Las técnicas de secuenciación masiva en el estudio de la diversidad biológica.

Massive sequencing techniques in the study of biological diversity.

Unai López de Heredia

Resumen

Las técnicas de análisis genómico que se basan en la secuenciación masiva han supuesto una revolución en la última década no sólo en biomedicina o agronomía, sino también en el estudio de la diversidad biológica. Las plataformas de secuenciación de segunda y tercera generación que vienen a complementar a la tradicional secuenciación por el método Sanger, permiten analizar un gran número de individuos obteniendo una profundidad de secuenciación significativa de sus genomas o transcriptomas. Así, aunque todavía de manera incipiente, se vienen realizando estudios sobre la flora y fauna silvestre a escala poblacional, con especial énfasis en la determinación de patrones adaptativos frente a los cambios ambientales. En la presente revisión se describe la situación actual de las plataformas de secuenciación masiva, señalando sus ventajas y limitaciones para el análisis en organismos no modelo, para a continuación detallar las bases de dos de las técnicas más populares que se benefician de la secuenciación masiva (RAD-seq y RNA-seq) y señalar algunos ejemplos de su uso para el estudio de la diversidad biológica.

Abstract

High-throughput sequencing based on genomic analysis techniques have not only revolutionized the fields of biomedicine or agronomy, but also the research into biological diversity.

Second and third generation sequencing platforms complement traditional Sanger sequencing, and allow the analysis of many individuals with a significant sequencing depth of their genomes and transcriptomes. Although only at the early stage, a number of studies have been conducted on plants and animals at a population scale, especially to discern the adaptive patterns of the species against environmental changes. This review aims to describe the current status of high-throughput sequencing platforms, highlighting their advantages and limitations for the analysis of non-model organisms and goes on to set out the methodological basis of two of the most popular techniques benefited by high-throughput sequencing (RAD-seq and RNA-seq), providing a number of examples of its use for the analysis of biological diversity.

Laburpena

Sekuenziatzio masiboan oinarritzen diren analisi genomikoko teknikek aurrerakuntza handia ekarri dute azken hamarkadan, ez bakarrik biomedikuntzan edo nekazaritzan, baita aniztasun biologikoaren ikerketan ere. Bigarren eta hirugarren belaunaldiko sekuenziatzio plataformek Sanger metodo tradizionala osatu eta izaki asko aztertzeko aukera ematen dute, horien genomen edo transkriptomen sakoneko sekuenziatzio esanguratsua lortzen delarik. Horrela, oraindik hastapenak badira ere, basa landare eta animali populazioak ardatz dituzten ikerketak egiten hasi dira, arreta berezia jarrita ingurumen aldaketen aurreko patroi adaptatiboen identifikazioan.

Gure berrikuspen honetan sekuenziatzio masiboko plataformen gaurko egoera deskribatzen da, ez-eredu diren izakien azterketarako dituzten abantailak eta eragozpenak aipatuaz; eta, ondoren, sekuenziatzio masiboan gehien erabilitako bi tekniken (RAD-seq eta RNA-seq) oinarri metodologikoak azaldu eta horien erabiltze adibide batzuk ematen dira, aniztasun biologikoaren ikerketa arloan.

Introducción

Los marcadores genéticos moleculares -ver revisión de los mismos en Rentaría Alcántara (2007)- y la secuenciación directa de fragmentos cortos de ADN mediante el método de Sanger[†] *et al.* (1977) han sido, y aún son, las principales herramientas para el estudio de la diversidad biológica. Estas metodologías supusieron en las últimas décadas del siglo XX una auténtica revolución en el estudio de los organismos biológicos. Por primera vez, se pudieron establecer diferencias entre especies, poblaciones e individuos sobre la base del genotipo en lugar de la apariencia externa o fenotipo. El fenotipo es la manifestación ex-

terna de un genotipo, y está condicionada por el efecto del ambiente. Así, las determinaciones fenotípicas requieren de ensayos relativamente complicados para descontar el efecto de éste y su interacción con el genotipo. Con el genotipado, y dada la posibilidad que proporcionan los marcadores moleculares de comparar un gran número de caracteres homólogos independientes del ambiente, se ha facilitado en gran medida la realización de estudios filogenéticos, tanto para delimitar la integridad de las especies y confirmar su estatus taxonómico, como para testar hipótesis evolutivas. Además, los marcadores moleculares han permitido la realización de estudios poblacionales a escala local o regional, ya que permiten acceder al genotipo de un número elevado de individuos. Sin ánimo de ser exhaustivos, algunos ejemplos de las principales aplicaciones de los marcadores genéticos moleculares y la secuenciación de Sanger en el ámbito de la diversidad biológica incluyen la delimitación de especies (Avise, 1994), el análisis de la distribución geográfica de la diversidad genética o filogeografía (Avise, 2000), la determinación de rutas migratorias de aves (Wink *et al.*, 2006) o de re-colonización post-glaciar por parte especies arbóreas (Hewitt, 1999) o la cuantificación del flujo genético y de la estructura genética espacial (Bossart y Prowell, 1998).

La obtención de variantes moleculares para el genotipado de individuos y la determinación de los perfiles de expresión de genes concretos tuvieron un fuerte impulso con el desarrollo de los *biochips* o *microarrays*[†] de ADN y ARN (Hoheisel et al. 2006), que todavía se utilizan con profusión. Los *microarrays* consisten en superficies sólidas con celdillas en las que se depositan una serie de secuencias de nucleótidos de manera ordenada y que se basan en la capacidad de hibridación entre dos cadenas de ADN complementarias entre sí.

Pese a que los *microarrays* permiten genotipar de manera menos sesgada que los marcadores genéticos tradicionales y permiten evaluar la expresión diferencial de miles de genes simultáneamente, presentan algunas complicaciones en su diseño derivadas de la necesidad de conocer a priori las secuencias de interés de los organismos de estudio, algo que en muchos casos no es posible cuando se trabaja con especies no modelo. Así, la siguiente revolución en el estudio de los seres vivos y de las interacciones entre éstos, y de éstos con el medio ha venido de la mano del desarrollo de las plataformas de secuenciación de alto rendimiento, o también llamada en su momento de nueva generación (*Next Generation Sequencing* - NGS-). Muchos procesos biológicos y ecológicos que anteriormente no podían ser abordados por su elevado coste, pueden ser estudiados en la actualidad a través de la secuenciación de genomas y transcriptomas o del genotipado masivo de individuos. Desde el punto de vista del estudio de la diversidad biológica, la secuenciación masiva permite, entre otras aproximaciones, obtener de manera relativamente sencilla un gran número de marcadores genéticos en especies no modelo (sin lugar a duda en número mucho mayor que por las técnicas tradicionales, e incluso más que con *microarrays*) para estudios filogenéticos, de diversidad, o de mapeo genético (Baird *et al.*, 2008; Davey *et al.*, 2011), la determinación de patrones de expresión y regulación génica (Wang *et al.*, 2009) o la identificación de especies de microorganismos presentes en muestras ambientales (Venter *et al.*, 2004).

Específicamente, las tecnologías de secuenciación masiva han supuesto un cambio enorme a la hora de abordar el estudio de especies silvestres, tanto de animales como de plantas, ya que éste se veía limitado con las técnicas tradicionales. En general, las especies silvestres no se contemplan como especies modelo, es decir, los recursos genómicos dedicados a su estudio son más bien escasos, de manera que la obtención y transferencia de marcadores era costosa y frecuentemente se obtenía baja resolución a nivel de cobertura genómica. Con mucha frecuencia, la caracterización de la diversidad genética en estos organismos se determinaba a partir de una representación muy reducida de sus genomas, pudiendo llevar a inferencias erróneas de parámetros poblacionales o en la determinación de fenómenos de adaptativos, evolutivos o de especiación.

El uso de las plataformas de secuenciación masiva consigue secuenciar en paralelo millones de fragmentos de ADN en múltiples individuos, lo cual redunda en un abaratamiento de costes y del tiempo de realización de los experimentos (Schloss, 2008). Además, la secuenciación masiva presenta otras ventajas (van Dijk *et al.*, 2014). En primer lugar, no es necesario clonar el ADN con bacterias, dado que las plataformas de secuenciación trabajan con bibliotecas genómicas† preparadas en sistemas libres de células. En segundo lugar, se elimina la electroforesis para detectar las bases secuenciadas, con la implementación de otras metodologías que permiten acelerar el proceso de obtención de secuencias. Además, mediante la secuenciación masiva, se puede determinar un amplio espectro de polimorfismos genómicos, desde la variación de un único par de bases o mutaciones puntuales - SNPs†- hasta inserciones y deleciones, o duplicaciones genómicas.

A pesar de sus innegables ventajas, la secuenciación masiva también presenta una serie de limitaciones que van superándose conforme se desarrollan nuevas mejoras tecnológicas. Precisamente, una de las mayores limitaciones es el manejo del elevado volumen de datos generado en cada experimento (Zhao *et al.*, 2013), que requiere un cierto manejo de técnicas bioinformáticas avanzadas y de una adecuada infraestructura computacional no siempre al alcance de los grupos de investigación (López de Heredia y Vázquez-Poletti, 2016). Además, generalmente hay una ausencia de bases de datos y recursos -ómicos† para especies no modelo, que dificultan el análisis para investigadores no especializados.

