Fundación JL Castaño SEQC

SEQC ML
Sociedad Española de Medicina de Laboratorio

2017-2018

APLICACIONES CLÍNICAS DE LAS TÉCNICAS ACTUALES DE BIOLOGÍA MOLECULAR

Ed. Cont. Lab. Clin 37: 33 - 40

TÉCNICAS DE SECUENCIACIÓN MASIVA (NGS).

María Santamaría González.

Servicio de Bioquímica Clínica. Hospital Clínico Universitario Lozano Blesa. Zaragoza.

José Miguel Lezana Rosales.

Unidad de Bioinformática. Centro de Genética Molecular Genetaq. Málaga.

1. FUNDAMENTO DE LA TÉCNICA

La secuenciación de nucleótidos que conforman la molécula de ADN representa el análisis más detallado de su estructura y una herramienta eficaz para identificar variantes en el material genético. La secuenciación Sanger supuso en los años 70 una revolución importante en Genética Humana, sentando precedente en el origen de la era genómica. Sin embargo, en los últimos años se han desarrollado nuevas plataformas de secuenciación denominadas de alto rendimiento o nueva generación (*Next-Generation Sequencing o NGS*) capaces de generar paralelamente, y de forma masiva, millones de fragmentos de ADN en un único proceso de secuenciación, elevando significativamente el rendimiento a un menor coste y ofreciendo ventajas significativas respecto a los sistemas convencionales.

Con la incorporación de las nuevas tecnologías de secuenciación a la práctica clínica, cada laboratorio gestiona el procedimiento a desarrollar en cada caso. Con la finalidad de mejorar el rendimiento del laboratorio se desarrollan paneles NGS que permiten el análisis en paralelo de múltiples genes o regiones seleccionadas del ADN que se relacionan con fenotipos parecidos o solapantes. Estos paneles pueden proporcionar un primer método de análisis genético. Si no se detecta ninguna alteración importante que explique el fenotipo reportado, el facultativo determinará si continúa ampliando el estudio con la secuenciación del exoma o del genoma completo.

El procedimiento general de las técnicas de secuenciación masiva incluye en términos generales:

1. Fragmentación del ADN: se generan pequeños fragmentos de distinto tamaño y posteriormente se unen a sus extremos unas moléculas llamadas adaptadores. El conjunto de todos los fragmentos de ADN unidos a sus adaptadores se conoce como librería.

2. Enriquecimiento (opcional): consiste en seleccionar exclusivamente las áreas de interés antes de la secuenciación.

   2.1. Métodos de captura por hibridación en fase sólida: se construyen microarrays con oligonucleótidos unidos covalentemente a una superficie sólida con secuencias complementarias a aquellas que se quieren seleccionar. Las áreas de interés del ADN fragmentado hibridan con estos oligonucleótidos y permanecen unidas a la superficie sólida. Roche/NimbleGen ha desarrollado y comercializa estos microarrays como el desarrollado para la captura del exoma (SeqCap EZ *Exome capture*)

   2,2. Métodos de amplificación mediante PCR: la amplificación se produce mediante el anclaje del fragmento de ADN, a través de sus adaptadores, a una superficie sólida. Las plataformas que requieren la amplificación clonal de los fragmentos de ADN se denominan de segunda generación. Existen plataformas, denominadas de tercera generación, que no requieren la amplificación de los fragmentos de ADN, reduciendo el tiempo de trabajo y abaratando el proceso de secuenciación.

3. Ligación del material amplificado a una superficie sólida donde se llevará a cabo la reacción de secuenciación.

4. Secuenciación del material genético: la secuenciación y detección de las bases ocurren al mismo tiempo en todas las moléculas de ADN (secuenciación masiva y paralela)

5. Creación de archivos con información de la secuenciación y alineamiento de las lecturas contra un genoma de referencia: los cientos de miles de lecturas obtenidas son almacenadas en archivos fastq (dos por paciente en el caso de *paired-end*). Se requieren el empleo de potentes herramientas informáticas para proceder al alineamiento de las secuencias, y posteriormente generar una base de datos con las variantes obtenidas.

