Dermatología Argentina Vol. 25 Nº 1 Marzo de 2019: 02-08 ISSN 1515-8411 (impresa) ISSN 1669-1636 (en línea)

EDUCACIÓN MÉDICA CONTINUA

Secuenciación masiva de ADN: la próxima generación

DNA massive sequencing: the next generation

Ana Mordoh

RESUMEN

El término secuenciación comprende el análisis de diferentes entidades moleculares a fin de conocer su composición nucleotídica precisa. Con ella es posible analizar un genoma completo (WGS, secuenciación genómica completa), un exoma (WES, secuenciación exómica completa), el transcriptoma (RNA-seq), paneles de genes y patrones genómicos de metilación. Los métodos de secuenciación masiva desarrollados después del método de Sanger, llamados de próxima generación, permiten secuenciar cantidades enormes de ADN e incluyen la pirosecuenciación, la secuenciación por unión y la secuenciación por síntesis. Su principal ventaja consiste en poder secuenciar gran cantidad de ADN en poco tiempo y a un costo relativamente bajo.

Todas las técnicas de secuenciación masiva de próxima generación constan de cuatro pasos: preparación del templado (molde de ADN a secuenciar), preparación de grupos de fragmentos de ADN clonal, secuenciación y análisis de datos. Mediante la secuenciación masiva se busca, por lo general, caracterizar variantes germinales y somáticas en pacientes individuales y, en cohortes con gran número de casos, identificar mutaciones tumorales *drivers*, mutaciones que predispongan al cáncer en la línea germinal o aquellas relacionadas con factores ambientales. Los grandes estudios poblacionales de asociación genómica (GWAS), de casos y controles, utilizan esta técnica.

Palabras clave: secuenciación masiva, secuenciación de próxima generación, secuenciación por síntesis, exoma, variantes génicas, polimorfismos de nucleótido único.

Dermatol. Argent. 2019, 25 (1): 02-08

ABSTRACT

Sequencing means the analysis of several molecular entities, to know their precise nucleotide composition. With this tool, it is possible to evaluate a complete genome (WGS), an exome (WES), a transcriptome (RNA-seq), gene panels and genomic methylation patterns.

Massive sequencing technologies, developed after Sanger sequencing method, known as next generation sequencing (NGS), can sequence huge quantities of DNA and include: pyro-sequencing, union sequencing and sequencing by synthesis. Their main advantage is their ability to sequence large quantities of DNA fast and at a relative low price. All NGS technologies have four basic steps: sample preparation (DNA template), cluster generation, sequencing and data analysis.

Massive sequencing technologies allow characterization of germinal and somatic variants in individual patients and driver mutations in big cohorts, identification of germinal predisposition mutations, as well as those related with environmental factors. Large case-controls population studies of genomic association (GWAS) employ massive sequencing for variant analysis.

Key words: *massive sequencing, next generation sequencing, sequencing by synthesis, exome, genomic variants, single nucleotide polymorphism.*

Dermatol. Argent. 2019, 25 (1): 02-08

Magister en Biología Molecular Médica
Investigadora Asociada, Centro de Investigaciones Oncológicas CIO-FUCA, Ciudad Autónoma de Buenos Aires, Argentina

Contacto del autor: Ana Mordoh
E-mail: ana.mordoh@gmail.com
Fecha de trabajo recibido: 10/12/2018
Fecha de trabajo aceptado: 9/4/2019
Conflicto de interés: la autora declara que no existe conflicto de interés.

El estudio del material genético (ADN) y de su expresión (ARNm) varió a lo largo del tiempo y, junto con estas variaciones, cambió la forma de entender los procesos biológicos. Inicialmente, en la década de 1980, la técnica de PCR posibilitó conocer si determinado fragmento de ADN estaba o no presente en una muestra de tejido; más tarde, la PCR en tiempo real (qPCR) permitió cuantificar la cantidad de ADN presente y, por último, gracias a la RTqPCR, se pudo cuantificar el ARNm expresado en una muestra dada mediante el uso de la transcriptasa-reversa asociada a la qPCR (es decir, que cantidad de genes se estaban expresando en determinado momento, cuantificando sus ARN mensajeros). El conocimiento avanzó un paso más y llevó a desarrollar herramientas que facilitaron conocer la composición nucleotídica precisa del ADN aislado, lo que dio lugar a la identificación de mutaciones y polimorfismos tanto germinales como somáticos.