En esta revisión se describen las herramientas disponibles para la realización de estudios que utilicen técnicas de secuenciación masiva, haciendo hincapié en aquellas técnicas de mayor aplicación para estudios ecológicos y evolutivos en especies silvestres. En primer lugar, se describirán las plataformas de secuenciación masiva disponibles en la actualidad, señalando sus ventajas y limitaciones para el análisis en organismos no modelo. Seguidamente, detallaremos las bases de dos de las técnicas más populares que se benefician de la secuenciación masiva (RAD-seq y RNA-seq) con algunos ejemplos de su uso para el estudio de la diversidad biológica. Finalmente, se señalarán las principales limitaciones y perspectivas de este campo de estudio.

Plataformas de secuenciación masiva

Las técnicas que se benefician de la secuenciación masiva deben abordar necesariamente dos fases: la fase *in vitro* y la fase *in silico* (Fig. 1). La primera fase (*in vitro*) consiste en la construcción de bibliotecas genómicas de acuerdo a la técnica ensayada, seguida de la subsiguiente secuenciación de dichas bibliotecas en una plataforma de secuenciación. La metodología para construir las bibliotecas es específica de cada técnica. La segunda fase (*in silico*) comprende el análisis bioinformático de los ficheros de lecturas generados por las distintas plataformas de secuenciación masiva. Precisamente la elección de la plataforma va a condicionar las dos fases de análisis debido a que cada una de ellas presenta unos fundamentos teóricos y unas características propias que permiten al investigador seleccionar la más adecuada para el tipo de análisis deseado (Tabla 1).

Frente a la secuenciación tradicional por el método de Sanger *et al.* (1977), considerado de primera generación, se desarrollaron a finales de la década pasada las plataformas de se-

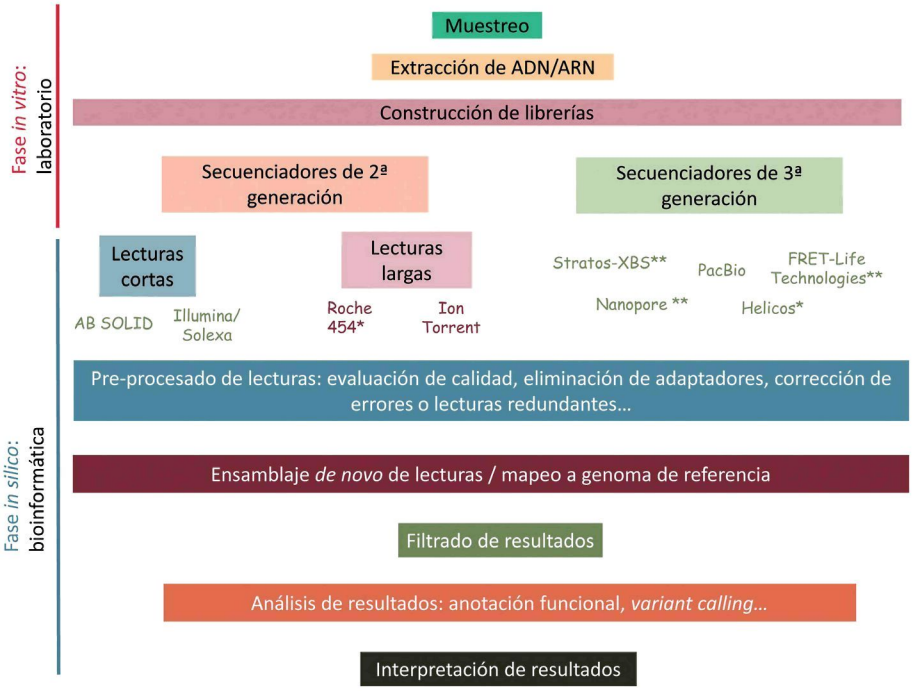

Fig. 1.- Esquema general de los procesos a llevar a cabo para realizar un experimento de secuenciación masiva.

Fig. 1.- General outline of the steps needed to conduct high-throughput experiment.

Generación	Plataforma	Tecnología	Tamaño máximo de lectura (pb)	Lecturas por corrida (ML)	GB totales	Compañía
1ª	Sanger	Longitud didesoxinucleótidos	~ 1000	-	-	Thermo Fisher Scientific, Waltham, Ma, US
2ª	Roche 454*	Pirosecuenciación	~ 700	1	14	Hoffmann-La Roche Ltd, Basilea, Suiza
	SOLiD	Ligación y codificación por dos bases	2 x 75	100	320	Thermo Fisher Scientific, Waltham, Ma, US
	Illumina/ Solexa	Secuenciación por síntesis	2 x 150	6.000	1800	Illumina Inc., San Diego, Ca, USA
	PGM/ Ion Proton	Tecnología de semiconductores	2.000	80	5	Thermo Fisher Scientific, Waltham, Ma, US
3ª	Helicos*	Secuenciación individual con moléculas fluorescentes	70	20†	20†	Helicos Biosciences, Cambridge, Ma, USA
	PacBio	Secuenciación de una única molécula l en tiempo real	30.000	0,55†	1†	Pacific Biosciencies, Menlo Park, Ca, USA
	Nanopore	Bioporos	200.000	1.250	12.000††	Oxford Nanopore Technologies, Oxford, Reino Unido
	FRET**	Transferencia energética de resonancia de la fluorescencia	-	-	-	Thermo Fisher Scientific, Waltham, Ma, US
	Stratos**	Secuenciación por expansión	-	-	-	Hoffmann-La Roche Ltd, Basilea, Suiza

* No se comercializan en la actualidad
** Actualmente en desarrollo
† Por celda o canal
†† Máximo teórico

* Not currently in the market
** Currently under development
† By cell or channel
†† Theoretical maximum

Tabla 1.- Plataformas de secuenciación de segunda y tercera generación. Se especifican los tamaños máximos de lectura, el número de lecturas máximo por corrida y el total de Gb alcanzado en los modelos más modernos de cada plataforma.

Table 1.- Second and third generation sequencing platforms. The maximum read sizes, maximum read number per run, and total Gb reached by the most up-to-date models for each platform.

cuenciación de segunda y tercera generación, también conocidas como plataformas de secuenciación masiva. Aparte de las diferencias en la metodología para diseñar las bibliotecas genómicas que se van a secuenciar y en los fundamentos de cada técnica para obtener las secuencias de nucleótidos, existe una división entre plataformas de secuenciación

masiva en función de la longitud de las lecturas (es decir, el número de pares de bases de cada lectura) que generan, de manera que se habla de plataformas de lecturas largas (> 300 pb) y de lecturas cortas (< 300 pb) (van Dijk *et al.*, 2014). La longitud de lectura es una de las principales variables y limitaciones que se deben tener en cuenta a la hora de diseñar los experimentos. Una longitud de lecturas corta implica necesariamente el mapeo a un genoma de referencia o el ensamblaje *de novo*† de dichas lecturas, mientras que las lecturas largas implican una mayor tasa de error en la identificación de los nucleótidos (Liu *et al.*, 2012). No obstante, a lo largo de los últimos años, las plataformas de lecturas cortas han ha ido consiguiendo un incremento progresivo en la longitud de lectura (> 250 pb en lecturas cortas) manteniendo la baja tasa de error (*c.* 0.01%), mientras que las plataformas de lecturas largas han conseguido reducir el error (c. 10-40%), llegando a conseguir la secuenciación de moléculas completas, por ejemplo de ARN mensajero. Estos avances metodológicos abren la puerta a la consecución de resultados más robustos, incluso con el uso combinado de distintas plataformas.

La primera plataforma de secuenciación masiva que se comercializó fue el 454 Genome Sequencer de Life Sciences (actualmente Hoffmann-La Roche Ltd, Basilea, Suiza). Esta plataforma se basaba en la técnica de pirosecuenciación† (Margulies *et al.*, 2005), y conseguía inicialmente lecturas simples† (*single-end*), de 100-150 pares de bases (pb), hasta completar 200 Mb† de secuencia. Tras sucesivas evoluciones, esta plataforma consiguió alcanzar longitudes de lectura de ~ 700 pb y un volumen total de secuencia de 14 Gb†. No obstante, aunque en sus inicios la plataforma 454 constituía una alternativa rápida y eficiente para conseguir lecturas largas con una precisión media (Pushkarev *et al.*, 2009), Roche anunció en octubre de 2013, que no seguiría dando soporte a esta plataforma, por su incapacidad de competir con las nuevas plataformas que se habían desarrollado.