A pesar de estas peculiaridades, cada plataforma de secuenciación masiva se basa en principios químicos distintos que generan diferencias cuantitativas y cualitativas. A continuación se detallan las técnicas de secuenciación más utilizadas por las plataformas disponibles en el mercado:

Secuenciación por síntesis o polimerización: sistemas dependientes de ADN polimerasa

- **Secuenciador semiconductor:** se basa en el registro de los cambios de pH producidos durante la incorporación de bases durante la síntesis de ADN. Esta tecnología es incorporada por las plataformas de Ion Torrent™ (Ion PGM, Ion PROTON) (https://www.thermofisher.com/es/es/home/brands/ion-torrent.html)

- **Terminación reversible cíclica:** se utilizan nucleótidos que incorporan un grupo terminador marcado con moléculas fluorescentes, de manera que la incorporación del siguiente nucleótido a la cadena no se efectúa hasta que el terminador es retirado

tras lectura de la señal fluorescente. Las plataformas que utilizan esta técnica son *Illumina* (https://www.illumina.com/) y *Helicos BioSciences* (http://seqll.com/).

La plataforma de *Illumina* permite realizar secuenciación de tipo *paired-end* mediante la cual es posible leer un fragmento de ADN por los dos extremos. Se secuencia el fragmento amplificado por los dos extremos en vez de por uno solo. Ofrece ventajas a la hora de realizar el alineamiento contra el genoma de referencia, ya que permite estimar el tamaño del fragmento original y situarlo con mayor precisión en el genoma mejorando así la cobertura de las zonas de interés y también resulta útil para la detección de eventos estructurales.

- **Secuenciación en tiempo real:** estas técnicas son muy rápidas porque no frenan la secuenciación tras la incorporación de cada base para llevar a cabo la lectura ni invierten tiempo en ciclos de lavado. Tampoco requieren amplificación, ni sufren fenómenos de desfase. Esta tecnología desarrollada por *Pacific Biosciences* (http://www.pacb.com/products-and-services/) e incorporada en su plataforma SMRT (*Single Molecule Real Time*) contiene un dispositivo que permite limitar el campo de observación lo suficiente como para captar exclusivamente la fluorescencia emitida por la incorporación de nucleótidos marcados en tiempo real.

Otras técnicas de secuenciación: todavía en fase de desarrollo

- **Basadas en nanoporos:** Se basan en la identificación de las distintas bases de la cadena de ADN gracias a una señal óptica o por la variación que se produce en una corriente eléctrica al pasar la cadena a través de un nanoporo anclado a una membrana. Desarrollada por Oxford Nanopore Systems (https://nanoporetech.com/)

- **Observación directa con microscopía:** desarrollada por compañías como *ZS Genetics* (http://alllseq.com/knowledge-bank/emerging-technologies/zs-genetics/) que utiliza la microscopía electrónica y permite leer la secuencia del ADN directamente por métodos ópticos sin necesidad de amplificación.

Las técnicas de secuenciación NGS generan principalmente, tres tipos de ficheros: FASTQ, SAM/BAM (alineamiento) y VCF (anotación).

Fichero FASTQ: texto de entrada estándar que contiene los datos crudos (lecturas o reads) obtenidos por el secuenciador. Permite almacenar la secuencia biológica junto con las calidades asociadas a cada nucleótido de la secuencia. Este fichero es reconocido por las herramientas bioinformáticas encargadas del alineamiento.

Fichero SAM (*Sequence Alignment/Map format*): representación de alineamientos de secuencias contra un genoma o secuencia de referencia.

Fichero BAM (*Binary Alignment/Map format*): versión comprimida del formato SAM. Permite realizar un indexado para tener acceso directo a las posiciones genómicas.