En la actualidad existen diversas herramientas para el estudio de polimorfismos o mutaciones, algunas acotadas y limitadas (PCR + secuenciación, FISH) y otras más abarcativas, que incluyen el estudio de fragmentos extensos de ADN. Para evaluar la presencia o la ausencia de determinada mutación conocida en un sitio puntual del genoma, es posible realizar técnicas como la PCR, en la que se amplifica un fragmento pequeño de un gen en particular, y luego secuenciarlo. Sin embargo, para evaluar todos los genes o un gran número de ellos, aun aquellos cuyas mutaciones se desconocen, es necesario utilizar técnicas más complejas, conocidas genéricamente como secuenciación masiva de próxima generación (NGS), capaces de abarcar grandes fragmentos de ADN.

El término secuenciación es amplio y comprende el análisis de diferentes entidades moleculares a fin de conocer su secuencia precisa de nucleótidos mediante diferentes técnicas. Con ella es posible analizar un genoma completo (WGS), todas las regiones codificantes del genoma o exoma (WES), el transcriptoma o regiones que se están transcribiendo activamente en un determinado momento (*RNA-seq*), paneles de genes y patrones genómicos de metilación[1,2].

El proceso de secuenciación del genoma humano fue evolucionando a lo largo del tiempo. Frederick Sanger inició, en 1977, el campo de la genómica con su revolucionario desarrollo de la secuenciación del ADN por el método de terminación de cadenas, conocida como secuenciación de primera generación o de Sanger. El desarrollo de instrumentos comerciales con este método produjo hitos tempranos que culminaron en la secuenciación del primer genoma humano. Esta tarea titánica, costosa y lenta, llamada Proyecto Genoma Humano (PGH), fue un gran paso, aunque fueron necesarias nuevas tecnologías que permitieran secuenciar el ADN a mayor velocidad y menor costo para que el conocimiento generado se pudiera aplicar en forma masiva. El método de Sanger fue el primer método de secuenciación y se lo conoce como la primera generación y es considerado, en algunos casos, el método de referencia (*gold standard*) de la secuenciación. Todos los métodos de secuenciación masiva diseñados posteriormente o métodos de próxima generación, son capaces de secuenciar cantidades enormes de ADN a una velocidad mucho mayor y a menor costo e incluyen la pirosecuenciación, la secuenciación por unión y la secuenciación por síntesis. Muchos miles de genomas y exomas han sido secuenciados hasta la fecha y los datos obtenidos han tenido gran impacto en el conocimiento de diversas áreas de la biología. La llegada de las técnicas de próxima generación al mercado ha cambiado el abordaje científico en la investigación básica, clínica y aplicada. Su principal ventaja consiste en poder secuenciar gran cantidad de ADN en poco tiempo y a un costo relativamente bajo[1,2].

La secuenciación masiva revolucionó el conocimiento y permitió la identificación de mutaciones somáticas en genomas de pacientes con cáncer que no habían sido identificadas previamente con técnicas como la secuenciación génica dirigida o la citogenética. Al ser aplicada a muchas muestras del mismo tipo tumoral permitió identificar mutaciones nuevas, recurrentes, no descriptas previamente (novel), algunas de las cuales resultaron ser importantes blancos terapéuticos. El descubrimiento del oncogén BRAF es un ejemplo, ya que este fue identificado al secuenciar genomas de pacientes con diversos tipos de tumores, incluido el melanoma cutáneo[3,4].

Todas las técnicas de secuenciación masiva de próxima generación constan de los siguientes pasos: preparación del templado (molde de ADN a secuenciar), preparación de los clones de cada fragmento de ADN (*clusters*), secuenciación propiamente dicha y análisis de datos.

La preparación del templado consiste en extraer el ADN a ser analizado y romperlo en fragmentos pequeños (de 200-300 pares de bases). De esta forma, se producen 100-200 millones de fragmentos de ADN, con extremos libres en los que se adaptan cebadores universales (*primers*) pasibles de ser hibridados con sus complementarios en el soporte empleado para iniciar la reacción de amplificación clonal. Luego se amplifica clonalmente cada uno de los fragmentos obtenidos mediante una PCR, en emulsión o en un soporte sólido. Así, se obtienen grupos idénticos (*clusters* o racimos) de cada uno de los fragmentos de ADN[2] (Gráfico 1).