Al poco tiempo de la aparición de la plataforma 454, surgieron dos plataformas caracterizadas por la identificación de las bases por fluorescencia y por la generación de un número muy superior de lecturas más cortas pero más precisas. Por un lado, Applied Biosystems (actualmente Life Technologies, una marca de Thermo Fisher Scientific, Waltham, Ma, USA) desarrolló en 2006 el sistema SOLiD (*Sequencing by Oligo Ligation Detection*) basado en la tecnología de secuenciación por ligación y codificación por dos bases† (Mardis, 2008). Está tecnología producía inicialmente 3Gb de lecturas de 35 pb, pero los modelos actuales consiguen generar outputs de hasta 300 Gb y lecturas pareadas† (*paired-end*) de 2 x 75 pb con una precisión del 99,9 %. Por otro lado, la compañía Illumina Inc. (San Diego, Ca, USA) ha desarrollado distintos modelos de secuenciadores tras adquirir el primer Genome Analyzer comercializado por Solexa en 2006. Todos los secuenciadores de Illumina (HiSeq, MiSeq, NextSeq, etc.) utilizan la tecnología de secuenciación por síntesis† (Ju *et al.*, 2006). En la actualidad, los secuenciadores de Illumina pueden conseguir lecturas de 2 x 300 pb hasta 15 Gb y 25 millones de lecturas (ML) (modelo MiSeq) o bien lecturas de 2 x 150 pb hasta 1800 Gb y 6.000 ML (modelo HiSeqX).

El incremento en el tamaño de lecturas es esencial para aumentar la profundidad de secuenciación†, particularmente a la hora de realizar ensamblajes *de novo* en especies de las

que no se dispone de un genoma de referencia. El uso de lecturas pareadas pertenecientes a la misma región genómica permite utilizar la posición de las mismas y su distancia entre ellas para mejorar la precisión del ensamblaje. Tanto SOLiD como Illumina son capaces de generar tanto archivos de lecturas simples como de lecturas pareadas, e incluso de lecturas con direccionalidad conocida (*stranded-specific*).

En 2010 Ion Torrent (actualmente Thermo Fisher Scientific, Waltham, Ma, USA) comercializó su sistema PGM (*Personal Genome Machine*), que incorporaba la tecnología de semiconductores† y no dependía del uso de fluorescencia, sino de los cambios en pH que se producen cuando se libera un protón al incorporarse a una molécula de ADN. El proceso se lleva a cabo en un chip con micro-pocillos cuya base actúa como un PH-metro microscópico para detectar dichos cambios. El ADN se fragmenta y se anilla a unas pequeñas bolas que se sumergen en una solución de cada uno de los nucleótidos en los micro-pocillos del chip. Dado que cada lámina del chip actúa como un semiconductor, se pueden detectar los cambios iónicos producidos por la incorporación de un nucleótido dado a la secuencia de ADN, y transferir la secuencia de nucleótidos a un ordenador. El sistema PGM es capaz de generar lecturas largas desde 400 pb hasta 2 Gb y 5 ML. La evolución de la tecnología de semiconductores ha desembocado en los secuenciadores Ion Proton (Thermo Fisher Scientific, Waltham, Ma, US), que son capaces de generar lecturas de 200 pb hasta10 Gb y 80 ML.

Más recientemente, se ha desarrollado la denominada secuenciación de la tercera generación. Las principales diferencias de las plataformas de tercera generación con las de segunda generación radican en la no necesidad de realizar una PCR† (reacción en cadena de la polimerasa) antes de la secuenciación (aunque existen protocolos para la construcción de bibliotecas genómicas sin PCR para su secuenciación en Illumina), reduciendo el tiempo de análisis y los errores derivados de ésta, y en que la detección de nucleótidos se realiza en tiempo real. La plataforma PacBio (Pacific Biosciencies, Menlo Park, Ca, USA) se basa en fluorescencia para identificar los nucleótidos mediante un procedimiento de secuenciación de una única molécula en tiempo real† (*single-molecule real-time* -SMRT-). Los secuenciadores PacBio RS operan acortando mucho los tiempos y consiguiendo una longitud media de secuencia de 4.200 a 8.500 pb, llegando hasta los 30.000 pb. Estas características hacen a esta plataforma muy útil para estudios de microbiología que analizan genomas muy pequeños porque permiten una profundidad de secuenciación suficiente para corregir la tasa de error, o para la obtención de transcritos de longitud completa en estudios de expresión génica, corrigiendo los errores cometidos en combinación con lecturas cortas más fiables, por ejemplo obtenidas con secuenciadores de Illumina (González-Ibeas *et al.*, 2015).

La secuenciación Nanopore, que se encuentra en un estado avanzado de desarrollo, utiliza un bioporo macroscópico de muy pequeño diámetro similar a una canal iónico de una membrana plasmática, por el que transitan las moléculas de ADN o proteínas, produciendo una interrupción del voltaje a través del canal (Clarke et al., 2009) y cuyo registro permite identificar la secuencia de nucleótidos. Oxford Nanopore Technologies (Oxford, Reino Unido) ha comercializado una serie de secuenciadores que utilizan esta tecnología (MinIon,

PromethIon y GridIon). Por ejemplo, utilizando el secuenciador portable MinIon, se ha sido capaz de secuenciar un cromosoma completo de la bacteria *Bacterioides fragilis* de 5.18 Mb (Risse *et al.*, 2015).

Existen sistemas de secuenciación de tercera generación, como el de la plataforma Helicos (Helicos Biosciences, Cambridge, Ma, USA), que no se comercializan actualmente, mientras que otros sistemas se encuentran en fase incipiente de desarrollo. Entre estos últimos, se encuentran el sistema Stratos, apoyado por Roche, y que utiliza la tecnología de secuenciación por expansión† -XBS- basado en nanoporos, o el sistema de transferencia energética de resonancia de la fluorescencia† -FRET-, que se basa en microscopía electrónica y que es la apuesta más firme de Thermo Fisher Scientific en las tecnologías de secuenciación de tercera generación. Sólo el tiempo dirá cuál de las plataformas mencionadas resultará más exitosa en el futuro de la secuenciación masiva.

Las técnicas de genotipado por secuenciación: RAD-seq

La secuenciación de genomas completos sólo se ha aplicado de manera marginal en estudios filogenéticos, filogeográficos o de genética de poblaciones de plantas y animales por el elevado coste de trabajar con un tamaño muestral alto. No obstante, la secuenciación masiva puede ser aplicada a bibliotecas genómicas que consigan una reducción del genoma de los organismos y a partir de la cual se obtenga un número elevado de variantes genómicas -*loci*† polimórficos- en múltiples individuos. La reducción del genoma se consigue en estas metodologías mediante el empleo de endonucleasas de restricción, que son enzimas que reconocen un patrón específico de bases en las secuencias de ADN y fragmentan la doble hélice en sitios específicos de cuatro a ocho nucleótidos: las dianas de restricción.

Esta capacidad de las endonucleasas de restricción para reconocer secuencias específicas de ADN es la base de la técnica RAD-seq (*restriction site associated DNA markers sequencing*) (Fig. 2a). Aunque existen multitud de modificaciones de esta técnica en la manera de construir las bibliotecas genómicas, todas tienen en común la digestión del ADN genómico por una o dos endonucleasas de restricción. En el caso de la técnica original, la digestión va seguida de una fragmentación aleatoria de los fragmentos resultantes, que usualmente se lleva a cabo mediante métodos mecánicos como la sonicación (fragmentación mediante ultrasonidos) o la nebulización (fragmentación a alta presión) (Baird *et al.*, 2008; Davey y Baxter, 2010). El siguiente paso consiste en ligar a los fragmentos unos adaptadores que van a llevar incluida una secuencia corta de 4-8 nucleótidos (*barcode*†) que va a ser específico para cada uno de los individuos del experimento, ya que es la manera de poder separar *a posteriori* las lecturas que corresponden a cada muestra. Los RADs, fueron empleados inicialmente en *microarrays* (Miller *et al.*, 2007), para ser luego secuenciados con plataformas NGS, principalmente de lecturas cortas (Baird *et al.*, 2008). Las lecturas van a cubrir los dos lados de la diana de restricción, y pueden ser simples o pareadas. Las dianas de restricción de la secuencia de ADN que se buscan con RAD-seq son poco frecuentes, de manera que el número de fragmentos que se producen sea tratable para una plataforma