Los datos de alineamiento se visualizan empleando herramientas de visualización interactiva como IGV (Integrative Genomics Viewer)

Fichero VCF (*Variant Call Format*) se obtiene a partir de los ficheros SAM/BAM. Permite almacenar las variaciones de la secuencia con respecto al genoma contra el que se alinea: SNPs, inserciones, deleciones y variantes estructurales, junto con ciertas anotaciones opcionales derivadas de diferentes bases de datos

Los pares de reads correctamente alineados se utilizan en la detección de SNP (*Single Nucleotide Polymorphism*), pequeñas inserciones y deleciones y en la estimación del número de copias. Los reads de un par no alineados que muestran una distancia o una orientación inesperada, se analizan como indicadores potenciales de variantes estructurales. Por último, el acoplamiento *de novo* de reads no alineados con la referencia proporciona predicciones de variantes estructurales.

Un aspecto importante en la NGS es el número de veces que cada base del genoma está presente en los reads de secuenciación producidos; es decir, el número de veces que se lee cada nucleótido. Este valor se denomina cobertura y es uno de los factores determinantes para evaluar la fiabilidad del nucleótido asignado a esa posición del genoma. El diseño de paneles de NGS debe garantizar una adecuada cobertura en todos los genes analizados. La gran capacidad de estas tecnologías para generar datos genómicos va a incrementar significativamente la información genética asociada que disponemos en la actualidad.

2. ÁMBITO DE APLICACIÓN

Debido a su gran rendimiento, este tipo de plataformas es idóneo para un sin fin de estudios a gran escala imposibles de abordar con ningún otro tipo de tecnología existente hasta la fecha debido al enorme coste que ello supondría. Entre las principales aplicaciones destacan:

◦ **Diagnóstico molecular de enfermedades hereditarias:** se pueden seguir varias estrategias pero recae en el laboratorio la responsabilidad de realizar una gestión adecuada del método a utilizar en cada caso.

> **Secuenciación dirigida:** la más utilizada, consiste en el aislamiento, enriquecimiento y secuenciación de regiones específicas del genoma (regiones codificantes de interés y zonas intrónicas flanqueantes). Se desarrollan paneles genéticos de apoyo al diagnóstico que permiten la identificación de variantes, polimorfismos, reordenamientos y variantes somáticas en baja frecuencia, en numerosas muestras de forma simultánea. Esta técnica es la más utilizada para el estudio de enfermedades con componente hereditario que ya tienen mutaciones o genes asociados a su etiología. También es de aplicación en la detección de mutaciones presentes en tumores, lo que permite evaluar si un tumor podría responder a una determinada terapia con fármacos dirigidos.

Secuenciación de exoma: consiste en seleccionar las secuencias codificantes del genoma, y finalmente secuenciarlas. Esta estrategia es empleada en la identificación de genes causantes de patología cuando se desconoce la causa genética de un fenotipo clínico concreto o para aquellas enfermedades con alta heterogeneidad fenotípica y genética. No indicada en aquellos casos en los que se conoce el gen causante de la patología o si ya se ha identificado la mutación responsable en otros miembros de la familia.

Secuenciación de Genoma completo: el reto de esta estrategia de análisis es obtener una interpretación eficiente y fiable antes de utilizarla con fines diagnósticos en genética humana. Sin embargo, se está aplicando en el campo de la Microbiología para obtener genomas completos tanto de bacterias como de virus o patógenos fúngicos. El estudio genómico que permite la identificación y caracterización de diferentes microogranismos de una comunidad microbiana por extracción directa de ADN, es lo que se conoce como metagenómica.

º **Diagnóstico prenatal de aneuploidías:** la secuenciación de ADN fetal presente en el plasma materno es un método no invasivo y eficaz utilizado para detectar las principales aneuploidías.

º **Análisis del transcriptoma** (*ARN-Seq* - Transcriptoma completo): con la secuenciación masiva del ADN complementario se genera información global sobre el contenido de ARN de una muestra, incluyendo mensajero, ribosomal y de transferencia y otros ARNs no codificantes. Proporciona una forma eficiente de medir niveles de expresión génica, identificar eventos de *splicing* alternativo, fusión génica SNPs de manera simultánea.

º **Estudio del microtranscriptoma:** identificación, cuantificación y caracterización de cientos de pequeñas moléculas de ARN no codificante (*SmallRNAs: miRNAs, snoRNAs, piRNAs*) cuya función puede ser clave para el funcionamiento celular.