El tercer paso consiste en la secuenciación propiamente dicha. En el Gráfico 2 se representa el método de secuenciación por síntesis que realiza el equipo de NGS más uti-

lizado (ILLUMINA®). Este incorpora en forma secuencial nucleótidos terminadores de cadena, marcados con cuatro fluoróforos diferentes (para A, T, C y G). Estos emiten una señal lumínica cuando se incorporan a la cadena de ADN naciente, la que es leída, y luego ese nucleótido marcado se lava y se incorpora un nucleótido sin marcar. Ese proceso se repite con cada uno de los nucleótidos incorporados en cada uno de los millones de fragmentos analizados en forma simultánea. El cuarto y último paso consiste en el análisis informático de todos los datos obtenidos[2].

Durante la secuenciación se generan muchos fragmentos (millones) de ADN secuenciado, llamados lecturas. El promedio de veces que es leída una base en determinada posición se denomina profundidad de lectura o cobertura y se designa con el símbolo x (p. ej., 100x indica que cada posición nucleotídica fue leída 100 veces)[5]. Una vez obtenidas las secuencias de esas lecturas es necesario conocer la ubicación individual de cada una de ellas para rearmar el genoma/exoma original. Para genomas de gran tamaño y cuyas secuencias de referencia ya se han publicado, el método utilizado consiste en mapear los datos obtenidos de las lecturas con las secuencias de un genoma conocido de alta confiabilidad para determinar su correcta ubicación. Este genoma de referencia humano es un genoma ensamblado representativo de la especie humana realizado por un consorcio internacional (GRC) que se utiliza como marco para comparar los genomas en estudio. Se estima que la diferencia de un genoma cualquiera con el genoma de referencia humano oscila entre 0,1 y 1%[6]. Es posible encontrar información de los genomas de referencia humanos, sus actualizaciones y datos concernientes en el siguiente link: https://www.ncbi.nlm.nih.gov/

En el paso siguiente, las lecturas ya mapeadas se colocan una al lado de la otra (alineado) para formar fragmentos cada vez más grandes hasta unir todos los fragmentos analizados (secuenciados). Una vez que las lecturas están mapeadas (con el genoma de referencia) y alineadas (entre sí), se aplican filtros para terminar analizando finalmente solo aquellas de buena calidad. Se obtiene entonces un archivo en formato SAM o BAM, que es una versión binaria de este. El archivo SAM es de suma importancia, ya que contiene toda la información relacionada con este proceso inicial. Mediante la secuenciación masiva se intenta caracterizar variantes germinales y somáticas en pacientes individuales y, en cohortes con gran número de casos, identificar mutaciones tumorales *drivers*, mutaciones que predispongan al cáncer en la línea germinal, aquellas relacionadas con factores ambientales o asociadas con diversas patologías. Al secuenciar un genoma, un exoma o un panel de genes, se buscan las diferencias que hay entre un genoma de referencia normal y el individuo estudiado. Estas diferencias se denominan variantes, tanto en la línea germinal como en la somática. Algunas variantes tienen asignado un *rs* (*rs: reference SNP* o polimorfismo de nucleótido único de referencia), identificado con un número con el que se puede buscar la variante en diversas bases de datos, mientras que otras no poseen un *rs* asignado.

Dado que algunas variantes son artificios de las técnicas utilizadas en el mapeo o en el proceso de secuenciación, se debe realizar un proceso de filtrado para eliminarlas. Con el objetivo de disminuir la cantidad de variantes reales no encontradas (maximizar la sensibilidad, reducir los falsos negativos) y de aumentar la cantidad de errores de secuenciación/alineamiento rechazados (mejorar la especificidad, reducir los falsos positivos), se divide el procesamiento en dos etapas: *variant calling* y *variant quality score recalibra-*

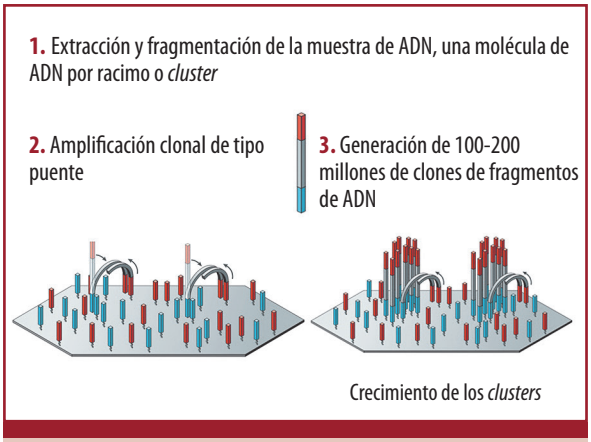

GRÁFICO 1: Preparación del templado de ADN en fase sólida. Unión de los cebadores a un fragmento de ADN monocatenario, que luego se une a su complementario en un soporte sólido para su amplificación por PCR. De esta forma, cada fragmento se amplifica clonalmente (adaptado de referencia 2).