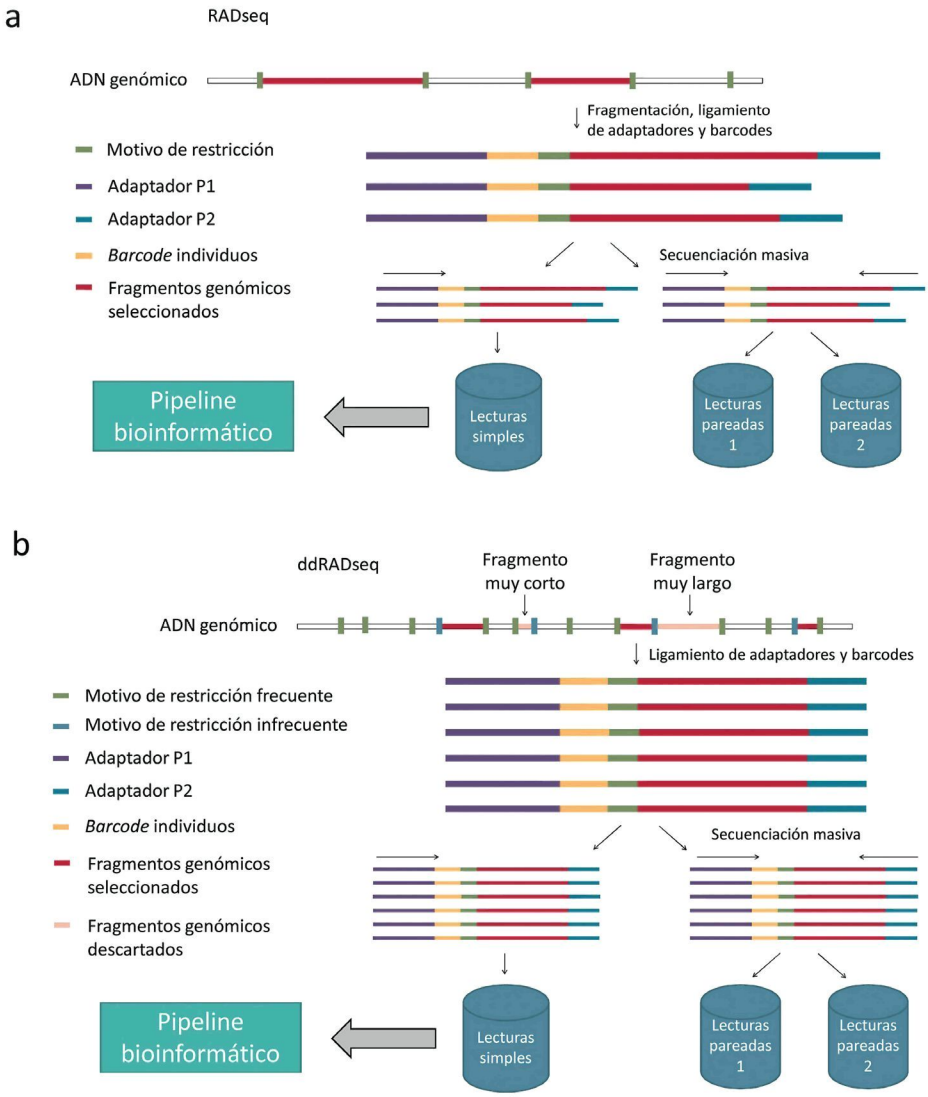

Fig. 2.- Esquema de las metodologías de genotipado RAD-seq (a) y ddRAD-seq (b).

Fig. 2.- Outline of RAD-seq (a) and ddRAD-seq (b) genotyping methodologies.

de secuenciación y no se genere un número elevado de datos perdidos por *locus* e individuo.

A día de hoy existen numerosas modificaciones de la técnica RAD-seq original, que permiten optimizar la obtención de marcadores genómicos a costes razonables, aumentar el número de individuos que se pueden analizar, reducir el volumen de datos perdidos o minimizar las fuentes de error. A modo de ejemplo, vamos a detallar dos de ellas, por su popularidad en el estudio de especies de plantas y animales: (1) *genotipado por secuenciación* -GBS- (Elshire *et al.*, 2012); y (2) *double-digested RAD-seq* -ddRAD-seq- (Peterson *et al.*, 2012). Otras modificaciones como *ezRAD-seq* (Toonen *et al.*, 2013), testada en especies de peces, corales y moluscos, o el genotipado por secuenciación con dos enzimas, como *2-enzyme GBS* (Poland *et al.*, 2012) o *2b-RAD* (Wang *et al.*, 2012), han tenido hasta ahora un menor uso.

La técnica de GBS se desarrolló para estudios de mapeo genético en plantas de maíz (Elshire *et al.*, 2012). Es muy similar a la técnica original, pero modula la composición del *barcode* utilizando nucleótidos degenerados† para minimizar el error en la identificación de los *loci* secuenciados. La principal modificación consiste en la utilización de endonucleasas de restricción resistentes a metilación† (originalmente *ApeK*I), de manera que se evitan regiones repetitivas del genoma, aumentando la profundidad de secuenciación de regiones con un bajo número de copias. Además, utiliza las propias lecturas obtenidas para realizar la fase de mapeo, por lo que no es necesario un genoma de referencia. Aunque esta técnica ha sido utilizada principalmente en especies con interés agronómico, hay algunos ejemplos de su uso en genética poblacional de especies silvestres de plantas (Chen *et al.*, 2013; Schilling *et al.*, 2014; Ilut *et al.*, 2015; Ratcliffe *et al.*, 2015) o animales (Swaegers *et al.*, 2015; Winger *et al.*, 2015; Cahill y Levinton, 2016; Dussex *et al.*, 2016; Underwood *et al.*, 2016).

Probablemente la metodología con más potencial para el estudio de la diversidad biológica es la técnica ddRAD-seq (Peterson *et al.*, 2012). La principal modificación que presenta respecto de la técnica original es el uso conjugado de dos endonucleasas de restricción, seleccionando únicamente los fragmentos que están flanqueados por cada una de las enzimas de restricción y que muestran un tamaño de inserto definido por el investigador. Esta modificación implica que no es necesario realizar una fase de fragmentación del genoma. La identificación de los individuos se realiza mediante la adición de uno o dos *barcodes* a los adaptadores que flanquean los sitios de restricción, cuya longitud va a determinar el número de individuos que se podrán analizar en cada biblioteca genómica (Figura 2b). Además, opcionalmente se puede añadir otra secuencia corta de nucleótidos degenerados (*degenerate base region*, DBR), que se puede utilizar como contador para detectar y eliminar duplicados de PCR (Schweyen *et al.*, 2014; Tin *et al.*, 2015). La presencia de duplicados de PCR es un artefacto que se produce en el paso previo a la secuenciación y cuyo efecto principal es condicionar la profundidad de secuenciación (DaCosta y Sorensen, 2014). Otra estrategia para cuantificar el impacto de potenciales errores de genotipado en experimentos de ddRAD-seq (y de RAD-seq en general) es la utilización de réplicas técnicas (aquellas en las que el material biológico es el mismo en cada muestra) para optimizar los

parámetros utilizados en pipelines† de análisis bioinformático y acotar dicho error (Mastretta-Yanes *et al.*, 2015).

En el ámbito del estudio de la diversidad biológica, la técnica ddRAD-seq se ha utilizado para estudios de hibridación y delimitación de especies, establecimiento de estructura poblacional o detección de regiones genómicas sometidas a selección en organismos tan diversos como invertebrados marinos (Saarman *et al.*, 2015; Lal *et al.*, 2016), aves (DaCosta *et al.*, 2016; Lavretsky *et al.*, 2016), reptiles (Leaché *et al.*, 2015), lepidópteros (Capblancq *et al.*, 2015; Lee y Mutanen, 2015), peces (Kai *et al.*, 2014; Saenz-Agudelo *et al.*, 2015); o coníferas (Friedline *et al.*, 2015) y frondosas (Mastretta-Yanes *et al.*, 2014). Usualmente, esta técnica ha ido asociada a la secuenciación con la plataforma Illumina, de lecturas cortas, ya que frecuentemente se desea contar con *loci* no ligados, y éstos aumentan con la proximidad en el genoma. Es decir, se prefiere obtener muchos *loci* con pocos polimorfismos, que pocos *loci* con muchos polimorfismos. No obstante, el diseño del experimento dependerá de los objetivos del mismo, y recientemente se ha desarrollado una adaptación del protocolo ddRAD-seq para su utilización con la plataforma Ion Torrent, de lecturas largas (Pukk *et al.*, 2015).

En los últimos años se han desarrollado diversos pipelines de análisis bioinformático, tanto para el diseño de experimentos de RAD-seq (*sensu lato*), como para el procesado de las lecturas obtenidas. Ejemplos de estos pipelines son Stacks (Catchen *et al.*, 2103), Pyrad (Eaton, 2014), AftrRAD (Sovic *et al.*, 2015), PredRAD (Herrera *et al.*, 2015) o ddRADseqTools (Mora-Márquez *et al.*, 2016). En general, todos los pipelines incluyen los siguientes pasos: (1) pre-procesado de las lecturas, incluyendo la evaluación de su calidad, el demultiplexado de individuos, es decir, la generación de ficheros de lecturas para cada individuo, y la eliminación de los adaptadores; (2) ensamblaje de los fragmentos secuenciados en grupos de lecturas correspondiendo cada grupo a un *locus*. (3) determinación de polimorfismos o *variant calling*; (4) anotación funcional de *loci* presentes en las regiones codificantes del genoma; y (5) generación de ficheros de salida en formatos utilizables por software de análisis de genética poblacional, filogenética o mapeo. Como ya hemos mencionado, el uso de estas herramientas requiere de unos mínimos conocimientos bioinformáticos por parte del investigador.