º **Identificación de zonas de interacción proteína-ADN** (ChIP-seq): esta aplicación es posible a través de la técnica *ChIP-Seq* que combina la inmunoprecipitación de la cromatina con la secuenciación masiva.

º **Identificación de patrones de Metilación** (metiloma): utilizada para el estudio de la regulación de procesos clave como la diferenciación celular, el desarrollo o la aparición de enfermedades.

Durante los próximos años, la lista de aplicaciones crecerá sin duda, al igual que la sofisticación con las que se realizan las aplicaciones existentes.

3. LIMITACIONES DE LA TÉCNICA

Los factores que influyen en la calidad de los resultados obtenidos son muy variados y van a influir de forma decisiva en el tipo de errores cometidos y en la fiabilidad de los datos derivados del análisis, incluyen: conservación de la muestra (tejido fresco o fijado), pureza de la muestra (en el caso de tumores, porcentaje de células no tumorales presentes en la muestra), técnica de secuenciación, cobertura y herramientas de análisis e interpretación de resultados.

Entre las limitaciones más importantes de estas técnicas a tener en cuenta, destacan:

- La secuenciación exómica no es útil para detectar inversiones, traslocaciones, grandes deleciones heterocigotas o zonas poliméricas con más de 8 bases repetidas.
- En función de la estrategia (panel, exoma) y de la metodología utilizada, en algunas ocasiones quedan zonas codificantes del genoma sin cubrir.
- El incremento vertiginoso en la cantidad de datos genómicos generados hace necesario el desarrollo de herramientas de almacenamiento de datos de elevada capacidad y con sistema de seguridad integrado.
- La obtención de un gran número resultados inesperados o no solicitados por el paciente, requiere de la aplicación de los métodos generales para interpretación de las variantes aisladas y el análisis integral del contexto genómico y fisiológico específico de cada caso. Es habitual realizar un análisis y consulta en bases de datos poblacionales para conocer la evidencia científica sobre las posibles implicaciones.
- La dificultad en la interpretación de los resultados: el análisis de datos no sigue un modelo único y la combinación de distintas herramientas de software y bases de datos pueden dar lugar a resultados variables. La bioinformática surge para dar respuesta a los problemas computacionales que aparecen con las plataformas de NGS. Esta disciplina aporta información importante para un análisis e interpretación eficiente de los datos. Incluye: Análisis primario: Análisis de imagen y asignación de bases Análisis secundario: Alineamiento, detección de variantes y anotación. Análisis terciario: Búsqueda y análisis de las variantes más probablemente patogénicas. Análisis conjunto de variantes y fenotipo del paciente.

BIBLIOGRAFÍA

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res [Internet]. Oxford University Press; 2010 Apr [cited 2017 Feb 22];38(6):1767–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20015970

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics [Internet]. Oxford University Press; 2011;27(15):2156–8. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet [Internet]. Nature Research; 2011 Apr 10 [cited 2017 Feb 12];43(5):491–8. Available from: http://www.nature.com/doifinder/10.1038/ng.806

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. Evaluation of next generation sequencing platforms for population targeted sequencing studies.Genome Biol.10(3):R32 (2009)

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics [Internet]. 2009 Jul 15 [cited 2017 Feb 12];25(14):1754–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19451168

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. Oxford University Press; 2009 Aug 15 [cited 2017 Feb 14];25(16):2078–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19505943

Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 11(1):31-46 (2010)

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics [Internet]. Oxford University Press; 2012 Nov 1;28(21):2747–54. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts526

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol [Internet]. Nature Research; 2011 Jan [cited 2017 Feb 14];29(1):24–6. Available from: http://www.nature.com/doifinder/10.1038/nbt.1754

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 94(3):441-8 (1975)

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74(12):5463-7 (1977)

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform [Internet]. Oxford University Press; 2013;14(2):178–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22517427

ISBN 978-84-697-4013-2 – Febrero 2018 (recibido para publicación Junio 2017).