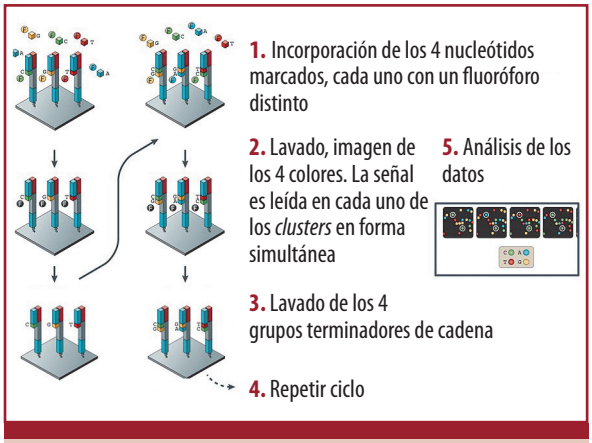

GRÁFICO 2: Ejemplo de método de secuenciación por síntesis. Los cuatro nucleótidos se marcan con fluoróforos distintos, que emiten una señal lumínica al ser incorporados, luego se lavan y se incorpora el mismo nucleótido sin la marca. Cada una de las señales que se incorpora es leída en forma simultánea (adaptado de referencia 2).

tion (VQSR), las cuales intentan resolver cada uno de los problemas mencionados, respectivamente.

En el caso de los tumores, las distintas lecturas obtenidas de muestras germinales y somáticas se mapean contra el genoma de referencia y, además, se comparan entre sí. Al mapearse contra el genoma de referencia se ven diferencias entre el genoma analizado y este, mientras que al comparar (restar) la secuenciación tumoral con la germinal del mismo paciente se obtienen las mutaciones somáticas, es decir, aquellas que adquirió el tumor en su aparición o en su evolución. En la práctica, la detección de todas las alteraciones somáticas y germinales constituye un desafío importante tanto por las limitaciones informáticas de los múltiples algoritmos utilizados para interpretar los datos como por la enorme cantidad de información para analizar. El proceso de encontrar las variantes (diferencias de la muestra analizada con el genoma de referencia) se denomina llamado de variantes o *variant calling* (Gráfico 3).

Del llamado de variantes se obtiene el archivo VCF, que es el archivo más importante de la genómica clínica, ya que es una representación condensada de un genoma/exoma particular. El VCF consiste en un encabezado que contiene la definición de los campos utilizados y una sección con las variantes. Entre los campos más importantes podemos citar: CHROM (número del cromosoma), POS (posición dentro del cromosoma), ID (identificador de la variante en dbSNP o rs), REF (nucleótido/s en el genoma de referencia), ALT (nucleótido/s alternativo/s en la muestra secuenciada), QUAL (calidad de la secuenciación en escala Phred), FILTER (filtro utilizado en la selección de variantes), INFO (información adicional sobre la variante), FORMAT (formato de la información sobre las lecturas y su genotipo). En el link http://samtools.github.io/hts-specs/VCFv4.3.pdf es posible encontrar todas las especificaciones concernientes a los VCF.

GRÁFICO 3: Flujo de análisis de una muestra, llamado y anotación de variantes. Adaptado de las recomendaciones de buenas prácticas de GATK (*Genome Analysis Toolkit*) (https://www.broadinstitute.org/gatk/).