El estudio de transcriptomas: RNA-seq

Si bien la genética de poblaciones se centra fundamentalmente en el estudio de regiones neutrales del genoma, no es menos importante el modo en que se expresan los genes de los distintos organismos. La secuenciación masiva también ha supuesto una revolución en los estudios de transcriptómica a través de la implementación de las técnicas de secuenciación directa de los ARN transcritos o RNA-seq (Wang *et al.*, 2009). El RNA-seq permite determinar y cuantificar los niveles de expresión de genes que se expresan bajo diferentes condiciones ambientales o en respuesta a diferentes estímulos (Ekblom y Galindo, 2011) y/o detectar genes previamente no identificados y sus variantes (Wang *et al.*, 2009), por lo que está pasando a ser la metodología más utilizada frente a los *microarrays* de expresión.

Hay varias cuestiones que deben considerarse a la hora de afrontar un experimento de RNA-seq. En primer lugar, es importante tener en cuenta el material de partida que se va a utilizar en función de la realidad biológica que se pretende estudiar. La expresión génica presenta una alta especificidad en función del tejido examinado (Brawand *et al.*, 2011). Así, hay que ser cauto en la interpretación de los resultados, ya que los transcriptomas generados a partir de tejidos diversos representan una mezcla heterogénea de los perfiles de expresión obtenidos de las células individuales. En este sentido, si bien la mayoría de estudios operan a nivel de individuo completo o de tejidos/órganos específicos, cabe destacar la posibilidad de obtener transcriptomas a nivel de célula (Hebenstreit, 2012). Estas técnicas, que utilizan la microdisección de células con láser para extraer su ARN y secuenciarlo, tienen su principal aplicación en el estudio del cáncer (Saliba *et al.*, 2014), pero también hay ejemplos de su uso en especies de plantas (Abbott *et al.*, 2010).

Otra decisión importante es la elección de la plataforma de secuenciación, ya que va a condicionar la construcción de bibliotecas genómicas, así como el posterior procesado bioinformático de las mismas, en función de tratarse de plataformas de lecturas cortas (Illumina) o de lecturas largas (454 Roche o PacBio). En un reciente estudio se ha cuantificado la frecuencia de uso de distintos tipos de plataforma para el análisis de RNA-seq de las principales especies de árboles forestales (López de Heredia y Vázquez-Poletti, 2016). Los resultados muestran un temprano uso de la plataforma 454 de Roche, que paulatinamente se ha visto sustituido por el empleo de la plataforma Illumina. El uso de otras plataformas como AB SOLiD o Ion Torrent ha sido marginal, aunque se prevé que el uso de los secuenciadores de tercera generación tendrán un papel preponderante en el análisis de RNA-seq en los próximos años, debido a que son capaces de secuenciar transcritos completos sin precisar del ensamblaje de las lecturas.

La longitud y el número de lecturas generadas va a tener una incidencia fundamental en la profundidad de secuenciación, que, de acuerdo a la ecuación de Lander y Waterman (1988), es directamente proporcional a la longitud y número de lecturas e inversamente proporcional a la longitud del genoma. En estudios de RNA-seq, la profundidad de secuenciación es difícil de determinar, y se va a producir una heterogeneidad en los niveles de expresión para todos los *loci*; es decir, los genes que se expresan en mayor medida presentarán un mayor número de lecturas. Además, para asegurar la reproducibilidad de los experimentos de RNA-seq no es usual incluir réplicas técnicas, pero sí se deben incluir al menos tres réplicas biológicas (Marioni *et al.*, 2008). Las réplicas técnicas evalúan la variabilidad técnica que se produce en el proceso de construcción de bibliotecas genómicas y secuenciación. Las réplicas biológicas consisten en muestras de diferentes individuos que se procesan de manera separada, en aras de una estimación de la varianza en los niveles de expresión y de una reducción del error experimental producido por la intrínseca variación entre los mismos (Robasky *et al.*, 2014). En este sentido, se prefiere incrementar el número de réplicas biológicas frente al número de lecturas por réplica (Liu *et al.*, 2014).

Un experimento genérico de RNA-seq (Figura 3) debe adaptar la construcción de bibliotecas genómicas a la plataforma utilizada y a la realidad biológica que se pretende capturar.

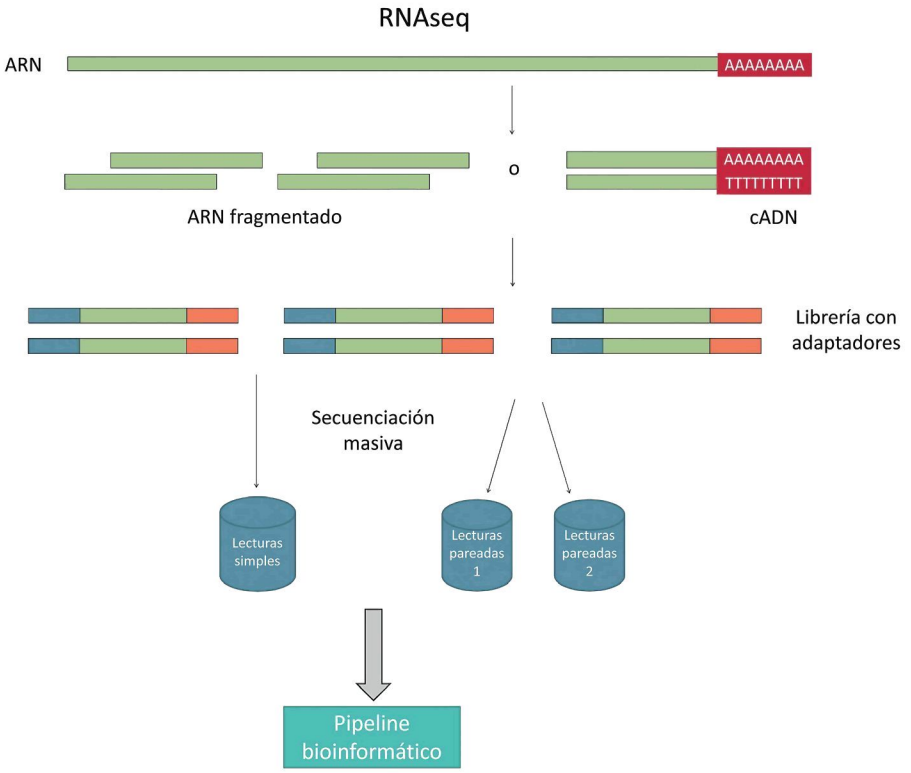

Fig. 3.- Esquema de la metodología para la análisis de RNA-seq (modificado de Mortazavi *et al.*, 2008).

Fig. 3.- Outline of the methodology for the analysis of RNA-seq (modified from Mortazavi *et al.*, 2008).

Así, existen multitud de modificaciones a las bibliotecas estándar de ARN total, como por ejemplo adaptaciones para analizar únicamente ARN mensajero (mRNA), microARNs (miRNA) o perfiles ribosómicos, o bien la secuenciación dirigida a ARN de rutas metabólicas conocidas (Vikman *et al.*, 2014).

Si bien normalmente los análisis de RNA-seq se realizan sobre muestras localizadas en ambientes controlados y sometidas a distintos estímulos (por ej. estrés abiótico por sequía, infección con algún patógeno, respuesta a traumatismos, etc.), también se han realizado algunos experimentos en los que se analizan poblaciones naturales para ver cómo varían los niveles de expresión a través de gradientes ambientales (Sork *et al.*, 2016), o para detectar marcadores diagnóstico con importancia adaptativa (fundamentalmente SNPs) entre diferentes poblaciones, con una estrategia muy similar a la utilizada en RAD-seq, pero para las regiones codificantes del genoma. En este sentido, es llamativo el estudio de Suárez-

González *et al.* (2016), en el que se muestran evidencias de regiones genómicas sometidas a introgresión adaptativa en especies de chopos.

Una aplicación de RNA-seq que es particularmente eficaz, tanto en el ámbito del estudio de la diversidad biológica como en la determinación de tratamientos frente a organismos patógenos, es el RNA-seq dual, mediante el cual se analizan en una sola biblioteca genómica muestras de hongos o bacterias y los organismos superiores infectados por ellas. Esta aproximación se ha llevado a cabo fundamentalmente para la detección de hongos patógenos en plantas (Hayden *et al.*, 2013; Meyer *et al.*, 2016). Usualmente, el RNA-seq dual se beneficia de la mayor facilidad de contar con un genoma de referencia para el patógeno en cuestión, de modo que se pueden separar los genes del huésped y el hospedante para así poder investigar los cambios en los patrones de expresión de diversos genes y relacionarlos con la virulencia del ataque.

Probablemente la mayor limitación de los experimentos de RNA-seq se encuentre en el análisis bioinformático de los ficheros de lecturas generados por las diferentes plataformas de secuenciación. El primer problema es cómo tratar el elevado volumen de datos que se generan en un proyecto de RNA-seq. Los algoritmos para analizar los ficheros de lecturas tienen un elevado consumo de memoria RAM y CPU, y además van a ir generando un ingente volumen de datos a lo largo del proceso de análisis. En el caso de experimentos de expresión diferencial que analicen varias réplicas biológicas con diversos tratamientos, el volumen de datos puede volverse inmanejable, a excepción de que se cuente con una infraestructura computacional adecuada (Zhao *et al.*, 2013).

