, Miles NW, Melkonian M, Deyholos MK, Leebens-Mack JH, Chase MW, Rothfels CJ, Stevenson DW, Graham SW, Wang X, Wu S, Pires JC, Edger PP, Yan Z, Xie Y, Carpenter EJ, Wong GKS, Wendehenne D, Nicolas-Francès V (2016) The Evolution of HD2 proteins in green plants. Trends Plant Sci. 21:1008-1016.
61. Sayou C, Monniaux M, Nanao MH, Moyroud E, Brockington SF, Thévenon E, Chahtane H, Warthmann N, Melkonian M, Zhang Y, Wong GK, Weigel D, Parcy F, Dumas R (2014) A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science 343:645-648.
62. Chun-Fang L, Yan Z, Yao Y, Qiong-Yi Z, Sheng-Jun W, Xin-Chao W, Ming-Zhe Y, Da L, Xuan L (2015) Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (Camellia sinensis) BMC Genomics 16:560.
63. Ibarra-Laclette E, Zamudio-Hernández F, Pérez-Torres CA, Albert VA, Ramírez-Chávez E, Molina-Torres J, Fernández-Cortes A, Calderón-Vázquez C, Olivares-Romero JL, Herrera-Estrella A, Herrera-Estrella L (2015) De novo sequencing and analysis of Lophophora williamsii transcriptome, and searching for putative genes involved in mescaline biosynthesis. BMC Genomics 16:657.
64. Der JP, Barker MS, Wickett NJ, Depamphilis CW, Wolf PG (2011) De novo characterization of the gametophyte transcriptome in bracken fern, Pteridium aquilinum. BMC Genomics 12:99.
65. Nagano AJ, Honjo MN, Mihara M, Sato M, Kudoh H (2015) Detection of plant viruses in natural environments by using RNA-Seq. Methods Mol Biol 1236:89-98.
66. Rodríguez-Alonso G, Matvienko M, López-Valle ML, Lázaro-Mixteco PE, Napsucialy-Mendivil S, Dubrovsky JG, Shishkova S (2018) Transcriptomics insights into the genetic regulation of root apical meristem exhaustion and determinate primary root growth in Pachycereus pringlei (Cactaceae). Sci Rep 8:8529.
67. Lieben L (2017) Plant genetics: Spatial transcriptomics in plants. Nat Rev Genet 18:394.
68. Tang F, Lao K, Surani MA (2011) Development and applications of single-cell transcriptome analysis. Nat Methods 8:S6-S11.