Las diferentes alteraciones analizadas por este método incluyen polimorfismos de nucleótido único (SNP), InDels (pequeñas inserciones o deleciones, por lo general menores de 50 pb) y alteraciones en el número de copias (CNV), cada una de las cuales se analiza según diferentes algoritmos. Los SNP son variaciones de la secuencia del ADN en las que está alterado un solo nucleótido y representan el 90% de toda la variación genética humana. Ocurren aproximadamente cada 100-300 pares de bases a lo largo 6000 millones (6×10^9) de bases del genoma humano diploide y son responsables de los múltiples fenómenos relacionados con la salud humana, como la predisposición a enfermedades o la respuesta a drogas. Se ha propuesto que las variaciones interindividuales respecto de la presencia de SNP en los genes de reparación del ADN contribuyen al riesgo individual de desarrollar tumores esporádicos, aunque esto no ha sido posible de demostrar en todos los casos debido a la naturaleza poligénica del cáncer[7]. La detección de los InDels por WES suele ser más dificultosa que la de los SNP debido a su menor frecuencia y a dificultades en el mapeo. Si bien hay herramientas de alineado de lecturas que contienen SNP, por lo general carecen de sensibilidad y especificidad para lecturas que se superponen con InDels o con alteraciones estructurales. En ocasiones, alteraciones importantes del ADN pueden ser pasadas por alto si la cobertura (X) de la secuenciación es demasiado baja, o si existen regiones genómicas repetitivas o complejas que dificultan la alineación y ensamblado de todos los fragmentos mapeados. Por eso, en los tumores malignos en particular, en los que la progresión suele ser un evento clonal, la cobertura de la secuenciación debe ser suficientemente grande (mayor de 100 X) como para permitir la detección de clones que están subrepresentados dentro de los tumores heterogéneos[1].

Las herramientas utilizadas para el llamado de variantes se siguen desarrollando en forma continua. Los diferentes algoritmos informáticos independientes se combinan, ya que una variante candidata llamada por diferentes algoritmos tiene menos probabilidad de ser un falso positivo que si se la encontró usando solo un algoritmo. Algunas de las herramientas empleadas para llamar las variantes incluyen GATK (*Genome Analysis Toolkit*), VarScan, SAM *tools*, para variantes somáticas y germinales y SomaticSniper, Mutec-2, Strelka, JointSNVMix y SNVMix para el llamado de variantes somáticas en las que se utilizan muestras germinales y somáticas pareadas[1].

Los archivos VCF presentan un gran número de variantes (miles, millones), según lo que se haya secuenciado (panel-exoma-genoma). Con el objetivo de quedarse solo con las variantes de buena calidad y disminuir la cantidad de variantes que pudieron haber sido falsos positivos se aplican filtros. Una vez finalizado el

proceso de filtrado, se recopila la información biológica correspondiente a cada una de las variantes encontradas para analizar su posible relevancia y buscar si han sido informadas previamente. Este procedimiento constituye la etapa denominada de anotación del VCF, en la que se adiciona información relevante a las variantes. En esta etapa se utilizan distintos *softwares*, según lo que se desee analizar. Por ejemplo, ANNOVAR y SNPeff se usan para anotar variantes transcriptas, SKIPPY predice sitios crípticos de *splicing*, *Ensembl Variant Effect Predictor* (VEP), FunSeq y SNPnexus anotan, además de las variantes transcriptas, aquellas en sitios regulatorios o no codificantes[1] (véase Gráfico 3).

Existe una herramienta para la anotación de las variantes en cáncer: CRAVAT (*Cancer-Related Analysis of Variants Toolkit*), que brinda información acerca del mapeo de las variantes anotadas en el genoma de referencia, los posibles transcriptos, el nucléotido de referencia y el alternativo que se encuentra en la mutación, la secuencia y estructura de la proteína codificada por esta. También clasifica las mutaciones en regiones codificantes de genes según su consecuencia funcional y predice su posible implicancia en el cáncer usando dos herramientas: CHASM-3.1 (*Cancer-specific High-throughput Annotation of Somatic Mutations*) y VEST-4 (*Variant Effect Scoring Tool*)[8].

SNP raros, no sinónimos, que alteren la secuencia proteica (codificantes) son candidatos fuertes como productores de enfermedad. La gran cantidad de SNP que se encuentran en los estudios de secuenciación masiva, del orden de 3.000 a 6.000 SNP no sinónimos por exoma, hace muy dificultosa la interpretación de todos ellos. La mayoría de las mutaciones encontradas con la NGS en tumores son pasajeras y solo una mínima parte son conductoras (*driver*), es decir, otorgan una ventaja selectiva a las células. Esto ha llevado a desarrollar métodos computacionales de predicción estadística a fin de evaluar el impacto funcional de aquellas, como las herramientas CHASM o VEST de CRAVAT, comentadas en el párrafo anterior.