Un flujo básico de análisis bioinformático de RNA-seq incluye los siguientes pasos: (1) preprocesado de las lecturas de secuencias: estimación de su calidad, eliminación de adaptadores y de datos redundantes; (2) mapeo de las lecturas a un genoma/transcriptoma de referencia o ensamblaje *de novo* de los transcritos; (3) anotación funcional de los transcritos; (4) cuantificación de los niveles de expresión de los distintos transcritos generados, incluyendo la determinación de *splicing* alternativo†; (5) análisis de expresión diferencial. En la actualidad existen multitud de aplicaciones para la realización de todas las fases de análisis bioinformático en experimentos de RNA-seq, que pueden consultarse en revisiones recientes (López de Heredia y Vázquez-Poletti, 2016).

Conclusiones

Como hemos visto, la secuenciación masiva se presenta como una herramienta fundamental para el estudio de la diversidad biológica, porque permite un análisis eficiente y profuso de genomas y transcriptomas en un gran número de individuos de especies no modelo obteniendo una profundidad de secuenciación significativa. Entre otras aplicaciones, la secuenciación masiva tiene aplicaciones en estudios poblacionales o filogenéticos mediante el genotipado de individuos o la obtención de genomas completo, en estudios de expresión diferencial y en la identificación de los patrones de expresión de genes concretos, o en la determinación de la composición genética de muestras ambientales. A día

de hoy, estas técnicas aún presentan un coste relativamente elevado (aunque menor por par de base al de la secuenciación de Sanger). No obstante, se prevé que su coste se reduzca en el futuro cercano. Previsiblemente, la secuenciación de tercera generación aumentará la fiabilidad de los genomas y transcriptomas que se generen en el futuro, entre otros, porque facilitarán el ensamblaje de lecturas. Paulatinamente, se está consiguiendo una reducción del error de lectura en las nuevas plataformas de secuenciación. Además, se están produciendo constantes avances para paliar las limitaciones de infraestructura computacional de los grupos de investigación que no tienen acceso a supercomputadores capaces de manejar los datos generados por las plataformas de secuenciación masiva, como por ejemplo los servicios de computación bajo demanda en la nube (*cloud computing*). En este sentido, es seguro que se desarrollarán metodologías para dar solución a algunos de los problemas existentes en la actualidad, pero surgirán nuevas cuestiones que merecerán la atención de los investigadores de lo natural. En la era del Big Data, la diversidad biológica no ha quedado exenta de integrarse y beneficiarse del inmenso conocimiento del que se dispone, y es labor de los investigadores filtrar esa ingente información para profundizar en el conocimiento, tanto básico como aplicado de la naturaleza.

Bibliografía

- **Aebischer, A.** 2009. *Der Rotmilan - Ein faszinierender Greifvogel.* Haupt Verlag. Bern.
- **Abbott, E., Hall, D., Hamberger, B., Bohlmann, J.** 2010. Laser microdissection of conifer stem tissues: isolation and analysis of high quality RNA, terpene synthase enzyme activity and terpenoid metabolites from resin ducts and cambial zone tissue of white spruce (*Picea glauca*). *BMC Plant Biol* 10: 106.
- **Avise, J.C.** 1994. *Molecular Markers, Natural History, and Evolution*. Chapman & Hall. New York.
- **Avise, J.C.** 2000. *Phylogeography: The History and Formation of Species.* Harvard University Press. Cambridge, MA.
- **Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A.** 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLos One* 3: e3376.
- **Bossart, J.L., Prowell, D.P.** 1998. Genetic estimates of population structure and gene flow: Limitations, lessons and new directions. *Trends Ecol. Evol.* 13: 202-206.
- **Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F.W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., Kaessmann, H.** 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478: 343-8.
- **Cahill, A.E., Levinton, J.S.** 2016. Genetic differentiation and reduced genetic diversity at the northern range edge of two species with different dispersal modes. *Mol. Ecol.* .25: 515-26.

- **Capblancq, T., Després, L., Rioux, D., Mavarez, J.** 2015. Hybridization promotes speciation in *Coenonympha* butterflies. *Mol. Ecol.* 24: 6209-6222.
- **Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., Cresko, W.A.** 2013. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22: 3124-40.
- **Chen, C., Mitchell, S.E., Elshire, R.J., Buckler, E.S., El-Kassaby, Y.A.** 2013. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genet. Genomes* 9: 1537-1544.
- **Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A, Reid, S., Bayley, H.** 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4: 265-270.
- **DaCosta, J.M., Sorenson, M.D.** 2014. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS One* 9: e106713.
- **Davey, J.W., Blaxter, M.L.** 2010. RADSeq: next-generation population genetics. *Brief. Funct. Genomics* 9: 416-423.
- **Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter. M.L.** 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12: 499-510.
- **Dussex, N., Chuah, A., Waters, J.M.** 2016. Genome-wide SNPs reveal fine-scale differentiation among wingless alpine stonefly populations and introgression between winged and wingless forms. *Evolution* 70: 38-47.
- **Eaton, D.A.R.** 2014. PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics* 30: 1844-1849.
- **Ekblom, R., Galindo, J.** 2011 Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107: 1-15.
- **Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E.** 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379.
- **Friedline, C.J., Lind, B.M., Hobson, E.M., Harwood, D.E., Mix, A.D., Maloney, P.E., Eckert, A.J.** 2015. The genetic architecture of local adaptation I: the genomic landscape of foxtail pine (*Pinus balfouriana* Grev. & Balf.) as revealed from a high-density linkage map. *Tree Genet. Genomes* 11: 49.
- **González-Ibeas, D., Martinez-Garcia, P.J., Famula, L., Loopstra, C.A., Puryear, J., Neale, D.B., Wegrzyn, J.L.** 2015. Survey of the Sugar Pine (*Pinus lambertiana*) Transcriptome By Deep Sequencing. *Plant and Animal Genome XXIII*. San Diego, CA. 10-14 ENE 2015.
- **Hayden, K.J., Garbelotto, M., Knaus, B.J., Cronn, R.C., Rai, H., Wright, J.W.** 2013. Dual RNA-seq of the plant pathogen *Phytophthora ramorum* and its tanoak host. *Tree Genet. Genomes* 10: 489-502.
- **Hebenstreit, D.** 2012. Methods, Challenges and Potentials of Single Cell RNA-seq. *Biology* 1: 658-667.
- **Herrera, S., Reyes-Herrera, P.H., Shank, T.M.** 2015. Predicting RAD-seq Marker Numbers across the Eukaryotic Tree of Life. *Genome Biol. Evol.*7: 3207–3225.

- **Hewitt, G.M.** 1999. Post-glacial re-colonization of European biota. *Biol. J. Linn. Soc.* 68: 87-112.
- **Hoheisel, J.D.** 2006. Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.* 7: 200-10.
- **Ilut, D.C., Sanchez, P.L., Costichc, D.E., Friebed, B., Coffeltb, T.A., Dyerb J.M., Jenkse, M.A., Gorea, M.A.** 2015. Genomic diversity and phylogenetic relationships in the genus Parthenium (Asteraceae). Ind. Crop. Prod. 76: 920-929.
- **Ju, J., Kim D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J., Edwards, J.R., Romu, A., Turro, N.J.** 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U.S.A.* 103: 19635-19640.
- **Kai, W., Nomura, K., Fujiwara, A., Nakamura, Y., Yasuike, M., Ojima, N., Masaoka, T., Ozaki, A., Kazeto, Y., Gen, K., Nagao, J., Tanaka, H., Kobayashi, T., Ototake, M. A.** ddRAD-based genetic map and its integration with the genome assembly of Japanese eel (*Anguilla japonica*) provides insights into genome evolution after the teleost-specific genome duplication. *BMC Genomics* 201415: 233.
- **Lal, M.M., Southgate, P.C., Jerry, D.R., Zenger, K.R.** 2016. Fishing for divergence in a sea of connectivity: The utility of ddRADseq genotyping in a marine invertebrate, the black-lip pearl oyster *Pinctada margaritifera*. *Mar Genomics* 25: 57-68.
- **Lander, E.S., Waterman, M.S.** 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2: 231-239.
- **Lavretsky, P., Peters, J.L., Winker, K., Bahn, V., Kulikova, I., Zhuravlev, Y.N., Wilson, R.E., Barger, C., Gurney, K., McCracken, K.G.** 2016. Becoming pure: identifying generational classes of admixed individuals within lesser and greater scaup populations. *Mol. Ecol.* 25: 661-74.
- **Leaché, A.D., Chavez, A.S., Jones, L.N., Grummer, J.A., Gottscho, A.D., Linkem, C.W.** 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biol. Evol.* 7: 706-19.
- **Lee, K.M., Mutanen, M.** 2015. Species delimitation of *Eupithecia* (Lepidoptera: Geometridae) using a ddRAD-Seq approach. *Genome* 58: 244-244.
- **López de Heredia, U., Vázquez-Poletti, J.L.** 2016. RNA-seq analysis in forest tree species: bioinformatics problems and solutions. *Tree Genet. Genomes* 12: 30. doi:10.1007/s11295-016-0995-x.
- **Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M.** 2012. Comparison of Next-Generation Sequencing Systems. *J. Biomed. Biotechnol.* 2012: 251364.
- **Liu, Y., Zhou, J., White, K.P.** 2014. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* 30: 301-4.
- **Mardis, E.R.** 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24: 133-141.
- **Margulies, M., Michael, E., Altman, W.E., Said, A., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L.,**

Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkes, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.

- **Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y.** 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18: 1509-1517.
- **Mastretta-Yanes, A., Zamudio, S., Jorgensen, T.H., Arrigo, N., Alvarez, N., Piñero, D., Emerson, B.C.** 2014. Gene duplication, population genomics, and species-level differentiation within a tropical mountain shrub. *Genome Biol. Evol.* 6: 2611-2624.
- **Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T.H., Piñero, D., Emerson, B.C.** 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* 15: 28-41.
- **Meyer, F.E., Shuey, L.S., Naidoo, S., Mamni, T., Berger, D.K., Myburg, A.A., van den Berg, N., Naidoo, S.** 2016. Dual RNA-Sequencing of *Eucalyptus nitens* during *Phytophthora cinnamomi* challenge reveals pathogen and host factors influencing compatibility. *Front Plant Sci.* 7: 191.
- **Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., Johnson, E.A.** 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* 17: 240-248.
- **Mora-Márquez, F., García-Olivares, V., Emerson, B.C., López de Heredia, U.** 2016. ddRADseqTools: a software package for *in silico* simulation and testing of double digest RADseq experiments. *Mol. Ecol. Res. Online version:* doi:10.1111/1755-0998.12550.
- **Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 5: 621-628.
- **Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E.** 2012. Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS One* 7: e37135.
- **Poland, J.A., Brown, P.J., Sorrells, M.E., Jannink, J.L.** 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7: e32253.
- **Pukk, L., Ahmad, F., Hasan, S., Kisand, V., Gross, R., Vasemägi, A.** 2015. Less is more: extreme genome complexity reduction with ddRAD using Ion Torrent semiconductor technology. *Mol. Ecol. Res.* 15: 1145-1152.
- **Pushkarev, D., Neff, N.F., Quake, S.R.** 2009. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* 27: 847-850.
- **Ratcliffe, B., El-Dien, O.G., Klápště, J., Porth, I., Chen, C., Jaquish, B., El-Kassaby, Y.A.** 2015. A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* × *glauca*) using unordered SNP imputation methods. *Heredity* 115: 547-555.
- **Rentaría Alcántara, M.** 2007. Breve revisión de los marcadores moleculares. En: *Ecología Molecular.* L. Eguiarte, V. Souza, X. Aguirre (Ed.): 541-566. Instituto Nacional de Ecología. México DF.

- **Risse, J., Thomson, M., Patrick, S., Blakely, G., Koutsovoulos, G., Blaxter, M., Watson, M.** 2015. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience* 4: 60.
- **Robasky, K., Lewis, N.E., Church, G.M.** 2014 The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15:56-62.
- **Saarman, N.P., Pogson, G.H.** 2015. Introgression between invasive and native blue mussels (genus *Mytilus*) in the central California hybrid zone. *Mol. Ecol.* 24: 4723-4738.
- **Saenz-Agudelo, P., Dibattista, J.D., Piatek, M.J., Gaither, M.R., Harrison, H.B., Nanninga, G.B., Berumen, M.L.** 2015. Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. *Mol. Ecol.* 24: 6241-6255.
- **Saliba, A.E., Westermann, A.J., Gorski, S.A., Vogel, J.** 2014. Single-cell RNA-seq: advances and future challenges. *Nucl. Acids Res.* 42: 8845-8860.
- **Sanger, F., Nicklen, S., Coulson, A.R.** 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74: 5463-5467.
- **Schilling, M.P., Wolf, P.G., Duffy, A.M., Rai, H.S., Rowe, C.A., Richardson, B.A., Mock, K.E.** 2014. Genotyping-by-sequencing for *Populus* population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS One* 9: e95292.
- **Schloss, J.A.** 2008. How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* 26: 1113-1115.
- **Schweyen, H., Rozenberg, A., Leese, F.** 2014. Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *Biol. Bull.* 227: 146-160.
- **Sork, V.L., Squire, K., Gugger, P.F., Steele, S.E., Levy, E.D., Eckert, A.J.** 2016. Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*. *Am. J. Bot.* 103: 33-46.
- **Sovic, M.G., Fries, A.C., Gibbs, H.L.** 2015. AftrRAD: a pipeline for accurate and efficient de novo assembly of RADseq data. *Mol. Ecol. Res.* 15: 1163-1171.
- **Suárez-González, A., Hefer, C.A., Christe, C., Corea, O., Lexer, C., Cronk, Q.C., Douglas, C.J.** 2016. Genomic and functional approaches reveal a case of adaptive introgression from Populus balsamifera (balsam poplar) in P. trichocarpa (black cottonwood). *Mol. Ecol.* 2016 doi: 10.1111/mec.13539.
- **Swaegers, J., Mergeay, J., Van Geystelen, A., Therry, L., Larmuseau, M.H., Stoks, R.** 2015. Neutral and adaptive genomic signatures of rapid poleward range expansion. *Mol Ecol.* 24: 6163-6176.
- **Tin, M.M.Y., Rheindt, F.E., Cros, E., Mikheyev, A.S.** 2015. Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Mol. Ecol. Res.* 15: 329-336.
- **Toonen, R.J., Puritz, J.B., Forsman, Z.H., Whitney, J.L., Fernandez-Silva, I., Andrews, K.R., Bird, C.E.** 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ* 1: e203.

- **Underwood, Z.E., Mandeville, E.G., Walters, A.W.** 2016. Population connectivity and genetic structure of burbot (Lota lota) populations in the Wind River Basin, Wyoming. *Hydrobiologia* 765: 329-342.
- **van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C.** 2014. Ten years of next-generation sequencing technology. *Trends Genet.* 30: 418-426.
- **Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66-74.
- **Vikman, P., Fadista, J., Oskolkov, N.** 2014. RNA sequencing: current and prospective uses in metabolic research. *J. Mol. Endocrinol.* 53: R93-101.
- **Wang, Z., Gerstein, M., Snyder, M.** 2009. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57-63.
- **Wang, S., Meyer, E., McKay, J.K., Matz, M.V.** 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Methods* 9: 808–810.
- **Wink, M.** 2006. Use of DNA markers to study bird migration. *J. Ornithol.* 147(2): 234-244.
- **Zhao, S., Prenger, K., Smith, L.** 2013. Stormbow: A cloud-based tool for reads mapping and expression quantification in large-scale RNA-seq studies. *ISRN Bioinform.* 2013: 481545.
- **Winger, B.M., Hosner, P.A., Bravo, G.A., Cuervo, A.M., Aristizábal, N., Cueto, L.E., Bates, J.M.** 2015. Inferring speciation history in the Andes with reduced-representation sequence data: an example in the bay-backed antpittas (Aves; Grallariidae; Grallaria hypoleuca s. l.). *Mol. Ecol.* 24: 6256-6277.

†Caja 1: Glosario de términos NGS

Barcode Secuencia específica de nucleótidos de corto tamaño que se incorpora en los extremos de los fragmentos a secuenciar. Los *barcodes* son necesarios para identificar el individuo o la muestra de la que proviene cada de unas de las secuencias.

Contig Primer nivel de asociación de los fragmentos secuenciadas en estructuras superiores tras un ensamblaje de lecturas. Fragmento >> *Contig*.

Profundidad de secuenciación Número medio de lecturas que se alinean a una secuencia de bases específica dentro de la muestra de ADN del genoma de referencia. Ej. un genoma secuenciado 30x indica que de media, cada base del genoma está representada en 30 lecturas.

Ensamblaje *de novo* Superposición y concatenación de un conjunto de lecturas de secuencias de ADN o ARN que se realiza en el caso de organismos que no disponen de un genoma de referencia. El ensamblaje *de novo* es un término heredado de la secuenciación de genomas completos, en los que se producen *contigs* y *scaffolds*, y que puede dar lugar a *loci* en el caso de RAD-seq o a transcritos completos en el caso de RNA-seq. Existen distintas aproximaciones metodológicas y algoritmos para realizar ensamblajes *de novo*, que varían en función de la longitud de las lecturas.

Gb Gigabase o 1000 millones de pares de bases.

Isoforma Cada una de las distintas variantes de una misma proteína.

Lecturas pareadas Ficheros de lecturas obtenidos primero de un extremo del fragmento y luego del otro, con o sin direccionalidad conocida. El coste de secuenciar lecturas pareadas es el doble frente a la secuenciación de lecturas simples, dado que se producen dos ficheros en lugar de uno.

Lecturas simples Ficheros de lecturas obtenidos únicamente de uno de los extremos del fragmento. Únicamente se produce un fichero de lecturas.

Biblioteca genómica Conjunto de fragmentos de ácidos nucleicos que tras una serie de procesos en laboratorio (fragmentación, ligación de adaptadores, etc.) está lista para la secuenciación masiva.

Locus/loci Lugar específico del cromosoma donde está localizado un nucleótido, un gen u otra secuencia de ADN. El plural de *locus* es *loci*. Los *loci* pueden presentar distintas formas o variantes que reciben el nombre de alelos.