Es importante señalar que la capacidad de las herramientas informáticas para identificar con certeza mutaciones nuevas (novel) con algún papel causal en el desarrollo tumoral es limitada. Lo que hacen es priorizar variantes candidatas para estudios posteriores (estudios funcionales en modelos experimentales), que puedan demostrar fehacientemente su implicancia en el desarrollo del fenotipo tumoral[9].

Para identificar las variantes funcionales y *drivers* pueden utilizarse diferentes abordajes. El primero consiste en anotar las variantes mapeadas e identificar aquellas que correspondan a funciones génicas definidas que hayan sido informadas y validadas experimentalmente. El segundo consiste en determinar una asociación estadística en grandes poblaciones entre la variante en estudio y un rasgo determinado (estudios de casos y controles). Los estudios de asociación genómica (GWAS) se basan en esta metodología[10]. El tercero consiste en predecir la naturaleza y la magnitud del impacto funcional de las variantes cuando se localizan regiones codificantes o regulatorias con base en su secuencia.

La secuenciación masiva, tanto WGS como WES, es una herramienta poderosa utilizada actualmente para descubrir mutaciones nuevas en el cáncer, dado que el rastreo (*screening*) de todo el genoma/exoma proporciona un método no sesgado capaz de descubrir alteraciones en genes inesperados, así como asociaciones de polimorfismos con determinadas patologías. Dado que la zona codificante del genoma es aproximadamente un 2% de este, la secuenciación exómica completa (WES) es la técnica de secuenciación masiva más utilizada. Esta técnica permite, a igual presupuesto, aumentar la cobertura (mayor profundidad de lecturas) de la secuenciación en regiones codificantes y, por ende, mejorar la sensibilidad para la detección de variantes de poca frecuencia. La WES es hoy la herramienta de secuenciación masiva más usada, considerando costos y beneficios. Por otro lado, si se tiene en cuenta, además, que los polimorfismos de nucleótido único (SNP) constituyen la mayor parte de la diferencia genética humana (90%), esta herramienta es útil como primera aproximación para el estudio de variantes tanto germinales como somáticas[1], las cuales pueden ser estudiadas *a posteriori* por secuenciación NGS mediante un panel específico de genes con mayor cobertura aún o por PCR cuantitativa.

Sin embargo, la secuenciación exómica (WES) presenta sus limitaciones. Las alteraciones en sitios de promotores génicos (p. ej., la mutación en el promotor del gen TERT en el melanoma) son eventos mutacionales descriptos y significativos en diversos tipos tumorales[11]. La secuenciación exómica puede pasarlos por alto, ya que se focaliza en la secuenciación de regiones codificantes. Por otro lado, tampoco permite detectar deleciones génicas grandes, mayores que el tamaño de las lecturas (aproximadamente 200-300 pares de bases). La pérdida de función en muchos genes supresores de tumores descriptos (CDKN2A en el melanoma o PTEN en los síndromes hereditarios) se produce por grandes deleciones, por lo que no es posible evaluarlas solo con la WES. Las amplificaciones génicas (amplificación de BRAFV600E o en el promotor de TERT en el melanoma) tampoco son identificadas con los datos de la secuenciación crudos. Ambos tipos de alteraciones génicas, amplificaciones y deleciones, pueden ser identificados mediante la ejecución de programas bioinformáticos para la detección del número de copias (CNV) que permiten cuantificar, en

más o en menos, el número de copias génicas a partir del VCF obtenido de la secuenciación.

La genómica, junto con la secuenciación masiva, ha invadido el campo del conocimiento científico y sus aplicaciones son tan extensas como variadas, desde la agricultura hasta la aplicación clínica en medicina. La enorme cantidad de datos que ha aportado generó nuevos cuestionamientos e interrogantes sobre cómo interpretar y jerarquizar la gigantesca cantidad de información obtenida, es decir, cómo extraer los datos relevantes de aquellos pasajeros, esto es, cómo separar la paja del trigo. Sus detractores la señalan como la responsable de que muchos investigadores hayan dejado de plantearse problemas a la manera tradicional (hipótesis) y de insumir mucho más tiempo y recursos que la investigación científica clásica, denostándola como "filatelia molecular", algo así como reemplazar la pregunta de una hipótesis por "secuenciemos a ver qué encontramos"[12].

A pesar de las controversias, la NGS, brazo operativo de la genómica actual, ha aportado herramientas invalorables para comprender la biología de numerosas enfermedades y una parte importante de este conocimiento se aplica hoy para el tratamiento de muchas de ellas, en el largo camino por recorrer de la medicina personalizada.

ABREVIATURAS

A: adenina.
ALT: nucleótido/s alternativo/s en la muestra secuenciada.
ARNm: ARN mensajero.
BAM: *binarian alignment map,* mapa de alineación binaria.
C: citosina.
CHROM: número de cromosoma.
CNV: *copy number variation*, variación del número de copias.
dbSNP: base de datos de SNP.
FILTER: filtro utilizado en la selección de variantes.
FISH: *fluorescence in situ hybridization*, hibridación fluorescente *in situ.*
FORMAT: formato de la información sobre las lecturas y su genotipo.
G: guanina.
GATK: *Genome Analysis Toolkit*, caja de herramientas para el análisis genómico.
GRC: *Genome Reference Consortium*, Consorcio de Genomas de Referencia.
GWAS: *Genome-Wide Association Studies*, Estudios de Asociación Genómica.
ID: identificación de la variante en dbSNP.
InDels: pequeñas inserciones o deleciones.
INFO: información adicional sobre la variante.
NGS: *next generation sequencing*, secuenciación de próxima generación.
PCR: *polymerase chain reaction*, reacción en cadena de la polimerasa.
PGH: proyecto genoma humano.
POS: posición dentro del cromosoma.
qPCR: PCR cuantitativa o en tiempo real.
REF: nucleótido/s en el genoma de referencia.
RNA-seq: secuenciación del transcriptoma.
rs: *reference* SNP, polimorfismo de nucleótido único de referencia.
RTqPCR: PCR cuantitativa por transcriptasa reversa.
SAM: *sequence alignment map*, mapa de secuencias alineadas.
SNP: *single nucleotide polymorphism*, polimorfismo de nucleótido único o simple.
T: timina.
VC: *variant calling*, llamado de variantes.
VCF: *variant calling format*, formato de llamado de variantes.
VQSR: *variant quality score recalibration.*
WES: *whole exome sequencing*, secuenciación exómica completa.
WGS: *whole genome sequencing*, secuenciación genómica completa.

BIBLIOGRAFÍA

1. Ding L, Wendl MC, McMichael JF, Raphael BJ. Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* 2014;7:1-15.
2. Metzker M. Sequencing technologies-the next generation. *Nat Rev Genet* 2010;11:31-46.
3. Davies H, Bignell GR, Cox C, Stephens P, *et ál*. Mutations of the BRAF gene in human cancer. *Nature* 2002;417:949-954.
4. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* 2014; 6:1-17.
5. Sims D, Sudbery I, Ilott N, Heger A, *et ál*. Sequencing depth and coverage: Key considerations in genomic analysis. *Nat Rev Genet* 2014;15:121-132.
6. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. *N Engl J Med* 2014; 370:2418-2425.
7. Barrett JH, Iles MM, Harland M, Taylor JC, *et ál*. Genome-wide association study identifies three new melanoma susceptibility loci. *Nat Genet* 2012;43:1108-1113.
8. González-Pérez A, Mustonen V, Reva B, Ritchie GRS, *et ál*. Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods* 2013;10:723-729.
9. Carter H, Douville C, Stenson PD, Cooper DN, *et ál*. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomic* 2013;14(Suppl 3):S3.
10. Teri M. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010;363:166-176.
11. Shain AH, Yeh I, Kovalyshyn I, Sriharan A, *et ál*. The genetic evolution of melanoma from precursor lesions. *N Engl J Med* 2015; 373:1926-1936.
12. Steensma DP. The beginning of the end of the beginning in cancer genomics. *N Engl J Med* 2013;368: 2138-2140.

CUESTIONARIO DE EVALUACIÓN

1) *Señale cuáles de las siguientes técnicas pueden utilizarse para conocer la composición nucleotídica precisa de un fragmento de ADN:*
A- PCR, FISH, Sanger.
B- qPCR, citogenética.
C- PCR + secuenciación, Sanger, secuenciación masiva de próxima generación.
D- PCR + secuenciación, qPCR.