Mb Megabase o un millón de pares de bases.

Metilación Modificaciones químicas en la molécula de ADN. Consiste en la unión de grupos metilos a las citosinas del ADN. Es un mecanismo de regulación génica que se asocia con una transcripción reducida del gen.

Microarray Superficies sólidas con celdillas en las que se depositan una serie de secuencias de nucleótidos de manera ordenada y que se basan en la capacidad de hibridación entre

dos cadenas de ADN complementarias entre sí. Se utilizan en estudios de alta densidad que analizan muchos genes, ARNs, SNPs, etc. simultáneamente en la muestra de interés. Antes de la eclosión de la secuenciación masiva como RAD-seq o RNA-seq, los *microarrays* constituían la principal alternativa para estudios de expresión o para la determinación de SNPs, y todavía son ampliamente utilizados. El análisis de la intensidad con que se registran determinadas sondas permite inferir la cantidad de secuencias complementarias presentes en una muestra determinada, lo que constituye la base de los estudios de expresión diferencial de ARN.

Nucleótidos degenerados Secuencias de nucleótidos sintetizadas artificialmente que son básicamente similares, pero que difieren en ciertas posiciones. Estos nucleótidos pueden realizar la misma función o incluso producir un mismo resultado que un nucleótido estructuralmente diferente. Se pueden utilizar como cebadores en PCRs o para estimar lecturas redundantes productos de duplicaciones en las PCRs en técnicas como ddRAD-seq.

PCR Reacción en cadena de la polimerasa. Metodología que permite obtener grandes cantidades de moléculas de ADN a partir de unas cantidades mínimas del mismo. La metodología se basa en el uso de unos oligonucleótidos diseñados como cebadores para iniciar la replicación *in vitro*, y en la adición de nucleótidos en cada una de las dos hebras molde mediante la acción de la enzima *Taq*-polimerasa. Todo el proceso se realiza a altas temperaturas para evitar la re-naturalización de las dos cadenas de ADN desnaturalizadas. La PCR es un proceso iterativo que se repite a lo largo de una serie de ciclos, consiguiendo amplificar exponencialmente el número de moléculas de ADN molde, y así poder realizar distintos tipos de análisis, entre ellos, la secuenciación de la molécula.

Pipeline bioinformático Conjunto de procesos de análisis *in silico* que se aplican mediante la concatenación secuencial de aplicaciones de software por la que van transformándose los datos procedentes de experimentos biológicos *in vitro*, entre ellos de secuenciación masiva, hasta conseguir los resultados específicos buscados por el investigador. A modo de ejemplo, existen numerosos pipelines para realizar el ensamblaje de lecturas, análisis de expresión diferencial, anotación funcional, etc.

Pirosecuenciación Método de secuenciación empleado por los sistemas 454 de Roche. El método se basa en la monitorización a tiempo real de la síntesis de ADN mediante la detección de los pirofosfatos liberados en el momento de la incorporación de los nucleótidos. La secuencia de ADN se determina por la intensidad lumínica emitida por la incorporación de cada nucleótido complementario. A la hebra simple sólo se puede añadir uno de cada uno de los cuatro nucleótidos complementarios a la vez, de manera que el nucleótido anterior se degrada antes de estimar la intensidad de luz. El proceso se repite hasta que se determina la secuencia completa.

Ómico/a Neologismo que se refiere al conjunto de metodologías de análisis biológico y recursos derivados de su aplicación a nivel de todo el genoma (genómica), transcriptoma (transcripómica), proteoma (proteómica), metaboloma (metabolómica), etc.

Scaffold Segundo nivel de asociación de los fragmentos secuenciados en estructuras de orden superior. Fragmento >> Contig >> Scaffold

Secuenciación de una única molécula en tiempo real Tecnología de secuenciación de tercera generación por síntesis de una molécula de ADN fija desarrollada por Pacific Biosciences. La técnica no necesita realizar PCR. Utiliza unas cavidades de aluminio y sílice para confinar nanofotones de ~70 nm de diámetro y ~100 nm de profundidad llamadas guías de onda "modo cero" (ZMW). El campo óptico de la cavidad disminuye exponencialmente dentro de la ZMW debido al comportamiento de la luz al atravesar una apertura pequeña. En el fondo de la ZMW se fija una ADN polimerasa con una molécula de ADN molde. Cada una de las cuatro bases nitrogenadas del ADN se marca con un fluoróforo diferente. Así, cuando un nucleótido se incorpora a la cadena incipiente, se libera el fluoróforo y se observa una señal que es recogida en un detector de luminiscencia.

Secuenciación por expansión Secuenciación de una sola molécula basada en el uso de nanoporos que tiene un paso previo en el que se transforma la molécula de ADN en un "expandómero" en un proceso similar a la replicación. El sistema utiliza una polimerasa especial capaz de incorporar cadenas de cuatro bases complementarias a la vez, llamadas "Xprobes" y que llevan incorporadas a la segunda y tercera base unas moléculas especiales coloreadas para representar cada una de las cuatro bases originales del ADN. Una vez sintetizado el expandómero, éste incrementa su tamaño 50 veces respecto del tamaño original de la molécula de ADN y se hace pasar por un nanoporo de estado sólido que lee los colores de las moléculas especiales para determinar la secuencia. El método se encuentra actualmente en desarrollo.

Secuenciación por ligación y codificación por dos bases Metodología de secuenciación de los sistemas SOLiD. Este método utiliza una ADN ligasa en lugar de una polimerasa para identificar la secuencia. En primer lugar se fragmenta la secuencia objetivo y se realiza una PCR para seleccionar los productos por tamaño. Seguidamente se añaden unas sondas de dos bases marcadas con fluorescencia que compiten para ligarse a los cebadores de secuenciación. La detección de la secuencia se consigue interrogando la primera y segunda base en la ligación. Se realizan varios ciclos de ligación, detección y fragmentación con distintos cebadores. La precisión del método es muy elevada

Secuenciación de Sanger Método para determinar la secuencia de nucleótidos de fragmentos de ADN de pequeño tamaño purificados. Se considera la secuenciación de primera generación. En el procedimiento original, el ADN se desnaturalizaba en cada una de sus hebras complementarias, se anillaba una secuencia corta de nucleótidos (cebador) a una de las hebras y la enzima DNA-polimerasa extendía el cebador añadiendo didesoxinucleótidos complementarios (uno cada vez), creando una copia de la hebra. Se precisaba de cuatro reacciones por separado conteniendo un único didesoxinucleótido cada una de ellas (ddG, ddA, ddT y ddC). Cada una de las cuatro reacciones se corría en un gel de poliacrilamida. La secuencia se determinaba a partir de la longitud del fragmento, que se correspondía con la posición en la que se incorporó un didesoxinucleótido. En la actualidad, la reacción de secuenciación se basa en una modificación de la PCR con dideoxinucleótidos

marcados con fluoróforos (moléculas que emiten fluorescencia tras ser excitadas con la luz) y se resuelve mediante una electroforesis capilar, resultando en cromatogramas que representan con colores la secuencia de nucleótidos.

Secuenciación por síntesis Es la metodología de secuenciación de la plataforma Illumina. La técnica utiliza polimerasas especiales y nucleótidos fluorescentes de terminador reversible. Las hebras de ADN y los cebadores de pegan a un portaobjetos donde se realiza un PCR, creando colonias locales de ADN. Mediante métodos ópticos se cuantifica la fluorescencia etiquetada de los nucleótidos, pudiendo así determinar qué nucleótido se une en una posición específica. Este proceso se repite en varios ciclos hasta que se completa toda la secuencia de ADN.

Sistema de transferencia energética de resonancia de la fluorescencia La transferencia energética de resonancia de la fluorescencia (FRET) es una interacción que ocurre solo a muy corta distancia entre dos estados de excitación electrónica de dos moléculas fluorescentes en la que la longitud de onda de emisión de una de ellas coincide con la de excitación de la otra. Ésta es la base para la identificación de secuencias genómicas o aminoacídicas de la plataforma de secuenciación masiva desarrollada por Thermo Fisher Scientific.

SNP Variación en la secuencia de ADN que afecta a una única base en posiciones concretas del genoma.

Splicing alternativo Proceso de edición posterior a la transcripción que se produce tras la obtención del ARN mensajero primario, y por el que, a partir de un pre-ARN mensajero único, se pueden obtener un número de diverso de ARNs maduros, que codificarán para diferentes isoformas de la proteína madura. Estas isoformas se producen mediante splicing alternativo, mecanismo que permite que un mismo gen pueda contener la información necesaria para sintetizar proteínas distintas.

Tecnología de semiconductores Es la tecnología utilizada por la plataforma Ion Torrent. El método se basa en la detección de protones liberados durante el proceso de polimerización del ADN. A diferencia de otros métodos de secuenciación, este método no usa nucleótidos modificados químicamente ni se usan métodos ópticos par la detección de nucleótidos sino por detección de pH.

Variant calling Proceso de detección de polimorfismos en las secuencias obtenidas (lecturas simples o ensambladas).