2) *La secuenciación masiva de próxima generación puede utilizarse para secuenciar:*
A- Un panel de genes.
B- Un genoma.
C- Un exoma.
D- Todas las opciones son correctas.

3) *La secuenciación masiva de próxima generación consiste en los siguientes pasos:*
A- Preparación del material, amplificación clonal, secuenciación, análisis de datos.
B- Preparación del material, amplificación clonal, pirosecuenciación y secuenciación por síntesis.
C- Preparación del material, amplificación clonal, PCR.
D- Preparación del material, amplificación clonal, PCR, Sanger.

4) *Señale la principal ventaja de la secuenciación masiva de próxima generación:*
A- Es más precisa que el método de Sanger.
B- Es más rápida que el método de Sanger.
C- Permite secuenciar grandes cantidades de ADN a un costo relativamente bajo.
D- Las opciones B y C son correctas.

5) *Las técnicas de secuenciación de próxima generación se usan en la actualidad en:*
A- Estudios de asociación genómica (GWAS), estudios de mutaciones somáticas en cáncer, estudios de predisposición a enfermedades.
B- Determinación de presencia o ausencia de determinados patógenos.
C- Cuantificación de la carga viral.
D- Estadificación de pacientes con cáncer.

6) *Una vez obtenidas todas las lecturas, luego de la secuenciación masiva, en el caso de genomas/exomas es necesario:*
A- Mapearlas contra un genoma de referencia y alinear los fragmentos entre sí.
B- Descartar todas las lecturas que hayan leído la misma posición nucleotídica.
C- Analizar cada lectura por separado antes de realizar el llamado de variantes.
D- Anotar las variantes para ver si están repetidas.

7) *Uno de los principales objetivos de la secuenciación de próxima generación es la detección de las variantes (rs) de la muestra analizada respecto de un genoma de referencia. En relación con esto señale la respuesta correcta:*
A- A la búsqueda de las variantes, respecto del genoma de referencia normal, se la denomina llamado de variantes (*variant calling*).
B- El archivo obtenido del llamado de variantes se denomina VCF (*variant calling format*) y representa un archivo condensado de todas las variantes encontradas, con sus datos asociados más relevantes.
C- Después de encontrar las variantes en determinada muestra hay que anotarlas, es decir, ver cuál es su función biológica (en qué base de datos están informadas, cuál es su función asociada, etc.).
D- Todas las opciones son correctas.

8) *La secuenciación exómica presenta las siguientes ventajas respecto de la secuenciación genómica:*
A- Se centra en las regiones codificantes del ADN, que corresponden a un 2% de todo el genoma, lo que permite, a igual presupuesto, aumentar la cobertura de secuenciación (profundidad de lectura) en esas regiones.
B- Permite obtener información de sitios intrónicos, promotores o reguladores génicos.
C- Permite obtener información de paneles de genes.
D- Es más difícil de interpretar, ya que brinda información de regiones no codificantes del ADN.

9) *El hecho de encontrar variantes nuevas (no reportadas previamente o novel) en la secuenciación (diferencias con el genoma de referencia):*
A- Es un indicio contundente de patología.
B- Indica que la secuenciación estuvo mal realizada.
C- Señala variantes candidatas de patología, que hay que confirmar a posteriori si efectivamente lo son con estudios funcionales (validación experimental), ya que podrían tratarse de SNP que formen parte de la variación genética normal entre las personas.
D- Si no fueron informadas previamente como asociadas a patología, es muy difícil que lo sean.

10) *La profundidad de lectura en la secuenciación masiva de próxima generación es la cantidad de veces que una misma posición nucleotídica es leída y se designa con el símbolo x. Esta constituye un dato importante para tener en cuenta, ya que:*
A- Indica nucleótidos repetidos en determinadas regiones.
B- Permite la identificación de subpoblaciones clonales no representadas en todo el tumor, lo que es particularmente importante en tumores sólidos malignos, ya que estos suelen estar compuestos por poblaciones celulares heterogéneas.
C- Señala sitios de deleción en genes supresores de tumor.
D- No es un dato relevante para tener en cuenta cuando se lee un trabajo de NGS en tumores.

Respuestas correctas, Vol. XXIV - Nº4, 2018
1. **B** / 2. **C** / 3. **D** / 4. **D** / 5. **C** / 6. **C** / 7. **D** / 8. **A** /9. **B** / 10. **C**